HAL
open science

COSMO : un modèle bayésien des interactions sensori-motrices dans la perception de la parole

Raphael Laurent

▶ **To cite this version:**

Raphael Laurent. COSMO : un modèle bayésien des interactions sensori-motrices dans la perception de la parole. Autre [cs.OH]. Université de Grenoble, 2014. Français. NNT : 2014GRENM063 . tel-01113286v2

HAL Id: tel-01113286

https://theses.hal.science/tel-01113286v2

Submitted on 3 Jul 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

UNIVERSITÉ DE GRENOBLE

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE

Spécialité : **Informatique**

Arrêté ministériel : 7 août 2006

Présentée par

Raphaël LAURENT

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

préparée au sein du **Laboratoire de Psychologie et de NeuroCognition**, du **Laboratoire d'Informatique de Grenoble**, du **Laboratoire GIPSA**, et de l' **École Doctorale de Mathématiques, Sciences et Technologies de l'Information, Informatique**

COSMO : un modèle bayésien des interactions sensori-motrices dans la perception de la parole

Thèse soutenue publiquement le **8 octobre 2014**, devant le jury composé de :

M. Pierre BESSIÈRE
DR CNRS, Institut des Systèmes Intelligents et de Robotique, Co-Directeur de thèse

M. Julien DIARD
CR CNRS, Laboratoire de Psychologie et de NeuroCognition, Directeur de thèse

M. Emmanuel DUPOUX
Professeur EHESS, Laboratoire de Sciences Cognitives et Psycholinguistique, Rapporteur

M. Yves LAPRIE
DR CNRS, Laboratoire lOrrain de Recherche en Informatique et ses Applications, Rapporteur

M. Roger MOORE
Professor, University of Sheffield, Examinateur

M. Pierre-Yves OUDEYER
DR INRIA, INRIA Bordeaux Sud-Ouest, Examinateur

M. Pascal PERRIER
Professeur Grenoble-INP, GIPSA-Lab, Président

M. Jean-Luc SCHWARTZ
DR CNRS, GIPSA-Lab, Co-Directeur de thèse

Table des Matières

Table des Matières **i**

Liste des Illustrations **vi**

1 Introduction **1**
1 La parole, un objet d'étude complexe 1
2 Le choix de la modélisation computationnelle 1
3 Contribution 2
4 Plan de lecture 4

2 Des théories et des modèles pour traiter de la variabilité en parole **7**
1 Le débat entre les théories motrices et auditives 8
1.1 Théories motrices 8
1.1.1 La production de la parole à partir d'invariants moteurs 8
1.1.2 La perception de la parole par récupération des invariants moteurs dans le signal acoustique 11
1.2 Théories auditives 12
1.2.1 La perception de la parole à partir des invariants acoustiques . . 13
1.2.2 La production de la parole à partir de gestes moteurs correspondant à des invariants acoustiques 14
2 Intégration perceptuo-motrice 15
2.1 Intégration théorique dans des modèles cognitifs conceptuels 15
2.1.1 Des modèles perceptuo-moteurs de la production de la parole . . 15
2.1.2 Des modèles perceptuo-moteurs de la perception de la parole . . 17
2.1.3 Taxonomie des différentes théories et modèles 19
2.2 Données expérimentales en faveur des théories intégratives 20
2.3 Des outils conceptuels pour réfléchir à l'intégration perceptuo-motrice . . 22
2.3.1 Modèles internes 22
2.3.2 Les conditions de l'émergence de la communication 24
3 Des modèles computationnels perceptuo-moteurs pour la reconnaissance et la perception de la parole 25
3.1 Travaux en Reconnaissance Automatique de la Parole 25
3.1.1 Les principes de base des systèmes de reconnaissance automatique 25
3.1.2 Limitations principales des systèmes actuels et apport potentiel des connaissances motrices 26

3.1.3 Paramétrisation articulatoire 28
3.1.4 Combinaison des paramètres acoustiques et articulatoires 28
3.1.5 Traitement temporel 29
3.2 Modèles cognitifs 29
3.2.1 Les travaux computationnels du groupe de Luciano Fadiga . . . 30
3.2.2 Le modèle neurocomputationnel *ACT* proposé par Bernd Kröger 31
3.2.3 Le modèle *PRESENCE* proposé par Roger Moore : un cadre de réflexion générale qui transcende la recherche en parole 33
4 Notre chemin de modélisation 35

3 *COSMO* : Un modèle formel et générique d'agent communicant pour l'étude quantitative des interactions sensori-motrices lors de la communication parlée 37
1 Modélisation conceptuelle de la situation de communication parlée 37
2 *COSMO* : Un modèle formel probabiliste d'agent communicant 40
2.1 Hypothèse d'internalisation de la situation de communication dans l'architecture cognitive de l'agent communicant 40
2.2 Dépendances probabilistes du modèle *COSMO* 41
3 Inférences probabilistes dans *COSMO* 43
3.1 Principe de l'inférence bayésienne 43
3.2 Implémentation probabiliste des tâches de production et de perception de la parole 44
3.3 Inférences probabilistes pour les tâches de perception 45
3.3.1 Implémentation dans le cadre d'une théorie motrice 45
3.3.2 Implémentation dans le cadre d'une théorie auditive 46
3.3.3 Implémentation dans le cadre d'une théorie perceptuo-motrice . 48
3.4 Inférences probabilistes pour les tâches de production 49
3.4.1 Implémentation dans le cadre d'une théorie motrice 49
3.4.2 Implémentation dans le cadre d'une théorie auditive 49
3.4.3 Implémentation dans le cadre d'une théorie perceptuo-motrice . 49
3.5 Résumé des calculs d'inférence 50
4 Théorème d'indistinguabilité des théories motrice et auditive en perception de la parole 51
4.1 Apprentissage par interactions avec un agent maître 52
4.2 Conditions suffisantes d'indistinguabilité 53
4.3 Discussion 55
5 Conclusion 56

4 Exploitation du modèle *COSMO*, dans un cadre théorique simplifié, pour comparer les approches motrice et auditive de la perception 59
1 Prise de position 60
2 Instantiation du modèle *COSMO* pour comparer quantitativement les approches auditives et motrices de la perception 61
2.1 Domaines de définition des variables du modèle 62
2.2 Formes paramétriques 62

3 Algorithmes d'apprentissage 64
3.1 Interactions avec un agent maître 64
3.1.1 Les propriétés du modèle de l'environnement 65
3.1.2 Les propriétés du modèle du maître 66
3.1.3 Données d'apprentissage 67
3.2 Apprentissage supervisé du classifieur auditif 68
3.3 Apprentissage de compétences motrices par accomodation 69
3.3.1 Paradigme d'imitation 69
3.3.2 Inférence probabiliste pour la tâche d'imitation 70
3.3.3 L'algorithme d'apprentissage de compétences motrices par accomodation : du babillage orienté vers des cibles 71
3.3.4 Dynamique de l'apprentissage 72
3.3.5 Propriétés des modèles appris 74
4 Comparaisons au sein de *COSMO* des prédictions des théories motrice, auditive et perceptuo-motrice de la perception 76
4.1 Apprentissage et indistinguabilité 77
4.2 Méthode de comparaison 78
4.2.1 Données d'évaluation 78
4.2.2 Implémentations de la tâche de perception 78
4.2.3 Un score global pour décrire les performances des modèles . . . 79
4.3 Résultats : comparaisons des performances des différents modèles 80
4.3.1 Des dynamiques d'évolution différentes dans la branche auditive et dans la branche motrice. 81
4.3.2 Comparaison de la robustesse aux dégradations des différents modèles 82
4.3.3 Impact de la non-linéarité sur les performances des modèles . . . 83
5 Conclusion 83

5 Synthèse de syllabes réalistes dans le cadre de *COSMO* avec un modèle géométrique de conduit vocal : *VLAM* 87
1 Modélisation de la syllabe 88
1.1 Différentes manières de modéliser le signal de parole 88
1.1.1 Comment caractériser la parole dans un espace auditif 88
1.1.2 Comment caractériser la parole dans un espace moteur 91
1.2 Le choix du modèle *VLAM* pour assurer le passage des représentations motrices aux représentations auditives 93
1.2.1 *VLAM* : the *Variable Linear Articulatory Model* 93
1.2.2 La transformation articulatori-acoustique et sa complexité . . . 94
1.3 Comment caractériser les syllabes de type plosive-voyelle 96
1.3.1 Variabilité des consonnes plosives et modèles de production de séquences plosive-voyelle 96
1.3.2 Invariants potentiels associés au lieu d'articulation et modèles de perception de séquences plosive-voyelle 98
1.3.3 Nos choix de modélisation 101

2 Synthèse de syllabes plosive-voyelle dans le cadre de *COSMO* 102
2.1 Synthèse de voyelles : de la variabilité autour de gestes prototypiques . . 102
2.2 Synthèse de plosives : un geste de perturbation superposée à la voyelle . . 103
3 Évaluation des syllabes plosive-voyelle produites 105
3.1 Résultats de simulation 105
3.2 Comparaisons avec des données réelles : étude des lois du locus 107
4 Conclusion 110

6 Apprentissages sensori-moteurs réalistes au sein de *COSMO* et application à des tâches de perception de syllabes 111
1 *COSMO-S* : le modèle *COSMO* étendu au traitement des syllabes 111
1.1 Principes, hypothèses de modélisation et enjeux de simulations 112
1.2 Variables du modèle et domaines de définition 113
1.3 Distribution de probabilité conjointe du modèle *COSMO-S* 114
1.4 Formes paramétriques 115
2 Implémentation d'algorithmes d'apprentissage dans le cadre de *COSMO-S* 119
2.1 Choix d'un cycle développemental 119
2.2 Données d'apprentissage 120
2.3 Apprentissage du système auditif par association 121
2.4 Apprentissage du système sensori-moteur par accomodation 122
2.4.1 Inférence probabiliste pour la tâche d'imitation de syllabes . . . 122
2.4.2 L'algorithme d'apprentissage des modèles internes par accomodation 123
2.4.3 Illustrations de la convergence de l'algorithme 124
2.5 Apprentissage du système moteur par accomodation 131
2.5.1 Inférence probabiliste pour la tâche d'imitation supervisée de syllabes 131
2.5.2 L'algorithme d'apprentissage des répertoires moteurs par imitation supervisée 133
2.5.3 Illustrations de la convergence de l'apprentissage : développement des idiosyncrasies 134
2.6 Synthèse des différents apprentissages réalisés 137
3 Principaux résultats 138
3.1 L'apprentissage de compétences motrices par un agent purement perceptif permet de faire émerger une ébauche de phonologie 138
3.1.1 Phonologie à partir de représentations auditives ? 138
3.1.2 Phonologie à partir de représentations motrices ! 140
3.1.3 Le rôle de la langue de la mère (*motherese*, mamanais) dans l'émergence de la phonologie 141
3.2 L'utilisation de connaissances motrices apporte de la robustesse pour la perception de syllabes en conditions dégradées 143
3.2.1 Inférences probabilistes pour les tâches de perception au sein du modèle *COSMO-S* 143
3.2.2 Méthode de comparaison 145

3.2.3 Comparaisons des performances des modèles moteur, auditif, et sensori-moteur lors de tests de perception de syllabes dans le bruit 146
4 Conclusion 147

7 Conclusion générale 149
1 Contributions principales 149
2 Discussion et perspectives 152
2.1 *COSMO* : un modèle de l'agent communicant, un modèle de la communication ou un modèle d'internalisation ? 152
2.2 Le débat entre les théories auditive, motrice et perceptuo-motrice 154
2.3 Performances et robustesse de *COSMO* 156
2.4 L'algorithme d'apprentissage par accommodation 157
2.5 L'apprentissage des premiers stades de la communication 158
2.6 Situations d'apprentissage 160
3 Le mot de la fin 161

Bibliographie 163

Liste des Illustrations

1.1 Le modèle *COSMO-S* de traitement des syllabes 3

2.1 Gestes de constriction, articulateurs et paramétrisation géométrique 9
2.2 Les trois étapes de la Phonologie Articulatoire 10
2.3 Évolution temporelle des degrés de constrictions lors de la production du mot anglais *'pan'* 11
2.4 L'architecture du modèle *DIVA* 16
2.5 Le modèle de Skipper 18
2.6 La Théorie de la Perception pour le Contrôle de l'Action 19
2.7 Les modèles internes pour le contrôle moteur 22
2.8 Un mélange de paramètres articulatoires et acoustiques pour la reconnaissance de la parole 30
2.9 *ACT* : un modèle neurocomputationnel du contrôle sensori-moteur de la parole . . . 32
2.10 Le modèle PRESENCE 34

3.1 Les trois phases de la communication 38
3.2 La situation de communication 39
3.3 L'agent communicant 41
3.4 Le modèle *COSMO* 50
3.5 Inférences probabilistes dans *COSMO* 51
3.6 Taxonomie des modèles de production et de perception 51
3.7 Le scénario d'apprentissage supervisé 52

4.1 Le modèle *COSMO* et ses différents sous-systèmes 63
4.2 Scénario d'interactions avec un agent maître 64
4.3 Modélisation des non-linéarités 65
4.4 Les distributions de probabilité sur les stimuli reçus par l'agent 67
4.5 Les prototypes auditifs de l'agent 68
4.6 Apprentissage du système sensori-moteur 74
4.7 Apprentissage du système moteur 75
4.8 Résumé des performance des modèles sous forme de matrice de confusion 80
4.9 Robustesse aux dégradations au cours de l'apprentissage 81
4.10 Comparaison de la robustesse des modèles aux dégradations 82
4.11 Impact de la non-linéarité et de la durée de l'apprentissage sur la robustesse des modèles aux dégradations 84

5.1 Représentation schématique de la cochlée 89
5.2 Représentation schématique des traitements auditifs 90
5.3 Les paramètres articulatoires du modèle *VLAM* 94
5.4 Différentes manières de réaliser un /u/ 95
5.5 Différentes caractéristiques acoustiques du /d/ en fonction du contexte vocalique . . 96
5.6 Les plosives sont sous-spécifiées articulatoirement 97
5.7 Les plosives caractérisées par un invariant acoustique 98
5.8 Les plosives caractérisées par un invariant moteur 99
5.9 Les plosives caractérisées par les équations du locus 100
5.10 Modification du paramètre *Apex* de *VLAM* 105
5.11 Nos voyelles synthétiques 106
5.12 Nos plosives synthétiques dans l'espace acoustique 107
5.13 Comparaison avec d'autres données synthétiques 108
5.14 Comparaison aux équations du locus de Sussman 109

6.1 Le modèle *COSMO-S* de traitement des syllabes 115
6.2 Chronologie du développement 119
6.3 Évolution de l'entropie des modèles internes de l'agent 125
6.4 L'entropie des distributions $P(S_V \mid M_V)$ apprises 127
6.5 Précision du modèle interne appris en fonction des zones de l'espace moteur des voyelles 128
6.6 Les gestes moteurs de voyelles les mieux appris 129
6.7 Les voyelles produites au cours de l'apprentissage par imitation 130
6.8 Les gestes moteurs de voyelles produits au cours de l'apprentissage par imitation . . 130
6.9 Évolution de l'entropie des répertoires moteurs de voyelles au cours de l'apprentissage 135
6.10 Les gestes moteurs de voyelles produits au cours de l'apprentissage par imitation supervisée 136
6.11 Les représentations auditives associées aux syllabes dans l'espace des voyelles 139
6.12 Les représentations auditives associées aux syllabes dans l'espace des consonnes . . . 140
6.13 L'apprentissage du lieu d'articulation des plosives 141
6.14 Le rôle du *motherese* dans l'émergence de la phonologie 143
6.15 Matrice de confusions pour la perception motrice sans bruit 146
6.16 Robustesse aux dégradations des différents modèles évalués sur des tâches de perception de syllabes 147

7.1 Le modèle *COSMO* 149
7.2 *COSMO*, un modèle intégrateur 150
7.3 Le modèle *COSMO-S* de traitement des syllabes 151

Chapitre 1

Introduction

1 La parole, un objet d'étude complexe

La parole est une faculté dont l'usage nous paraît tout à fait naturel. Pourtant, si du point de vue physiologique la parole semble aujourd'hui assez bien comprise, autant en perception qu'en production, il nous reste encore beaucoup de choses à comprendre sur la nature des représentations et des processus cognitifs qui la gouvernent.

Historiquement, les recherches à ce sujet ont tout d'abord été relativement cloisonnées, avec d'une part l'étude de la parole du point de vue du locuteur, et d'autre part l'étude de la parole du point de vue de l'auditeur. Plus récemment, les liens entre perception et action ont pris une place croissante dans de nombreux domaines des sciences cognitives. C'est également le cas dans cette thèse où nous centrons notre approche sur les interactions entre perception et action en parole.

La parole et le langage sont classiquement structurés autour de deux niveaux d'articulation. Le premier concerne les combinaisons de mots et les connaissances lexicales, syntaxiques et sémantiques ; nous ne nous y intéressons pas. En revanche, nous centrons notre travail sur le second niveau d'articulation, qui concerne les combinaisons de sons, de phonèmes et de syllabes.

Nous étudions donc les interactions entre perception et action dans la production et la perception de phonèmes et de syllabes. Ce champ d'étude est encore très large, et l'on y trouve par exemple les questions suivantes. Comment caractériser les phonèmes, une unité majeure du second niveau d'articulation, en terme de propriétés sensori-motrices ? Peut-on quantifier l'apport des connaissances motrices pour la perception de la parole ?

Dans ce domaine de recherche, trois ensembles de théories sont le centre d'un débat ancien : les théories motrices, les théories auditives et les théories perceptuo-motrices.

2 Le choix de la modélisation computationnelle

De nombreux arguments sont avancés de part et d'autre dans ce débat théorique, mais tant que le débat reste théorique justement, il y a peu de moyens de comparer leur poids relatif. Nous ne connaissons pas d'observation expérimentale qui, considérée isolément, soit décisive.

La modélisation conceptuelle permet d'organiser les données de la littérature et la connaissance en général. Dans le cadre de travaux multidisciplinaires, comme c'est le cas pour l'étude

de la parole, elle permet de faire émerger dans un domaine un petit nombre d'idées clés autour desquelles s'articule ce que l'on sait, et ce sont ces idées qui passent d'un domaine à l'autre.

Le travail réalisé dans cette thèse trouve son origine dans le pari selon lequel la modélisation computationnelle fournit un cadre rigoureux rendant possible des comparaisons systématiques des différentes théories et de leurs propriétés. La modélisation bayésienne apporte un formalisme qui permet l'écriture de modèles cognitifs arbitrairement complexes, dans un cadre mathématique unique. C'est ce formalisme que nous adoptons comme langage d'expression de nos modèles.

Dans ce cadre, nous héritons de deux travaux précurseurs. Il s'agit tout d'abord de la thèse de Clément Moulin-Frier (Moulin-Frier, 2011 ; Moulin-Frier *et al.*, 2010, 2011). Contrairement à notre travail, centré sur la modélisation d'un agent en situation de communication en ligne, Moulin-Frier s'est intéressé au problème de l'émergence des codes phonologiques dans des sociétés d'agents sensori-moteurs en interaction. Son objectif était de rendre compte des régularités évidentes dans les langues du monde, malgré leur variété, ce que l'on appelle les « universaux du langage ». Pour atteindre cet objectif, il a proposé des modèles bayésiens d'agents en interaction, incluant notamment un lien fort entre leurs connaissances sensorielles et leurs connaissances motrices. Ces modèles et la réflexion qui est à l'origine de leur conception constituent une base de notre travail.

Le second travail dont nous héritons est la thèse d'Estelle Gilet (Gilet, 2009 ; Gilet *et al.*, 2011). Ce travail portait sur un enjeu différent du nôtre : la modélisation de la boucle perception-action impliquée dans la lecture et l'écriture de lettres manuscrites. Mais, comme dans notre contexte, la question des relations entre connaissances perceptives et connaissances motrices y était centrale. L'inférence bayésienne a permis à Gilet de prédire le fonctionnement du système de reconnaissance de lettres selon que les connaissances motrices étaient mises en jeu ou non. Nous reprenons ce thème, ainsi que la méthode générale, pour comparer la perception de syllabes lorsque les connaissances de production de parole sont activées ou non.

3 Contribution

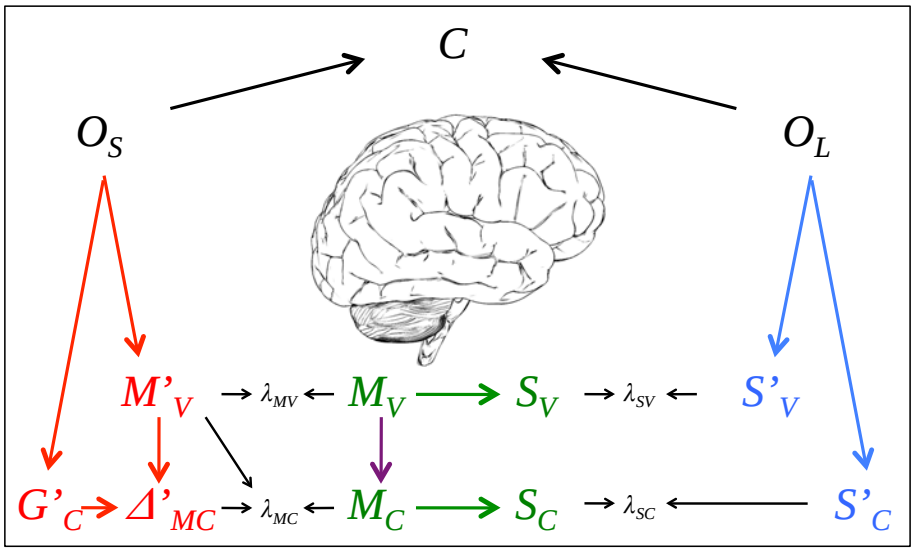
Nous identifions trois principaux axes de notre contribution, et rappelons les références des articles de journaux ou d'actes de conférences où les travaux correspondants ont été publiés.

1. La première contribution consiste en un travail de formalisation : *COSMO* et *COSMO-S* sont deux modèles mathématiques que nous avons définis. Ces deux modèles sont exprimés dans un langage probabiliste et définis selon la méthodologie de la Programmation Bayésienne.

   Le modèle *COSMO* (chapitre 3) est un modèle général et abstrait d'un agent en situation de communication en ligne, qui permet de définir formellement ce que sont les tâches de production et de perception pour des agents qui suivent les théories auditives, motrices, ou perceptuo-motrices. Ce travail (Moulin-Frier *et al.*, 2012) est une variante du travail préalable de Moulin-Frier *et al.* (2010).

   Le second modèle, *COSMO-S* (qui est résumé par la figure 1.1 et sera décrit en détails au chapitre 6), est l'instanciation de *COSMO* pour le cas des syllabes de type plosive-voyelle. C'est un modèle réaliste, basé sur un dictionnaire de syllabes obtenu par simulations dans

le modèle de conduit vocal *VLAM*, et qui fait intervenir des grandeurs articulatoires et acoustiques justifiées vis-à-vis de la littérature (chapitre 5).



$$
\begin{aligned}
&P(O_S\ G'_C\ M'_V\ \Delta'_{MC}\ \lambda_{MV}\ \lambda_{MC}\ M_V\ M_C\ S_V\ S_C\ \lambda_{SV}\ \lambda_{SC}\ S'_V\ S'_C\ O_L\ C)\\
&\quad = P(O_S) \times P(M'_V \mid O_S) \times P(G'_C \mid O_S) \times P(\Delta'_{MC} \mid M'_V\ G'_C) \times\\
&\qquad P(\lambda_{MV} \mid M'_V\ M_V) \times P(\lambda_{MC} \mid M'_V\ \Delta'_{MC}\ M_C) \times\\
&\qquad P(M_V) \times P(S_V \mid M_V) \times P(M_C \mid M_V) \times P(S_C \mid M_C) \times\\
&\qquad P(\lambda_{SV} \mid S_V\ S'_V) \times P(\lambda_{SC} \mid S_C\ S'_C) \times\\
&\qquad P(O_L) \times P(S'_V\ S'_C \mid O_L) \times\\
&\qquad P(C \mid O_S\ O_L)\ .
\end{aligned}
$$

Figure 1.1: **Le modèle *COSMO-S* de traitement des syllabes**, décrit schématiquement par un modèle graphique (en haut), et par sa distribution de probabilité conjointe (en bas). Les connaissances motrices (en rouge) et auditives (en bleu) sont reliées par un modèle cognitif de la transformation articulatori-acoustique (en vert et violet). Pour plus de détails, se reporter au chapitre 6.

(2) La seconde contribution principale est un théorème, que nous nommons « thérorème d'indistinguabilité » (chapitre 3). Ce théorème montre que les voies auditives et motrices de traitement de la parole d'un agent communicant peuvent encoder exactement la même information, les rendant alors indistinguables expérimentalement. En d'autres termes, sous certaines hypothèses, il est imposible de savoir par l'observation extérieure du processus si un agent communicant fait appel à des connaissances auditives ou motrices pour réaliser une tâche de perception de parole (Laurent *et al.*, 2013a).

Ce constat explique en partie la stagnation du débat autour des théories auditives et motrices de la parole (chapitre 2). Nous avons défini formellement les hypothèses qui conditionnent ce théorème (chapitre 3). Les mettre en défaut nous permet d'identifier des

conditions qui rendraient ces théories distinguables expérimentalement.

La première condition d'indistinguabilité concerne une hypothèse d'après laquelle la perception est réalisée en conditions idéales. Nous retombons sur l'idée selon laquelle étudier la perception de la parole en conditions adverses, par exemple en présence de bruit, permet de mettre au jour l'utilisation de connaissances motrices (voir la section 2.2 du chapitre 2). Mais nous dépassons ce cadre : en effet, grâce à notre modélisation mathématique, nous pouvons de plus étudier *comment* les connaissances motrices sont utilisées en conditions bruitées (voir Moulin-Frier *et al.* (2012) et le chapitre 4).

La seconde condition de l'indistinguabilité concerne l'hypothèse d'un apprentissage parfait. Nous explorons donc cette voie, plus originale, de l'apprentissage d'un agent communicant, son décours temporel, et l'évolution du contenu informationnel des voies auditives et motrices (chapitre 4). Nous proposons un algorithme, suivant des étapes inspirées du développement de l'enfant (chapitre 6). Son cœur est un modèle nouveau d'apprentissage de connaissances motrices en parole (incluant le babillage canonique), que nous nommons « apprentissage par accomodation », permettant d'obtenir un agent communicant qui n'a pas une connaissance parfaite (irréaliste) de la transformation articulatori-acoustique, qui développe des idiosyncrasies de production de parole, et qui pourtant s'adapte au bain acoustique ambiant.

(3) Enfin, la troisième contribution principale de notre travail est une mise en œuvre expérimentale des modèles *COSMO* et *COSMO-S*. Les principales observations résultant de nos simulations sont les suivantes :

- une étude de l'émergence de la phonologie pendant l'apprentissage des syllabes (chapitre 6). Nous montrons que la notion de catégorie vocalique est implicitement contenue dans les connaissances auditives, mais pas la consonne qui apparaît en revanche dans les connaissances motrices (Laurent *et al.*, 2012).
- une étude quantitative de la perte de performance des traitements auditifs, moteurs et perceptuo-moteurs dans la perception en conditions bruitées, dans le cas général (chapitre 4) ou dans le cas des syllabes (voir Laurent *et al.* (2013b) et le chapitre 6). Nous montrons notamment que des connaissances motrices apportent de la robustesse pendant la perception.
- une étude de la vitesse de convergence des apprentissages des voies auditives et motrices (chapitre 6). Nous montrons que les connaissances motrices sont plus lentes à se mettre en place, et mènent à des choix idiosyncratiques de production.

4 Plan de lecture

Nous choisissons dans ce document de concentrer l'essentiel de l'étude bibliographique (chapitre 2), mais de repousser l'exposition de certains pans de littérature pour les introduire aux moment opportun : nous rappelerons ainsi quelques éléments sur la modélisation des plosives en contexte au chapitre 5, et résumerons des modèles conceptuels de l'apprentissage de la parole par le bébé au chapitre 6.

Le reste de cette thèse comporte six chapitres.

Chapitre 2 Des théories et des modèles pour traiter de la variabilité en parole

Nous y présentons les grandes lignes du débat entre théories auditives, théories motrices, et théories perceptuo-motrices de la communication parlée. Nous rappelons les principaux résultats expérimentaux et les modèles conceptuels. Nous évoquons les solutions d'ingénierie dans un domaine connexe, celui de la reconnaissance automatique de la parole, et les distinguons du problème de la proposition d'un modèle cognitif des processus d'intégration perceptuo-motrice en perception de la parole.

Chapitre 3 *COSMO* : Un modèle formel et générique d'agent communicant pour l'étude quantitative des interactions sensori-motrices lors de la communication parlée

Nous décrivons le développement théorique du modèle computationnel d'agent communiquant *COSMO*, qui fournit un cadre unificateur générique permettant d'étudier les interactions perceptuo-motrices en production et en perception de la parole. Nous mettons en évidence des hypothèses garantissant l'indistinguabilité des théories motrice et auditive de la perception, et proposons des principes permettant de s'en écarter.

Chapitre 4 Exploitation du modèle *COSMO*, dans un cadre théorique simplifié, pour comparer les approches motrice et auditive de la perception

Nous présentons une instanciation de ce modèle *COSMO* dans un cadre théorique abstrait grâce auquel nous présentons et illustrons en détail un algorithme original d'apprentissage de compétences motrices, par accomodation, à partir d'entrées perceptives uniquement. Nous étudions le développement d'idiosyncrasies motrices et comparons quantitativement les vitesses de convergence des apprentissages auditif et moteur ainsi que la baisse de performance des modèles auditif et moteur sur une tâche de perception en conditions dégradées.

Chapitre 5 Synthèse de syllabes réalistes dans le cadre de *COSMO* avec un modèle géométrique de conduit vocal : *VLAM*

Un travail de modélisation de syllabes de type plosive-voyelle nous permet de proposer une synthèse de données articulatoires et acoustiques grâce à un modèle de conduit vocal. Ces données montrent des patrons de variabilité réalistes, nous les utilisons au chapitre suivant.

Chapitre 6 Apprentissages sensori-moteurs réalistes au sein de *COSMO* et application à des tâches de perception de syllabes

L'implémentation d'une version du modèle *COSMO* étendue au traitement des syllabes permet de montrer que des connaissances motrices apprises de manière réaliste grâce à notre algorithme d'apprentissage par accomodation ont deux propriétés intéressantes : elles permettent l'émergence d'une ébauche de phonologie, et elles viennent complémenter efficacement des connaissances purement perceptives, notamment en conditions adverses (dans le bruit).

Chapitre 7 Conclusion générale

Ce dernier chapitre résume notre contribution et discute de quelques questions ouvertes.

Chapitre 2

Des théories et des modèles pour traiter de la variabilité en parole

1 Le débat entre les théories motrices et auditives **8**

2 Intégration perceptuo-motrice **15**

3 Des modèles computationnels perceptuo-moteurs pour la reconnaissance et la perception de la parole **25**

4 Notre chemin de modélisation **35**

La parole est la principale forme orale de la communication humaine. Elle est basée sur la combinaison suivant une certaine syntaxe d'unités lexicales, les mots du langage, qui sont porteurs de sens, afin de former des phrases. Ces mots sont eux-mêmes formés en combinant un petit ensemble de sons : les phonèmes, voyelles et consonnes, qui sont les unités de base du langage parlé. La linguistique, c'est-à-dire la discipline scientifique qui étudie le langage, fait donc intervenir de nombreux domaines d'études, centrés sur la sémantique, la syntaxe, le lexique, la morphologie, la phonologie, la phonétique. Dans le cadre de cette thèse, nous nous concentrons sur l'étude des représentations et des processus cognitifs permettant de faire le lien entre les sons du langage et les catégories phonétiques.

Cet espace scientifique est régi par deux disciplines partenaires : la phonologie et la phonétique. La phonologie étudie les systèmes sonores du langage du point de vue de leurs structures et de leurs propriétés contrastives pour la communication. Une langue est alors caractérisée au niveau phonologique par un ensemble d'unités distinctives : les phonèmes. La phonétique quant à elle étudie les propriétés physiques et cognitives de ces unités : leur contenu acoustique, perceptif, moteur, et donc leur fonctionnement cognitif. L'articulation entre phonologie et phonétique, visant à expliciter ce qu'est un phonème, comment il se caractérise cognitivement, à la fois individuellement et en référence aux autres phonèmes de la langue, est au cœur du présent travail.

Se pose alors la question de savoir ce qui confère son identité à une catégorie phonétique. Qu'est-ce qui fait qu'un /a/ est un /a/[1], quelles sont les propriétés physiques associées à la catégorie phonologique correspondante ? Deux types d'approches ont été proposés pour répondre à

[1] Conformément à l'usage, les phonèmes sont notés /./ dans cette thèse.

cette question, qui ont donné naissance à deux ensembles de théories cognitives. Les premières proposent de chercher des invariants auditifs dans le domaine acoustique pour caractériser le son du /a/. Les secondes proposent au contraire de chercher des invariants moteurs, en argumentant que le son du /a/ n'a pu être produit que grâce à un ensemble de commandes motrices puis musculaires envoyées aux articulateurs orofaciaux et permettant le contrôle du conduit vocal, et que donc la caractérisation première du /a/ est articulatoire. Ainsi, les travaux de recherche sur la parole font intervenir deux grands courants cognitifs, les théories auditives et les théories motrices, pour lesquelles l'unité minimale de référence est de nature acoustique et motrice respectivement.

1 Le débat entre les théories motrices et auditives

Cette section décrit les théories motrices et auditives de la parole, en production et en perception, ainsi que les principaux arguments sur lesquels elles reposent. L'objectif n'est pas de présenter en détails l'ensemble des travaux des cinquante dernières années, mais plutôt de mettre en lumière les principaux enjeux du débat.

1.1 Théories motrices

Les théories motrices considèrent que dans le signal de parole l'unité minimale de référence est de nature motrice. Bien que l'information qui transite par le signal de parole soit de nature acoustique, ce signal n'est que la conséquence des gestes moteurs à partir desquels il a été produit. Du point de vue des théories motrices, c'est la représentation motrice de ces gestes qui d'une part sert de cible pour la production, et constitue d'autre part une connaissance nécessaire à la perception de la parole. Tous les développements théoriques autour des théories motrices sont partis d'un même lieu, les Laboratoires Haskins[2], et nous verrons que les différentes propositions sont fortement imbriquées.

1.1.1 La production de la parole à partir d'invariants moteurs

La principale théorie de production de la parole ne faisant intervenir que des représentations motrices est la Phonologie Articulatoire. La présentation que nous en faisons ici s'appuie sur les articles fondateurs de Browman et Goldstein (Browman et Goldstein, 1986, 1989, 1992 ; Goldstein et Fowler, 2003). Nous présentons cette théorie dans ses grandes lignes, le lecteur intéressé pourra se référer à Fougeron (2005) pour plus de détails.

Dans le cadre de la Phonologie Articulatoire, l'unité de base est le geste articulatoire permettant de produire les constrictions[3] souhaitées dans le conduit vocal du locuteur. Cette unité intervient à deux niveaux de description. D'une part elle sert à la représentation de l'action articulatoire permettant de contrôler le processus de production de parole. D'autre part elle sert également à la représentation des contrastes phonologiques. Par exemple, la différence entre « bas » et « cas » vient du fait que le /b/ de « bas » est une consonne bilabiale produite

[2] http://www.haskins.yale.edu/

[3] Une constriction est un rétrécissement appliqué localement dans le conduit vocal et qui se traduit par une signature acoustique majeure. Ainsi, le mot « bas » commence par une constriction labiale, avec une fermeture des lèvres pour le /b/.

en fermant les lèvres, alors que le /k/ de « cas » est une consonne vélaire obtenue grâce à une constriction au niveau de l'arrière du palais. Ce que défend la Phonologie Articulatoire, c'est que le geste permettant d'obtenir la constriction (fermeture des lèvres ou remontée de la langue à l'arrière du palais dans notre exemple) fournit également une bonne caractérisation de l'unité phonologique produite, dans la mesure où ce geste est d'une part invariant au sein de la catégorie phonologique, et d'autre part contrastif par rapport aux autres catégories.

Ainsi, le geste est l'unité motrice minimale qui, décrivant des dynamiques de mouvements articulatoires, caractérise les unités phonologiques produites. Il faut cependant noter que tout mouvement articulatoire n'est pas un geste au sens de la Phonologie Articulatoire. Seules les actions articulatoires ayant un but distinctif sont considérées comme des gestes, car le geste est également l'unité du contraste phonologique.

Le but du geste est de contrôler la formation ou le relâchement d'une constriction cible. Pour cela, la réalisation d'un geste peut nécessiter l'intervention de plusieurs articulateurs. Le modèle décrit dans Browman et Goldstein (1992) utilise des variables contrôlant la géométrie du conduit vocal pour caractériser les constrictions cibles. Le Tableau 2.1 présente la correspondance entre le type de constriction, les paramètres géométriques et les articulateurs sollicités.

Organe contrôlant la constriction	**Paramètres géométriques décrivant la constriction**	**Articulateurs sollicités**
Lèvres	Protrusion des lèvres Aperture des lèvres	Lèvre supérieure Lèvre inférieure Mandibule
Pointe de la langue	Position de la constriction de la pointe de la langue Degré de constriction de la pointe de la langue	Pointe de la Langue Corps de la langue Mandibule
Dos de la langue	Position de la constriction du corps de la langue Degré de constriction du corps de la langue	Corps de la langue Mandibule
Velum	Aperture vélaire	Velum
Glotte	Aperture glottale	Glotte

Figure 2.1: **Gestes de constriction, articulateurs et paramétrisation géométrique.** Les cinq types de gestes sont réalisés en coordonant plusieurs articulateurs pour obtenir des constrictions spécifiées par des variables géométriques. (Cette figure est adaptée de Fougeron (2005).)

La Phonologie Articulatoire est implémentée dans un système computationnel comprenant trois étapes, qui sont décrites Figure 2.2.

1. La première étape assure le passage de l'énoncé à produire à une partition[4] des gestes à effectuer. Cette partition décrit la répartition temporelle des différentes constrictions permettant de réaliser la succession d'unité phonologiques présentes dans l'énoncé de départ. Ces constrictions sont caractérisées par leur lieu (labiale, dentale, alvéolaire, palatale, vélaire, uvulaire, pharyngale) et par leur degré de fermeture (occlusion, critique, étroite,

[4]Ici « partition » est à comprendre au sens musical : chaque geste a son rôle à jouer au cours du temps.

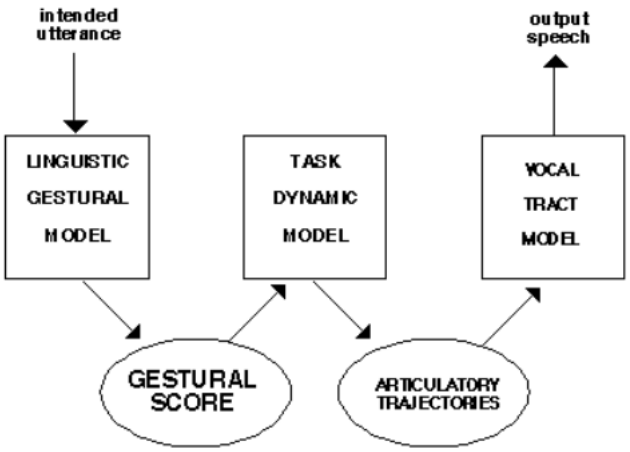


Figure 2.2: **Les trois étapes de la Phonologie Articulatoire**, d'après Browman et Goldstein (1990).

moyenne, large). Aux objectifs articulatoires spécifiés par la dynamique de ces gestes sont associées des contraintes de phasage (Kelso et Tuller, 1987), de raideur et d'amortissement (Browman et Goldstein, 1990). Plusieurs gestes peuvent donc se recouvrir dans le temps, totalement ou partiellement. Ces recouvrements sont obtenus à partir d'un ensemble de règles issues de la linguistique et de la phonétique articulatoire.

(2) Dans la seconde étape, le passage de la partition de gestes aux trajectoires des différents articulateurs se fait grâce au *task dynamics model* (Saltzman et Kelso, 1987). Ce modèle dynamique repose sur les deux hypothèses suivantes : (a) la principale tâche dans la production de la parole est de contrôler et de coordonner les mouvements d'un ensemble d'articulateurs (plutôt que les mouvements individuels d'articulateurs considérés séparément) ; (b) la coordination de ces mouvements peut être caractérisée en utilisant des équations dynamiques. La figure 2.3 montre l'évolution temporelle des variables de degrés de constrictions pour la production du mot anglais *'pan'*. On y voit apparaître en rectangles grisés les éléments phonologiques de la partition gestuelle définie par le *linguistic gestural model*, puis en lignes continues les trajectoires articulatoires spécifiées par le *task dynamics model* en prenant en compte les contraintes biomécaniques des articulateurs impliqués.

Il est intéressant de remarquer que sur la Figure 2.3, l'aperture vélaire augmente avant le geste de constriction de la pointe de la langue. Il s'agit d'un phénomène connu de nasalisation de la voyelle précédant une consonne nasale (Anderson, 1976) qui apparaît ici comme la simple conséquence de la coordination des gestes articulatoires au sein de la partition. De même, l'utilisation des gestes comme représentations de base et leur coordination en partition permet d'expliquer les phénomènes de coarticulation (Öhman, 1966) ou de variations allophoniques (Trager et Bloch, 1941), qui traduisent le fait que la réalisation d'un phonème est liée à son contexte phonétique.

(3) Dans la troisième étape, un modèle de conduit vocal (Rubin *et al.*, 1981) permet de génèrer le signal acoustique à partir des trajectoires articulatoires. Il s'agit d'un modèle

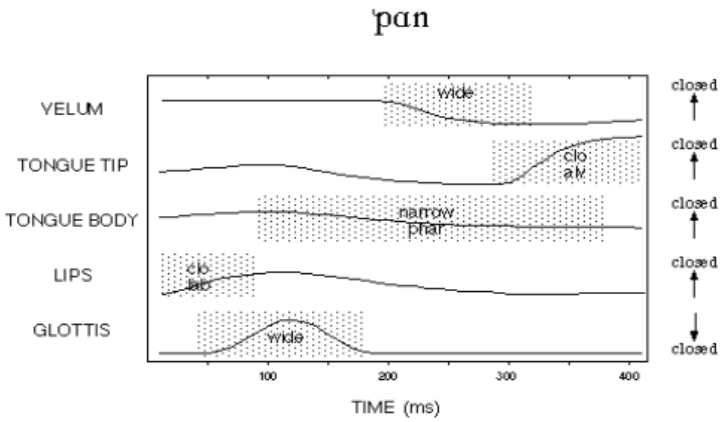


Figure 2.3: **Évolution temporelle des degrés de constrictions lors de la production du mot anglais *'pan'*.** Cette figure provient de `http://www.haskins.yale.edu/research/gestural.html`.

géométrique dans le cadre duquel les paramètres articulatoires permettent de calculer la fonction d'aire du conduit vocal, de laquelle est déduite la fonction de transfert acoustique utilisée pour produire l'onde sonore.

Ainsi, la Phonologie Articulatoire défend l'hypothèse selon laquelle la cible de la production est une unité motrice, le geste de constriction du conduit vocal, qui suffit pour spécifier la production des trajectoires articulatoires et de l'onde sonore. Les sous-systèmes computationnels réalisant les étapes (2) et (3) ont pour but de décrire des systèmes physiques et le font en tirant parti des équations qui les régissent. L'étape (1) quant à elle est au contraire de nature purement cognitive puisqu'elle assure la correspondance entre ce qui doit être énoncé et les représentations motrices permettant de le faire. Il peut être argumenté que, dans le cas d'énoncés fréquents, les partitions de gestes correspondantes sont regroupées dans un répertoire moteur que le locuteur s'est constitué.

Nous qualifions la Phonologie Articulatoire de théorie motrice de la production au sens où elle ne fait intervenir que des représentations et processus moteurs.

1.1.2 La perception de la parole par récupération des invariants moteurs dans le signal acoustique

La célèbre *théorie motrice de la perception de la parole* (Liberman et Mattingly, 1985) défend l'hypothèse selon laquelle la parole serait spéciale. La perception de l'information phonétique se ferait au sein d'un système biologique séparé, un module cognitif spécifique qui permettrait d'utiliser des connaissances sur la production pour réaliser la tâche de perception de la parole. Ce module spécialisé permettrait de recouvrer les gestes intentionnels du locuteur, qui constitueraient la base des catégories phonologiques. Pour la théorie motrice de la perception (Liberman et Mattingly, 1985, page 2), « les objets de la perception de la parole sont les gestes phonétiques intentionnels du locuteur, représentés dans le cerveau en tant que commandes motrices invariantes contrôlant les mouvements des articulateurs [pour pouvoir atteindre

les configurations correspondant aux phonèmes produits]. Ces commandes de gestes[5] sont la réalité physique sur laquelle reposent les notions phonétiques traditionnelles »[6] (Liberman et Mattingly, 1985, page 2). L'unité minimale considérée pour la parole est donc le geste, et percevoir la parole c'est percevoir les gestes.

Une variante de la théorie motrice a été proposée par Fowler (1986) dans le cadre de la théorie réaliste de la perception directe. La différence principale porte sur le premier point : la théorie de Fowler se réfère au courant général de la théorie gibsonienne de la perception directe, selon laquelle tout acte perceptif réfère aux causes physiques qui l'ont induit (Gibson, 1986). A ce titre, la théorie motrice de la parole ne serait qu'un cas particulier d'une théorie plus générale de la perception, et la parole n'est en aucun cas *spéciale*. Une conséquence, fournissant une autre différence entre Fowler et Liberman, est que pour la première l'objet de la perception est directement la configuration de tous les articulateurs du conduit vocal qui est directement encodée dans le signal acoustique, alors que pour la théorie motrice l'objet de la perception est une représentation motrice abstraite du phonème : le geste de constriction intentionnel obtenu grâce à un décodage du signal par un module spécialisé.

Comme le montrent Galantucci *et al.* (2006) dans leur revue de questions sur les théories motrices, l'hypothèse selon laquelle les objets de la perception de la parole sont les gestes présente l'intérêt majeur de fournir une explication unique pour un large ensemble de données, que l'on peut regrouper en trois catégories :

1. Le phénomène de coarticulation, qui implique deux choses : d'une part la réalisation d'un phonème dépend fortement au niveau acoustique des phonèmes voisins, le phonème étant du coup « mieux caractérisé » par ses propriétés articulatoires qu'acoustiques (Liberman *et al.*, 1954) ; inversement, les « traces » de l'influence d'un phonème sur ses voisins doivent être correctement réassociées aux voisins et non au phonème contaminé, ce qui pour Fowler (2006) suggère que l'auditeur utilise bien une représentation du signal acoustique dans un espace qui capture également les gestes qui en sont la cause.

2. L'apport d'information motrice par d'autres modalités (visuelle ou haptique, voir par exemple Massaro (1987) ; Fowler et Dekle (1991)) que le signal acoustique, et qui impliquerait des mécanismes d'intégration multisensorielle naturellement compatibles avec une théorie des causes motrices.

3. La rapidité des mécanismes d'imitation et de répétition (*close shadowing*) qui serait liée au fait que l'information motrice soit déjà récupérée lors de la perception (Fowler *et al.*, 2003).

1.2 Théories auditives

Les théories auditives se sont opposées aux théories motrices. Elles considèrent que dans le signal de parole, l'unité minimale de référence est de nature acoustique. Ce sont donc des

[5] La définition de geste proposée ici est tout à fait semblable (et antérieure) à celle sur laquelle repose la Phonologie Articulatoire.

[6] Le texte original est en anglais : « The objects of speech perception are the intended phonetic gestures of the speaker, represented in the brain as invariant motor commands that call for movements of the articulators through certain linguistically significant configurations. These gestural commands are the physical reality underlying the traditional phonetic [notions]. »

représentations auditives qui d'une part servent de cibles pour la production, et qui contiennent d'autre part l'information suffisante à la perception de la parole.

1.2.1 La perception de la parole à partir des invariants acoustiques

A l'inverse des théories motrices, les théories auditives défendent l'idée selon laquelle ce sont des processus auditifs généraux qui permettent d'expliquer la perception de la parole ; la parole n'est donc pas *spéciale*, et sa perception ne fait pas intervenir une récupération explicite des gestes moteurs par l'auditeur.

Pour les théories auditives (voir par exemple Diehl et Kluender (1989) ; Kuhl (2000); ou encore Diehl *et al.* (2004) pour une revue) la perception de la parole ne fait intervenir que des représentations auditives ou multisensorielles, et des processus qui ne sont pas spécifiques à la parole. Plus précisément, la perception de la parole fait appel à des processus généraux d'analyse temporelle et fréquentielle du signal acoustique qui sont similaires pour la perception de signaux de parole et de non-parole (Stevens et Klatt, 1974 ; Miller *et al.*, 1976 ; Pisoni, 1977), et chez l'humain et chez certains animaux (Kuhl *et al.*, 1975 ; Kuhl et Miller, 1978 ; Kuhl et Padden, 1983 ; Kluender *et al.*, 1987 ; Dooling *et al.*, 1995). L'ensemble de ces travaux conduit les défenseurs des théories auditives à réfuter l'hypothèse proposée par les théories motrices selon laquelle la parole serait spéciale.

Les théories auditives opposent un autre argument aux théories motrices : le principe d'équivalence motrice. L'argument repose sur l'observation selon laquelle le système moteur est redondant, c'est-à-dire qu'un même signal de parole peut correspondre à plusieurs formes du conduit vocal et à plusieurs configurations articulatoires (voir par exemple Ladefoged *et al.* (1972) ; Perkell *et al.* (1997)). Si, comme le défendent les théories motrices, l'unité phonologique est le geste intentionnel du locuteur en production aussi bien qu'en perception, comment expliquer qu'un même phonème puisse-t-être réalisé avec des configurations du conduit vocal très différentes ? (Par exemple Delattre et Freeman (1968) proposent huit classes différentes pour le /r/ américain.)

Par ailleurs, les théories auditives défendent l'idée que les relations de dépendance entre un phonème et son contexte, bien que conséquences d'actions motrices, sont encodées dans le signal acoustique, et que des mécanismes d'apprentissage perceptif suffisent à expliquer le fait que l'auditeur sache ensuite associer à la même catégorie phonétique des occurrences d'un même phonème réalisé dans des contextes différents, sans avoir besoin de recourir à quelque information motrice que ce soit.

Un tel effet de contexte est mis en évidence par Mann (1980) qui montre que des stimuli synthétiques situés sur un continuum entre /da/ et /ga/ sont plus souvent reconnus par des sujets humains comme des /ga/ quand ils sont présentés après /al/ que après /ar/. Les théories motrices expliquent cet effet perceptif comme étant une compensation de la coarticulation qui a lieu durant la production. Cependant Lotto et Kluender (1998) reproduisent le même effet en remplaçant les contextes /al/ et /ar/ par des signaux synthétiques de non parole ayant des transitions formantiques similaires. De plus, Lotto *et al.* (1997) reproduisent chez les oiseaux le même type de décalage que celui observé par Mann (1980) sur des sujets humains. Ces résultats suggèrent là encore que ces effets de contexte semblent ne pas être spécifiques à la perception de la parole, ou même au genre humain, mais qu'ils seraient au contraire une manifestation de mécanismes perceptifs plus généraux. De plus, ces deux expériences contredisent l'idée de

l'utilisation de connaissances motrices pour la perception puisque, dans le premier cas, le signal influençant la perception est très éloigné de ce que peut produire un conduit vocal, et puisque, dans le second cas, les oiseaux n'ont bien sûr pas de connaissance du conduit vocal humain.

1.2.2 La production de la parole à partir de gestes moteurs correspondant à des invariants acoustiques

Alors que pour la Phonologie Articulatoire les buts de la production de la parole sont des constrictions à réaliser dans le conduit vocal dans un certain ordre, les théories auditives de la production se donnent comme objectif une séquence de cibles dans l'espace acoustique (Perkell *et al.*, 1997 ; Guenther *et al.*, 1998 ; Perrier, 2005).

Quatre arguments principaux sont avancés par Guenther *et al.* (1998) pour défendre l'hypothèse selon laquelle la production de la parole consiste à choisir des gestes permettant de réaliser au mieux des cibles acoustiques.

1. Le premier argument est une attaque destinée aux théories qui, comme la Phonologie Articulatoire, sont basées sur des représentations internes des constrictions. Guenther *et al.* (1998) prétendent que dans le cas des voyelles ou des semi-voyelles, les retours somatosensoriels ne permettent pas de connaître directement le lieu et le degré des constrictions de manière précise. L'absence de ce type d'information rend difficile d'envisager la construction au cours de l'apprentissage de représentations internes des constrictions. En revanche, le locuteur dispose de retours proprioceptifs et auditifs des conséquences de ses commandes motrices, ce qui est suffisant pour apprendre à associer des gestes moteurs à des cibles acoustiques.

2. L'invariance du lieu et de la taille des constrictions peut émerger de systèmes basés sur l'utilisation de cibles auditives en production. La transformation articulatori-acoustique est une transformation *many-to-one* (non injective), c'est-à-dire que pour une même cible, il peut y avoir de nombreuses configurations articulatoires permettant de l'atteindre. Même si ce phénomène d'équivalence motrice rend possible une certaine variabilité dans les gestes moteurs qu'un locuteur peut produire pour atteindre une cible, garantissant ainsi une certaine flexibilité, ajouter des contraintes de relaxation ou d'effort minimal permet d'assurer l'invariance relative des degrés et lieux de constriction obtenus pour une cible donnée (Guenther *et al.*, 1998).

3. Un ensemble de travaux s'est donné comme but d'étudier comment des locuteurs pouvaient compenser des contraintes extérieures, que ce soient des perturbations appliquées à la mandibule (Lindblom *et al.*, 1977), ou aux lèvres (Abbs et Gracco, 1984 ; Perkell *et al.*, 1993, 1994 ; Savariaux *et al.*, 1995, 1999). Ces expériences, partant du principe que les sujets vont essayer de s'approcher le mieux possible de leur cible, cherchent à savoir si la cible se caractérise en terme de constrictions ou en terme de buts auditifs, et tranchent la question en faveur de la seconde possibilité. Par exemple, dans l'expérience réalisée par Savariaux *et al.* (1999), les sujets, contraints par un tube rigide placé entre leurs lèvres, ne peuvent pas effectuer le geste d'arrondissement des lèvres caractéristique du /u/. Ils ont alors tendance à complètement réorganiser leur conduit vocal, changeant donc les

constrictions, ce qui confirme l'idée selon laquelle la production de la parole se fait en sélectionnant des gestes permettant d'atteindre des cibles acoustiques.

(4) Le principe d'équivalence motrice permet également au locuteur de faire des choix en situation non contrainte. C'est le cas par exemple pour le /r/ américain, qui a beaucoup été étudié (par exemple Delattre et Freeman (1968) ; Lindau (1980) ; Espy-Wilson et Boyce (1994) ; Hagiwara (1994) ; Ong et Stone (1998) ; Westbury *et al.* (1998)). Les chercheurs proposent de distinguer plusieurs classes de production du /r/, même si ils n'en proposent pas le même nombre : certains en voient deux (Espy-Wilson et Boyce, 1994 ; Ong et Stone, 1998), d'autres trois (Hagiwara, 1994), et d'autres préfèrent parler d'un continuum (Delattre et Freeman, 1968 ; Westbury *et al.*, 1998) qui sépare deux positions extrêmes fortement opposées. D'après ces travaux, en fonction du contexte, un même locuteur peut choisir différentes manières de produire le /r/, avec des constrictions très différentes. C'est un argument de plus en faveur de l'idée que le locuteur essaye d'atteindre une cible acoustique de /r/, et que pour ce faire il choisit la configuration la moins coûteuse à réaliser au vu du contexte.

2 Intégration perceptuo-motrice

Bien qu'elles apportent des cadres théoriques qui ont bien fait progresser la recherche en parole, il semble que ni les théories motrices ni les théories auditives ne suffisent à expliquer à elles seules tous les phénomènes qui interviennent. Ainsi, d'un côté les théories motrices ont du mal à fournir une explication au phénomène d'équivalence motrice, et de l'autre les théories auditives n'expliquent pas clairement la variabilité dans la réalisation des phonèmes due à la coarticulation. Puisqu'il semble que considérer une unité minimale de nature purement motrice ou de nature purement perceptive ne suffise pas à décrire la complexité de la parole, les théories perceptuo-motrices proposent que l'unité de base soit de nature plus complexe, et intègrent pour ce faire des aspects sensoriels et des aspects moteurs.

2.1 Intégration théorique dans des modèles cognitifs conceptuels

Les théories perceptuo-motrices proposent d'étudier la production et la perception de la parole en faisant intervenir explicitement un lien entre les représentations motrices et les représentations perceptives, et en insistant sur le rôle clé de ce lien sensori-moteur dans la co-construction des représentations motrices et auditives.

2.1.1 Des modèles perceptuo-moteurs de la production de la parole

Les approches perceptuo-motrices de la production de la parole sont basées sur la notion de boucle rétroactive de contrôle introduite en parole par Levelt avec la *Perceptual Loop Theory* (Levelt, 1983, 1992, 1993 ; Levelt *et al.*, 1999).

C'est le cas par exemple du modèle *DIVA (Directions Into Velocities of Articulators)* (Guenther, 2006) qui propose un modèle cognitif du contrôle moteur pour la production de la parole

et une implémentation informatique de cette architecture, à base de réseaux de neurones artificiels.[7] Ce modèle fait intervenir deux sous-systèmes qui agissent de concert :

un sous-système de contrôle direct (*feedforward*) qui associe les sons à produire (mots, phonèmes ou syllabes) aux commandes motrices (position et vitesse des articulateurs) ;

un sous-système de contrôle par rétroaction (*feedback*) qui permet de comparer les conséquences auditives et somatosensorielles des commandes motrices aux buts désirés et d'adapter dynamiquement ces commandes motrices pour pouvoir atteindre les cibles.

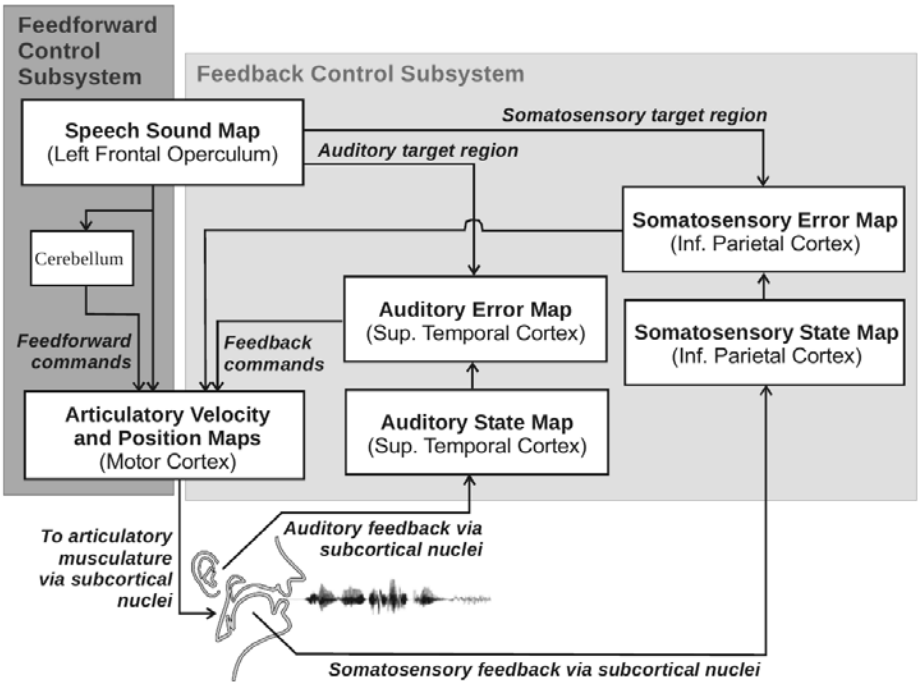


Figure 2.4: **L'architecture du modèle *DIVA***. Les rectangles représentent les cartes neuronales, et les flèches les connexions synaptiques qui les relient. Cette figure est tirée de Guenther (2006).

La figure 2.4 représente l'architecture du modèle *DIVA* (Guenther, 2006). La carte des sons (*Speech Sound Map*) contient des unités de paroles (mots, sons, phonèmes). Lorsque l'une de ces unités est activée, le système de contrôle feedforward envoie des consignes motrices (*feedforward commands*) aux articulateurs. Ces commandes sont ensuite modifiées dynamiquement grâce au système de contrôle feedback qui compare (au sein des *Error Maps*) les retours auditifs et somatosensoriels (*auditory feedback* et *somatosensory feedback*) aux cibles (*target regions*) auditives et somatosensorielles correspondant à l'unité de parole à produire.

Ce modèle fait intervenir trois phases successives d'apprentissage. Dans la première phase, que Guenther juge similaire à du babillage, des activations semi-aléatoires des articulateurs permettent d'apprendre les liens entre les représentations motrices, auditives et somatosensorielles,

[7] À noter que nous avons pris successivement deux versions du modèle de Guenther comme illustrations respectivement des approches auditives et perceptuo-motrices, ce qui correspond bien à l'évolution des prises de position, la version de 1998 centrée sur les représentations auditives se trouvant complétée en 2006 par une véritable architecture perceptuo-motrice.

qui sont stockés dans les cartes d'erreur (*Error Maps*). Dans la seconde phase, tel un enfant exposé aux sons de sa langue, le modèle apprend le lien entre la carte des sons et les régions cibles dans l'espace acoustique. Dans la troisième étape, le modèle essaye de produire les gestes moteurs correspondant aux cibles acoustiques. Lors des premiers essais, le système de contrôle direct est assez mauvais, et l'exécution du bon geste moteur dépend grandement du mécanisme de feedback. Par la suite, chaque nouvelle tentative contribue à améliorer le système de contrôle direct, qui petit à petit apprend le bon mapping jusqu'à ne plus avoir besoin de feedback, si ce n'est en cas de perturbations.

Ainsi, en combinant un feedback perceptif, essentiel durant l'apprentissage, avec un système feedforward de production motrice, le modèle *DIVA* permet de rendre compte de plusieurs propriétés caractéristiques de la production de parole, comme les principes d'économie articulatoire, de coarticulation, d'équivalence motrice ou les mécanismes sensori-moteurs de régulation, et de proposer des étapes pour l'acquisition des représentations phonémiques.

Perrier (2005) défend également l'hypothèse selon laquelle le lien sensori-moteur joue un rôle clé lors de l'apprentissage, ou en conditions perturbées. Mais, de plus, Perrier (2005) insiste sur le fait qu'après apprentissage, et en conditions normales, les différents niveaux de représentation (moteur, orosensoriel et acoustique) sont équivalents. « Ainsi, après apprentissage de la parole, les représentations de la tâche de production pourraient regrouper des composantes dans le domaine du contrôle moteur, dans le domaine orosensoriel et dans le domaine acoustique, dont l'importance augmente quand on va du moteur vers l'acoustique. Dans des conditions normales de production de la parole, ces trois niveaux de représentation sont équivalents. La planification et le contrôle de la production de la parole pourraient ainsi être réalisés dans chacun de ces domaines, ou dans un domaine hybride basé sur une combinaison complexe, éventuellement dépendante des phonèmes, de ces trois niveaux. Cependant, lorsque les conditions de parole sont perturbées, de telle sorte que les objectifs de différentes natures ne puissent pas tous être remplis, la priorité est donnée à la satisfaction des objectifs acoustiques »[8] (Perrier, 2005, page 127).

2.1.2 Des modèles perceptuo-moteurs de la perception de la parole

Les approches perceptuo-motrices de la perception de la parole sont basées sur l'idée que les connaissances motrices sont sollicitées et sont utiles en perception. À notre connaissance, le premier modèle qui utilisait conjointement des connaissances auditives et des connaissances motrices pour des tâches de perception a été proposé au *MIT* au début des années 60 par Stevens (1960). Ce modèle procède par génération d'hypothèse et test : il contient une branche de perception auditive pour la partie *feedforward* (génération d'hypothèses) et un modèle de production motrice pour la partie *feedback* (test de l'hypothèse). Il s'agit là d'un mécanisme plus général connu sous le nom d'*analyse par la synthèse* (Halle et Stevens, 1959 ; Stevens, 1960 ; Halle et Stevens, 1962 ; Stevens et Halle, 1967) par lequel la récupération des gestes moteurs

[8] « Thus, after speech learning, the representation of the task could consist of components in the motor control domain, in the orosensory domain and in the acoustic domain, with an increasing importance from the motor control component to the acoustic one. In normal speech production, these three levels of representation are equivalent. Planning and monitoring of speech production could thus be made in either of these domains, or in a hybrid domain based on complex, possibly phoneme-dependent combinations of the three components. However, when perturbations modify the speech conditions, so that the goals of different nature cannot be matched simultaneously, priority will be given to the achievement of the acoustic goals. »

susceptibles d'avoir permis de produire le signal de parole donné en entrée influence la manière dont ce signal est perçu.

Voici la manière dont Skipper *et al.* (2007) inscrivent ce mécanisme d'analyse par la synthèse au cœur de leur modèle de la perception audiovisuelle. Dans un premier temps, l'auditeur extrait du signal de parole et des mouvements du visage de son interlocuteur des représentations multisensorielles, qui à ce stade ne sont que des hypothèses (et pas encore des interprétations) sur les phonèmes produits par le locuteur. À ces hypothèses de phonèmes sont ensuite associées les commandes motrices qui permettraient de les produire. Les commandes motrices ainsi activées permettent alors de prédire les conséquences acoustiques et somatosensorielles qui en résultent (voir la notion de modèle interne (Jordan et Rumelhart, 1992 ; Miall, 2003 ; Callan *et al.*, 2004), qui est liée aux notions de copie d'efférence (Von Holst, 1954) et de décharge corollaire (Sperry, 1950)). Finalement, ces conséquences perceptives prédites viennent contraindre l'interprétation phonémique du signal de parole en venant renforcer ou réduire la plausibilité des hypothèses faites en amont. De plus, Skipper *et al.* (2007) proposent une description des différentes aires corticales intervenant dans leur modèle sensori-moteur de la perception de parole, en s'appuyant sur des données d'imagerie dont on pourra trouver une revue dans Skipper *et al.* (2006). La boucle de simulation/prédiction perceptuo-motrice du modèle de Skipper *et al.* (2007) est présentée de manière schématique par la figure 2.5.

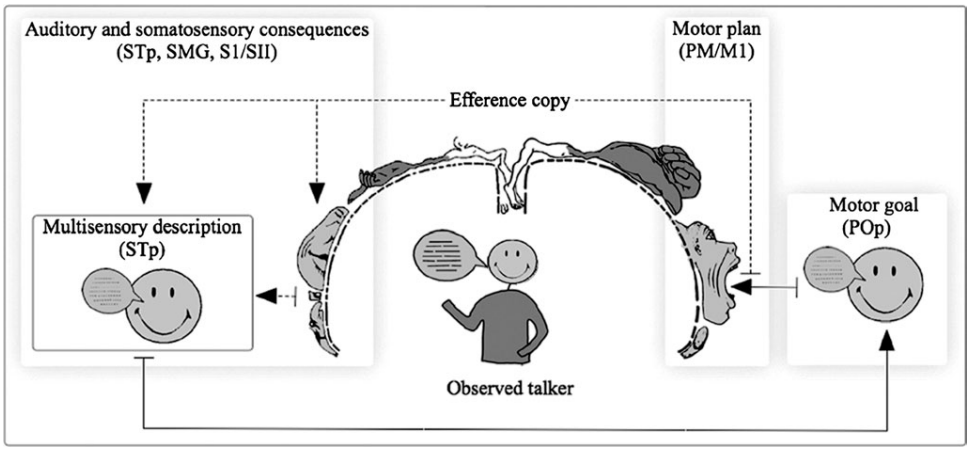


Figure 2.5: **Le modèle de Skipper** (d'après Skipper *et al.* (2007)) décrit une boucle de simulation/prédiction perceptuo-motrice, dans laquelle les hypothèses générées par les entrées sensorielles au niveau des aires visuelles, des aires auditives temporales supérieures (STp), du cortex somatosensoriel (SI/SII) et du gyrus supramarginal (SMG), qui sont résumées en une description multisensorielle au niveau du cortex temporal supérieur postérieur (STp), génèrent des hypothèses motrices dans le cortex prémoteur frontal (*pars opercularis* POp) se traduisant par des plans moteurs dans le cortex prémoteur (PM) et moteur primaire (M1) puis, par copie d'efférence, par des hypothèses multisensorielles qui viennent raffiner les descriptions multisensorielles initiales.

Ces approches basées sur l'analyse par la synthèse défendent l'idée selon laquelle la perception de la parole fait intervenir à la fois des représentations sensorielles ou multisensorielles, et des représentations motrices. La Théorie de la Perception pour le Contrôle de l'Action (*Perception for Action Control Theory (PACT)*, voir Schwartz *et al.* (2002a, 2007, 2012a)) va plus loin en proposant que la perception et l'action se co-structurent au cours de l'apprentissage

de la parole, qui passe à la fois par la perception et la production de sons de parole. Dans ce cadre, les unités de la parole peuvent être vues soit comme des « percepts multimodaux régularisés par des contraintes motrices », soit comme des « gestes moteurs dont la mise en forme est déterminée par des traitements multimodaux » (Schwartz *et al.*, 2002a). La figure 2.6 montre de manière schématique le liage qui intervient lors de la perception en ligne entre des représentations motrices et sensorielles qui se sont co-construites lors de l'apprentissage.

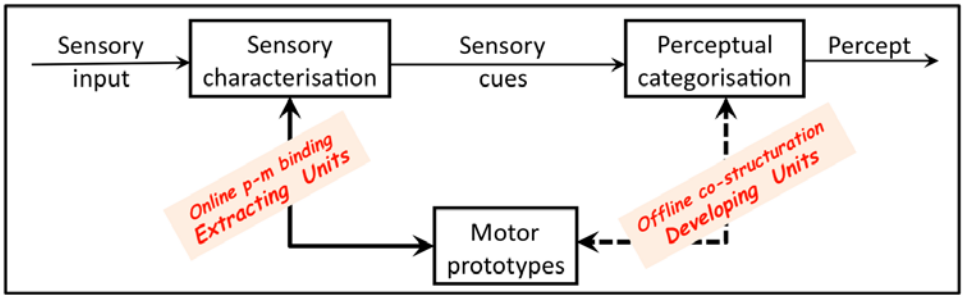


Figure 2.6: **La Théorie de la Perception pour le Contrôle de l'Action** (d'après Schwartz *et al.* (2012a)). La voie sensorielle assurant la caractérisation sensorielle (par extraction d'indices) et la catégorisation perceptive, est complétée par une voie motrice qui permet de raffiner l'extraction des unités (« online p-m binding » : le liage perceptuo-moteur mis en jeu en permanence dans les processus de traitement perceptif). Les prototypes moteurs sont co-construits avec les prototypes perceptifs au cours du développement (« offline co-structuration »).

Un des arguments sur lesquels s'appuie la *PACT* pour défendre la nature perceptuo-motrice des unités de parole vient de l'analyse de la genèse des systèmes phonologiques (Schwartz *et al.*, 2007). À partir de l'observation selon laquelle les voyelles présentes dans les langues du monde sont bien séparées dans l'espace acoustique, la théorie de la dispersion propose dans sa version initiale (Liljencrants et Lindblom, 1972) que les voyelles des langues du monde se répartissent dans l'espace acoustique de manière à maximiser le contraste perceptif. Plus tard, cette théorie de la dispersion est modifiée : il ne suffit pas seulement d'avoir de grands écarts dans l'espace acoustique, mais il faut également essayer de minimiser le coût articulatoire (Lindblom, 1986, 1990b). Pour le dire autrement, l'émergence des systèmes phonologiques n'est pas conditionnée par des contraintes perceptives uniquement, mais également par des contraintes motrices. Par ailleurs, la théorie quantique de la parole (Stevens, 1972) fournit un autre argument qui va dans ce sens en défendant que les non-linéarités de la transformation articulatori-acoustique fournissent des frontières perceptives qui participent également à la structuration des systèmes phonologiques. Ainsi, un geste ne peut, dans le contexte de la *PACT*, être défini sans référence à ses propriétés perceptives, et potentiellement contrastives : le geste est une unité perceptuo-motrice.

2.1.3 Taxonomie des différentes théories et modèles

Le Tableau 3.6 présente une classification des principaux modèles et théories de la littérature en fonction de la tâche à laquelle ils s'intéressent (production ou perception) et du type de représentations sur lesquelles ils reposent (motrices ou auditives).

	Tâche	
Théorie	Production	Perception
Motrice	Articulatory Phonology, Browman et Goldstein (1989)	Motor Theory, Liberman et Mattingly (1985)
Auditive	Auditory reference frames for speech planning, Guenther *et al.* (1998)	Auditory theories, Diehl *et al.* (2004)
Perceptuo-motrice	DIVA model, Guenther (2006)	Perception for Action Control Theory, Schwartz *et al.* (2012a)

Tableau 2.1: **Taxonomie des modèles de production et de perception**, adaptée de Moulin-Frier *et al.* (2010).

Nous ne citons dans cette table que certains travaux emblématiques ou représentatifs de chaque configuration possible, sans souci d'exhaustivité, et en renvoyant aux sections précédentes pour plus d'exemples. Ce qui est essentiel pour nous est la possibilité d'organiser ainsi les différentes propositions en trois grands courants théoriques (théories motrices, auditives/sensorielles et perceptuo-motrices), déclinés à la fois du point de vue de la production et de la perception. Cette taxonomie, à notre sens très éclairante, nous servira de guide tout au long de cette thèse. Elle est reprise de la thèse de Moulin-Frier (2011) qui l'a utilisée dans le cadre de ses recherches sur l'émergence des systèmes sonores des langues du monde dans des sociétés d'agents communicatifs en interaction.

2.2 Données expérimentales en faveur des théories intégratives

Les débats théoriques entre les théories perceptives et motrices de la communication parlée ont fait rage dans les années 70 et 80, suscitant la mise au point d'un riche ensemble de nouveaux paradigmes expérimentaux, ainsi que le recueil de données abondantes.

Dans le domaine de la production de la parole, que nous détaillons peu car il n'est pas au cœur de cette thèse, les principaux arguments en faveur de l'une et de l'autre hypothèse (cibles articulatoires ou cibles auditives) ont été présentés dans la section 1. Dans le domaine de la perception de la parole, le débat a été nourri par de nombreux travaux expérimentaux, qui ont conduit à l'émergence ou au développement de nouveaux paradigmes portant sur la perception catégorielle (Repp, 1984), les relations de compensation (*trading relations* ; Repp (1983)), les effets d'adaptation perceptuo-moteurs (Cooper, 1979), la répétition rapide (*close shadowing* ; Porter et Lubker (1980)), l'intégration audiovisuelle et multisensorielle (Dodd et Campbell, 1987 ; Campbell *et al.*, 1998), l'effet duplex (Liberman *et al.*, 1981), etc... Néanmoins, force est de constater qu'à chaque argument a succédé un contre-argument et que le débat a finalement semblé s'enliser quelque peu – et lasser la communauté scientifique – jusqu'au renouveau d'intérêt produit par les avancées des neurosciences cognitives dans les années 90.

Les théories intégratives s'inscrivent en effet dans un courant récent des neurosciences de la parole qui ont montré de manière récurrente et cohérente à la fois l'implication dans des tâches de production de parole d'aires corticales du lobe temporal, classiquement associées à la perception de la parole ; et réciproquement l'implication dans des tâches de perception de parole d'aires corticales des lobes pariétal et frontal, classiquement associées à la production de

la parole.

L'implication des aires perceptives en production de parole, dont on trouvera une revue précise dans la thèse de Grabski (2012), est mise en évidence à la fois par l'existence d'une modulation des réponses du cortex auditif en production ouverte de parole par rapport aux réponses observées dans l'écoute passive des stimuli correspondants, et par l'existence de réponses auditives en production de parole silencieuse.

À l'inverse, depuis la mise en évidence des neurones miroirs chez le singe (Rizzolatti *et al.*, 1996a) puis d'un système miroir chez l'humain (Rizzolatti *et al.*, 1996b), l'implication des aires pariétales et frontales sensori-motrices ou prémotrices en perception de parole a été largement démontrée dans les 15 dernières années. Ainsi, de nombreux résultats expérimentaux montrent que le système moteur est activé pendant des tâches de perception. Par exemple, des études utilisant la stimulation magnétique transcranienne (TMS) ont montré que l'activité des muscles de la langue augmente lors de l'écoute de phrases comprenant des consonnes linguales (Fadiga *et al.*, 2002), et que l'activité des muscles des lèvres augmente lorsque les sujets écoutent ou visionnent des séquences comprenant des consonnes labiales (Watkins *et al.*, 2003). Des études d'imagerie fonctionnelle par résonance magnétique (*fMRI*) ou par magnétoencéphalographie (*MEG*) ont montré l'existence d'un recouvrement entre les aires corticales actives lors de la production de la parole, et celles qui s'activent lors d'une écoute passive (Wilson *et al.*, 2004 ; Pulvermüller *et al.*, 2006 ; Grabski *et al.*, 2013) ou active, avec la mise en évidence de corrélations entre réponses prémotrices précoces et acuité de la décision phonétique (Alho *et al.*, 2012, 2014).

Ainsi, toutes ces données convergent vers l'hypothèse de l'existence d'une « voie dorsale », associant aires temporales perceptives et aires frontales motrices à travers les aires pariétales de convergence sensori-motrice (Hickok et Poeppel, 2000, 2004). La question posée est alors celle du rôle fonctionnel de ces corrélations neuroanatomiques : s'il apparaît maintenant tout à fait incontestable que la perception de la parole engage l'activation de régions pariéto-frontales associées à la production de la parole, cette activation est-elle le simple reflet de processus d'association perceptuo-motrice nécessaires au développement de la parole, comme l'ont proposé Hickok et Poeppel de façon récurrente (Hickok et Poeppel, 2007), ou joue-t-elle au contraire un rôle fonctionnel nécessaire au bon fonctionnement des mécanismes perceptifs (Pulvermüller et Fadiga, 2010) ?

Les données expérimentales ont apporté peu à peu des éléments en faveur de la seconde hypothèse. Il apparaît ainsi que des paradigmes visant à modifier temporairement le fonctionnement du système moteur, soit par application de stimulation magnétique transcranienne répétitive (rTMS) appliquée pendant quelques minutes sur le cortex prémoteur ventral ou sur le cortex moteur primaire, soit par mise en œuvre de processus de fatigue motrice (adaptation sélective), pouvaient produire une modification des processus de décision perceptive dans des tâches de catégorisation phonétique (voir Meister *et al.* (2007) ; D'Ausilio *et al.* (2009) ; Sato *et al.* (2009, 2011) ; Möttönen et Watkins (2009) ; Möttönen *et al.* (2013).

Toutefois, les effets sont quantitativement faibles, et généralement confinés à des tâches de perception difficile, sur des stimuli ambigus ou en présence de bruit (D'Ausilio *et al.*, 2012). Ainsi, une étude récente par imagerie fonctionnelle fMRI (Du *et al.*, 2014) montre que, si les aires temporales auditives semblent présenter dans une analyse multivariée de bonnes capacités de discrimination entre phonèmes présentés sans bruit, en présence de bruit les aires frontales (cortex prémoteur ventral et aire de Broca) semblent plus discriminatives, avec une activité

augmentant avec le niveau de bruit et la difficulté de catégorisation. Peu à peu émerge ainsi une hypothèse générale selon laquelle les aires perceptives et motrices joueraient un rôle complémentaire dans la perception de la parole, avec un rôle accru des secondes en situation adverse, notamment dans le bruit acoustique.

2.3 Des outils conceptuels pour réfléchir à l'intégration perceptuo-motrice

2.3.1 Modèles internes

D'après Perrier (2012), le concept de modèle interne semble avoir été introduit par Francis et Wonham (1976) sous le nom de *Internal Model Principle* dans le cadre de la théorie du contrôle. Le principe en est que pour assurer un contrôle robuste, le contrôleur a besoin d'avoir un modèle des réactions du système aux perturbations et aux changements dans les paramètres de contrôle. Le concept de modèle interne a par la suite été mis au coeur de leurs développements théoriques par Kawato *et al.* (1987), puis il a pénétré le domaine du contrôle moteur des gestes de parole (voir une revue dans Perrier (2012)).

Ces modèles internes permettent de prévoir les conséquences perceptives des commandes motrices sans avoir besoin de réaliser l'action correspondante. La figure 2.7 illustre le fonctionnement des modèles internes du point de vue cognitif.

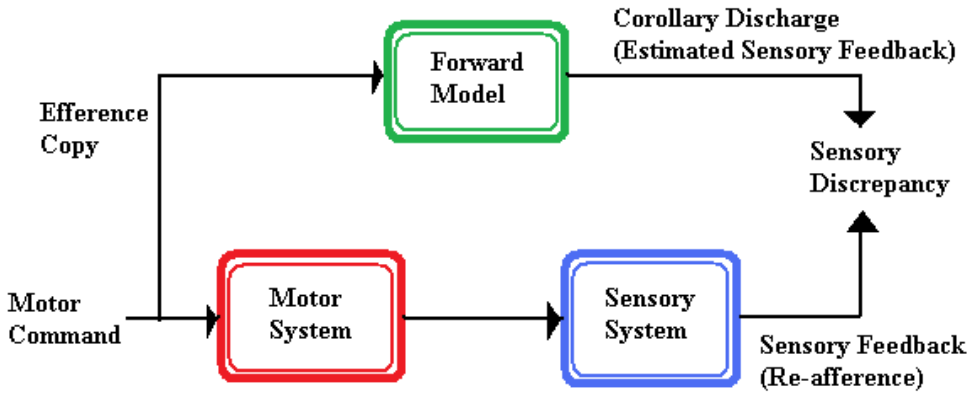


Figure 2.7: **Les modèles internes pour le contrôle moteur.** Lorsqu'une commande motrice (*Motor Command*) est envoyée au système moteur, une copie de cette commande (*Efference Copy*) est envoyée au modèle interne (*Forward Model*). À partir de la copie d'efférence, le modèle interne génère une prédiction du feedback sensoriel estimé (*Corollary Discharge*). Parallèlement, le système moteur (*Motor System*) a réalisé l'action correspondant à la commande motrice, et cette action a des conséquences à la suite desquelles le système perceptif (*Sensory System*) produit un signal de feedback sensoriel (*Re-afference*) qui est comparé au feedback sensoriel estimé (*Corollary Discharge*) par le modèle interne.[9]

Un modèle interne permet ainsi, à partir (d'une copie d'efférence) d'une commande motrice, de prévoir les conséquences perceptives de l'action correspondante. On considère en général deux types de modèles internes : les modèles directs, permettant d'associer une possible conséquence sensorielle à une cause motrice, et les modèles inverses permettant d'inférer en retour une action

[9]Cette figure est tirée de `http://en.wikipedia.org/wiki/Efference_copy`.

motrice susceptible d'avoir généré une stimulation sensorielle donnée. Les modèles internes sont discutés en détail dans un cadre cognitif général par Grush (2004) (ainsi que dans les commentaires). Ici, nous nous contentons de proposer quelques interprétations de ce que peut apporter la notion de modèle interne dans le cadre de la parole.

En production de la parole

Ainsi que le démontre clairement la revue de questions présentée par Perrier (2012), les modèles internes en production de la parole jouent un rôle crucial pour contrôler un système tel que le système articulatoire qui ne peut se permettre d'attendre le retour de ses *feedbacks* sensoriels pour déterminer la séquence de ses actions. Les modèles internes permettent ainsi de simuler les propriétés cinématiques et dynamiques du système de production – préalablement apprises par divers processus d'entraînement – pour mettre en œuvre des commandes adaptées à la réalisation de séquences complexes. Les retours sensoriels peuvent en parallèle renseigner le système sur la présence de perturbations ou d'erreurs de production lui permettant le cas échéant de modifier sa programmation ultérieure, voire de modifier dynamiquement les modèles internes eux-mêmes.

En perception de la parole

La théorie motrice de la perception fait appel de manière implicite à cette notion de modèle interne, dans la mesure où elle suppose que l'auditeur dispose d'un processus permettant, à partir du signal acoustique, de retrouver les gestes articulatoires qui l'ont produit. Liberman *et al.* (1967) proposent ainsi dès les années 60 l'hypothèse selon laquelle le module de décodage de la parole opère « en quelque sorte en effectuant le processus [de production de la parole] à l'envers »[10] (Liberman *et al.*, 1967, page 454).

Il s'agit en l'occurrence de ce que l'on appelle, nous l'avons vu, un modèle inverse, par opposition à un modèle direct qui permettrait d'obtenir directement le signal acoustique correspondant à chaque geste articulatoire, comme c'est le cas par exemple dans le modèle proposé par Stevens (1960) qui fait de l'*analyse par la synthèse*.

Plus tard, dans la version révisée de la théorie motrice, Liberman et Mattingly semblent préférer l'idée d'un modèle direct, qu'ils décrivent d'une manière assez proche de l'idée d'*émulateur* proposée par Grush (2004) : « Les descriptions potentielles du signal sont calculées par un processus analogue à celui utilisé pour la production – un synthétiseur interne de conduit vocal, inné [...] – qui dispose de toutes les informations sur les caractéristiques anatomiques et physiologiques du conduit vocal, et sur les conséquences articulatoires et acoustiques des gestes de parole significatifs du point de vue linguistique »[11] (Liberman et Mattingly, 1985, page 26).

Finalement, le mécanisme d'analyse par la synthèse, qui repose sur la notion de modèle interne, suscite depuis quelques années un intérêt renouvelé (Poeppel *et al.*, 2008 ; Poeppel et Monahan, 2011 ; Bever et Poeppel, 2010). Ainsi, Bever et Poeppel (2010) proposent

[10] "somehow running the process [of speech production] backward"

[11] « [T]he candidate signal descriptions are computed by an analogue of the production process – an internal, innately specified vocal-tract synthesizer [...] – that incorporates complete information about the anatomical and physiological characteristics of the vocal tract and also about the articulatory and acoustic consequences of linguistically significant gestures. »

une revue récente de ce qu'une telle approche est susceptible d'apporter à la recherche sur la parole et le langage (et sur la vision).

2.3.2 Les conditions de l'émergence de la communication

La question de l'émergence de la communication dans l'évolution, jusqu'à l'apparition et la stabilisation du langage humain, est d'une complexité considérable et très au-delà des enjeux de la présente thèse. On en trouvera quelques éléments de présentation dans une perspective historique et épistémologique chez Boë *et al.* (2011).

Dans une perspective computationnelle beaucoup plus proche de nos propres travaux, Moulin-Frier (2011) a proposé un cadre de modélisation s'appuyant sur des mécanismes précurseurs de la communication, en nombre réduit. Nous voulons ici en mentionner deux, qui joueront un rôle dans la suite de notre travail.

Les théories de l'esprit

Les théories motrices font l'hypothèse que l'auditeur a accès aux processus de production motrice de son interlocuteur, voire à ses intentions motrices (Liberman et Mattingly, 1985), et donc qu'il est capable de simuler une partie de ses mécanismes cognitifs. Cette hypothèse peut être intégrée dans le cadre d'une théorie de la simulation plus générale, la « théorie de l'esprit » (Baron-Cohen *et al.*, 1985). Dans ce cadre conceptuel très large et d'une importance théorique considérable, il est fait l'hypothèse que chacun dispose d'un ensemble de connaissances et d'hypothèses sur les processus cognitifs en oeuvre chez ses congénères, et peut ainsi, en supposant ces processus actifs dans la communication, inférer chez ses partenaires de communication des états mentaux, interpréter leurs actions et prédire leurs actions futures. La notion de modèle interne a d'ailleurs été progressivement insérée dans ce cadre pour fournir des bases computationnelles et prédictives fortes (voir par exemple Wolpert *et al.* (2003)).

Curieusement, le lien entre les débats sur les processus perceptuo-moteurs dans la communication parlée et les développements de la théorie de l'esprit est rarement fait. Il semble pourtant que la théorie de l'esprit soit une composante potentielle majeure dans le chemin vers l'émergence du langage humain (Tomasello *et al.*, 2005). Les neurones miroirs et le système miroir chez l'humain ont d'ailleurs été proposés comme un mécanisme central dans l'interprétation des intentions (Gallese et Goldman, 1998) et dans l'émergence du langage (Rizzolatti et Arbib, 1998).

La deixis

La deixis désigne la capacité de partager une référence commune en montrant des choses du geste ou du regard. Là encore, nous ne nous engagerons pas dans une revue de questions sur ce mécanisme qui a fait l'objet d'un nombre considérable d'études, à la fois dans le contexte du développement langagier chez l'enfant, et de la communication chez les primates non humains (voir une revue chez Ducey-Kaufmann (2007)). Mais il sera important pour la suite de souligner que la deixis a été mentionnée par Abry *et al.* (2009) comme un possible précurseur du langage, et qu'elle fournit en tout état de cause un mécanisme précurseur plausible pour traiter du problème de la référence, c'est-à-dire de la mise en cohérence d'un *signifiant* acoustico-articulatoire avec un *signifié* qui est un objet du monde

extérieur (dans un sens très général) auquel la matière sonore doit référer. La deixis est utilisée par Moulin-Frier (2011) comme un processus clé pour assurer la communication entre deux agents sensori-moteurs en interaction : nous lui ferons jouer par la suite un rôle équivalent.

3 Des modèles computationnels perceptuo-moteurs pour la reconnaissance et la perception de la parole

La présentation des théories auditives et motrices de la parole que nous avons faite ayant abouti à la nécessité d'approches intégratives, nous nous intéressons maintenant à ce que la modélisation computationnelle peut apporter dans ce cadre. Nous allons pour ce faire passer en revue les différents modèles computationnels proposés dans la littérature, qui allient des connaissances motrices et des connaissances acoustiques pour simuler les mécanismes de la perception ou tenter d'améliorer les performances des systèmes de reconnaissance automatique de la parole.

De tels travaux computationnels intégrant informations motrices et informations acoustiques s'inscrivent en effet dans deux types d'approches. D'un côté l'on trouve des modèles qui visent la performance dans le domaine de la Reconnaissance Automatique de la Parole (*Automatic Speech Recognition, ASR*), et de l'autre on trouve dans le domaine de la modélisation cognitive des modèles qui tentent de fournir une description plausible de l'organisation des représentations et processus cognitifs intervenant dans la perception de la parole.

3.1 Travaux en Reconnaissance Automatique de la Parole

Les sytèmes de Reconnaissance Automatique de la Parole sont en général composés de deux parties complémentaires : un modèle acoustique et un modèle de langage. Le modèle acoustique calcule à partir d'une description – typiquement spectrale – du signal d'entrée une séquence d'unités – typiquement des phonèmes – plausibles, et le modèle de langage utilise des connaissances contextuelles spécifiques décrivant la probabilité d'apparition de séquences de phonèmes (ou de mots) dans un contexte donné pour déterminer le choix de la séquence finale – typiquement textuelle – qui constitue la sortie du système.

Ainsi, les modèles de langage guident la transcription du signal acoustique en phonèmes ou autres unités phonologiques en utilisant les régularités statistiques du langage. Leur étude sort du cadre de cette thèse. En revanche, la présente section explore les apports potentiels de connaissances motrices issues de la production de la parole aux modèles acoustiques. Nous verrons que les processus d'évaluation de ces systèmes de reconnaissance automatique de la parole incluant des connaissances articulatoires allient évaluation quantitative des taux de performance dans divers types de situation, et simple « preuve de concept » de la faisabilité de la démarche, qui est effectivement relativement complexe.

3.1.1 Les principes de base des systèmes de reconnaissance automatique

Avant d'envisager comment ces systèmes pourraient intégrer des connaissances motrices, décrivons d'abord la structure des modèles acoustiques traditionnels.

Tout d'abord, un modèle acoustique prend en entrée une image compacte du signal de parole. Il s'agit en général d'un résumé sous forme de vecteurs acoustiques des principales

caractéristiques spectrales du signal au cours du temps. Typiquement, chaque vecteur résume une tranche de 20 à 30 millisecondes du signal de parole, qui est échantillonné avec un pas de 10 millisecondes. Cela revient à supposer le signal de parole stationnaire sur cette période. Les coefficients de ces vecteurs acoustiques sont obtenus par des techniques d'analyse fréquentielle. Le signal de parole est ainsi classiquement résumé par les MFCC (*Mel-Frequency Cepstral Coefficients*, voir Zheng *et al.* (2001)) ou les LPCC (*Linear Prediction Cepstral Coefficients*, voir par exemple Davis et Mermelstein (1980) pour une comparaison des deux), et éventuellement leurs dérivées première et seconde.

Si les MFCC et les LPCC sont les caractéristiques d'entrée les plus classiques et les plus couramment utilisées, de nombreux travaux visent à en trouver de meilleures. On peut citer notamment les publications proposant des principes de paramétrisation inspirés des traitements auditifs (des recherches pionnières d'Hermansky (1990) ; Hermansky et Morgan (1994) sur le système « RASTA-PLP » aux recherches plus récentes sur l'utilisation de modèles auditifs comme étape de traitement initiale : voir une revue récente dans Stern et Morgan (2012)).

Le problème de la reconnaissance automatique de la parole consiste alors à inférer, à partir de l'évolution temporelle des caractéristiques spectrales du signal, la succession de phonèmes la plus probable. Pour ce faire, la plupart des systèmes de reconnaissance de la parole utilisent des modèles de Markov cachés (HMM, *Hidden Markov Models* ; voir Baker *et al.* (2009) pour une revue historique de l'utilisation des modèles de Markov en reconnaissance automatique de la parole, et Rabiner (1989) pour un tutoriel) pour rendre compte de la variabilité temporelle du signal de parole. Chaque réalisation de chaque phonème étant décrite par une séquence d'états, c'est-à-dire une séquence temporelle des vecteurs de coefficients acoustiques mentionnés plus haut, un HMM est alors un résumé statistique des différentes réalisations des occurrences d'un phonème dans un corpus d'apprentissage. À chaque état du modèle de Markov est donc associée une distribution de probabilité modélisant la génération de vecteurs acoustiques *via* cet état. Cette distribution de probabilité peut être représentée de deux manières : pendant longtemps ce sont des modèles de mixtures de gaussiennes (GMM, *Gaussian Mixture Model* ; voir Juang *et al.* (1986)) qui ont été utilisés, mais plus récemment il semble que des réseaux de neurones artificiels (ANN, *Artificial Neuron Networks* ; Hinton *et al.* (2012) ; Graves *et al.* (2013)) permettent aux systèmes de reconnaissance qui les utilisent d'obtenir de meilleures performances.

Une des raisons contribuant à expliquer la popularité de ces approches réside sans doute dans l'efficacité des algorithmes sur lesquels elles reposent (Rabiner, 1989) : une variante de l'algorithme *Expectation-Maximization* (Dempster *et al.*, 1977) rend l'implémentation de l'apprentissage des paramètres très simple, et l'algorithme de Viterbi (Viterbi, 1967) permet de calculer à partir du signal d'entrée les séquences de phonèmes les plus probables selon le HMM appris.

3.1.2 Limitations principales des systèmes actuels et apport potentiel des connaissances motrices

Les performances des systèmes de reconnaissance automatique de la parole se sont considérablement améliorées au cours des vingt dernières années, le consensus parmi les spécialistes du domaine portant sur le fait que les avancées proviennent moins de bouleversements théoriques que de la mise en oeuvre d'outils d'apprentissage massif sur des bases de données énormes, et exploitant au mieux les progrès systématiques des performances des calculateurs disponibles.

Néanmoins, les systèmes butent encore sur des limites sévères de performance, essentiellement dès que les matériaux proposés à la reconnaissance s'écartent trop des données d'apprentissage. Globalement, c'est donc le problème de l'insuffisante prise en compte de la variabilité des données qui est en jeu ici. Les sources de variabilité sont multiples.

Benzeghiba *et al.* (2007) proposent une revue de ces différentes sources de variabilité et des approches adoptées en ASR (*Automatic Speech Recognition*, ou Reconnaissance Automatique de la Parole) pour les prendre en compte. Il y a de la variabilité extrinsèque liée aux conditions environnementales (bruit ambient, réverbération, distance de l'interlocuteur ou du capteur...), de la variabilité interlocuteur (genre, âge, différences physiologiques, vitesse d'élocution, accents régionaux, différences socio-linguistiques...) et de la variabilité intralocuteur liée au contexte (état émotionnel, adaptation à l'interlocuteur et à l'environnement). Cooke *et al.* (2014) proposent une revue des études comportementales portant sur ces effets d'adaptation de la production de la parole au contexte environnant. Parmi les 46 types de stratégies utilisées par des locuteurs humains ou par des algorithmes pour obtenir une meilleure intelligibilité, Cooke *et al.* (2014) en recensent une bonne vingtaine qui sont utilisées par les humains et se traduisent par des modifications des paramètres temporels ou spectraux du signal de parole.

Les sources de variabilité intrinsèque sont également multiples, liées aux variantes de production d'un locuteur à l'autre ou pour un même locuteur en fonction de son état émotionnel, de stress, ou de difficultés spécifiques rendant son élocution particulière, de façon temporaire ou permanente. C'est l'ensemble de ces sources de variabilité (qui sont décrites plus en détail dans la thèse de Camus (2011)) que les modèles acoustiques doivent être à même de capturer.

C'est dans ce double contexte de variabilité intrinsèque et extrinsèque que les spécialistes de reconnaissance automatique ont été amenés à proposer que des connaissances motrices ou des entrées articulatoires complémentaires pourraient jouer un rôle important et produire des améliorations de performance. Des améliorations pourraient survenir pour deux types de raisons différentes.

Hypothèse d'invariance motrice

D'abord, si l'on suit l'hypothèse des théories motrices selon laquelle les invariants seraient moteurs plutôt qu'auditifs, des paramètres articulatoires pourraient fournir une meilleure caractérisation des unités phonétiques, et donc réduire les phénomènes de variabilité liés à la coarticulation.

Hypothèse de génération motrice

Ensuite, les processus de génération motrice sous-jacents aux stimuli acoustiques pourraient s'avérer capables de prédire certaines sources de variabilité, de reconstituer de l'information manquante, et donc d'éliminer une part de variabilité liées aux troubles de la production. C'est dans ce contexte que l'on peut intégrer les travaux de Rudzicz (2010a,b, 2011a,b) ou de Young et Mihailidis (2010) sur la reconnaissance de parole mal articulée (parole dysarthrique, troubles du vieillissement). A l'inverse, Soltau *et al.* (2002) utilisent des connaissances motrices pour compenser l'effet de surarticulation des utilisateurs de systèmes automatiques cherchant à corriger des erreurs.

3.1.3 Paramétrisation articulatoire

L'idée d'utiliser des connaissances motrices en Reconnaissance Automatique de la Parole n'est pas nouvelle (voir Rose *et al.* (1994) ; Deng et Sun (1994) ; Deng *et al.* (1997) ; Ostendorf (1999)). L'enjeu principal de ces systèmes est alors avant tout de disposer effectivement d'entrées articulatoires pour configurer le système dans la phase d'apprentissage ou augmenter les entrées dans la phase de test. Il y a deux types d'entrées disponibles. D'abord, ce que l'on pourrait appeler les entrées réelles, c'est-à-dire les entrées fournies, à l'apprentissage ou au test, par des capteurs susceptibles de fournir des informations sur le conduit vocal. Les travaux suivant cette approche utilisent à peu près tous les capteurs disponibles et utilisés dans l'étude expérimentale de la production de la parole (voir des revues récentes dans King *et al.* (2007) ; Mitra *et al.* (2012)).

On trouve notamment des données de radiographie, de cinéradiographie ou d'articulographie électromagnétique (*electro-magnetic midsagittal articulometer, EMMA*). Frankel *et al.* (2000) utilisent ainsi des données EMMA recueillies sur 400 phrases du célèbre corpus TIMIT, fournissant des mesures à 500 Hz sur la position de la pointe, du corps et du dos de la langue, des lèvres inférieure et supérieure, de la mandibule et du velum ; Al Bawab (2009) adapte quant à lui les données EMMA du corpus MOCHA vers un modèle articulatoire sur lequel nous reviendrons au chapitre 5. Des systèmes sont également élaborés sur des données d'ultrason décrivant les mouvements de la langue (Hueber *et al.*, 2010). Et puis il faut bien évidemment intégrer à ce niveau tous les systèmes de reconnaissance audiovisuelle de la parole, utilisant en entrée, complémentairement au flux acoustique, une paramétrisation des mouvements des lèvres et de la face (voir une revue dans Potamianos *et al.* (2012)).

On trouve également des approches « sans données », dans lesquelles les concepteurs des systèmes tentent d'adjoindre aux entrées acoustiques des descripteurs articulatoires inspirés soit de classifications articulatoires abstraites (voir les travaux de Kirchhoff *et al.* (2002), avec des données classificatoires extraites directement de l'Alphabet Phonétique International, IPA), soit de théories de la production. Dans ce second registre, la Phonologie Articulatoire, dont nous avons parlé précédemment, a fourni le support à de nombreuses réalisations utilisant des descriptions associées à la position et au degré des constrictions du conduit vocal (voir par exemple Schmidbauer (1989) ; Deng et Sun (1994) ; Zweig (1998) ; Kirchhoff (1999) ; Sun *et al.* (2000) ; Stüker *et al.* (2003) ; Livescu *et al.* (2003) ; Scharenborg *et al.* (2007) ; Mitra *et al.* (2011, 2014)).

De manière beaucoup plus marginale, certains travaux sont basés sur des données synthétiques. C'est le cas de ceux de Huang et Er (2011, 2012) ; Huang (2012) qui apprennent une transformation des paramètres acoustiques vers des paramètres articulatoires grâce à un modèle de synthèse (le logiciel *PRAAT*, Boersma (1998)), et font ensuite des tests sur des corpus de parole réelle.

3.1.4 Combinaison des paramètres acoustiques et articulatoires

Une fois spécifiées les variables articulatoires insérées dans le modèle, il s'agit pour le concepteur de déterminer comment intégrer ces variables dans les processus d'apprentissage et de test. Au niveau du test, une forte ligne de partage sépare les systèmes disposant effectivement au niveau de leur entrée de variables articulatoires complémentaires, des systèmes qui ne comportent en

réalité qu'une entrée auditive.

Dans le premier groupe de systèmes, que l'on pourrait baptiser de multicapteurs, figure avant tout l'ensemble des systèmes de reconnaissance audiovisuelle de la parole, pour lesquels un enjeu majeur est celui de la fusion des capteurs auditif et visuel (Schwartz *et al.*, 2002b ; Potamianos *et al.*, 2012). Le même type de problème se pose pour les systèmes disposant de données de capteurs ultrason sur la langue (Hueber *et al.*, 2010).

Par contre, pour les « systèmes sans données », il s'agit d'être capable dans la phase de test de restituer des paramètres articulatoires (traits articulatoires caractéristiques de constrictions, etc...) à partir du son incident. Ceci passe donc par la mise en place d'associateurs de formes diverses : réseaux de neurones multicouches pour Frankel *et al.* (2000), *Support Vector Machines (SVMs)* et perceptrons multicouches (MLPs) pour Scharenborg (2007) ; réseaux bayésiens dynamiques pour Zweig (1998) ; Frankel *et al.* (2007) ; Neiberg *et al.* (2009) ; *Deep Neural Networks (DNNs)* pour Mitra *et al.* (2014). L'évaluation vise alors à savoir si le remplacement des données acoustiques par les traits articulatoires estimés, ou l'apport de ces estimations en sus des entrées acoustiques, peut produire des gains de performance. Globalement, tous ces modèles proposent des preuves de concept souvent convaincantes, mais les gains de performance sont en général assez aléatoires.

3.1.5 Traitement temporel

Il peut être intéressant à ce stade de mentionner brièvement quelques outils de traitement de séquences temporelles à l'oeuvre dans ces différents modèles. Classiquement, la prise en compte des effets de contexte temporel est gérée dans les systèmes de reconnaissance automatique de la parole par les modèles de Markov cachés qui permettent de modéliser des effets d'ordre un, et s'avèrent, par leur souplesse d'emploi, bien adaptés au traitement des séquences de parole. D'autres outils ont été régulièrement introduits dans les vingt dernières années et évoluent très vite : réseaux de neurones récurrents (*Time-Delayed Neural Networks*, voir un exemple récent dans Graves *et al.* (2013)), *Deep Neural Networks* (Bacchiani *et al.*, 2014), *Long Short Term Memory Networks* (Sak *et al.*, 2014), réseaux bayésiens dyamiques (voir Mitra *et al.* (2012, 2014)). Enfin, la dynamique est souvent introduite directement dans la paramétrisation, sous forme de dérivées première ou seconde de coefficients spectraux, ou encore en donnant en entrée non pas des valeurs de paramètres instantanées, mais plutôt toute une fenêtre de paramètres, ce qui revient à prendre en compte tout un contexte.

3.2 Modèles cognitifs

Il s'agit ici, non pas d'être guidés par une démarche de développement et d'amélioration de systèmes (ou en tout cas pas exclusivement), mais de développer des modèles quantitatifs visant à servir de support à des hypothèses théoriques sur le rôle des connaissances motrices dans les mécanismes de perception.

Les travaux dans ce domaine sont plus rares. Nous en présenterons trois, car ils nous semblent assez illustratifs, qu'ils représentent des approches assez différentes, et qu'ils sont décrits chacun avec un niveau de détail ou de conceptualisation suffisant pour en permettre une description explicite. A contrario, nous ne présenterons pas ici le cadre intégrateur introduit et défendu depuis de nombreuses années par Pickering et Garrod (2013), car il est plus un cadre

de pensée et d'argumentation qu'un réel modèle quantitatif : il n'est d'ailleurs jamais fourni de simulations quantitatives par les auteurs.

3.2.1 Les travaux computationnels du groupe de Luciano Fadiga

Outre de nombreux travaux dans le domaine des neurosciences, dont nous en avons présenté certains à la section 2.2 qui visent à mettre en évidence le rôle fonctionnel du système moteur en perception, un groupe de chercheurs travaillant autour de Luciano Fadiga a récemment publié un certain nombre d'études utilisant des modèles computationnels pour défendre le rôle des connaissances motrices en perception (Castellini *et al.*, 2011 ; Badino *et al.*, 2012 ; Canevari *et al.*, 2013c,b,a ; Badino *et al.*, 2014).

Comme l'illustre la figure 2.8, l'approche qu'ils adoptent est de regarder à quel point l'ajout de paramètres moteurs en entrée d'un système de reconnaissance automatique permet d'améliorer les performances du système.

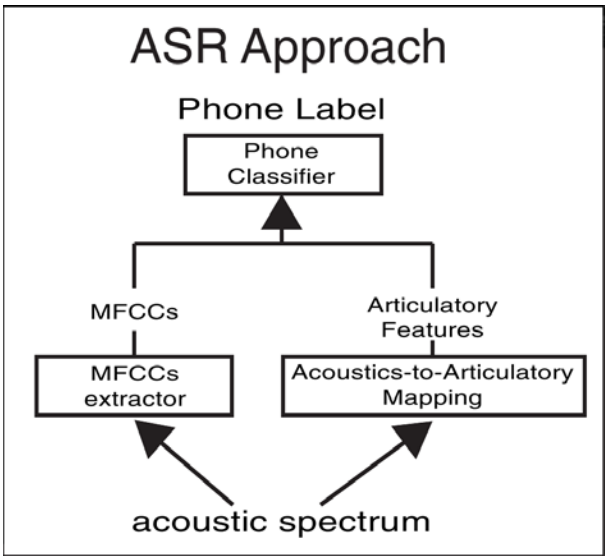


Figure 2.8: **Un mélange de paramètres articulatoires et acoustiques pour la reconnaissance de la parole** (Cette figure est tirée de Canevari *et al.* (2013a)).

Se posent alors les questions de savoir de quels ensembles de données les paramètres articulatoires sont extraits, comment le système y a accès (sont-ils donnés en entrée ou sont-ils inférés à partir des entrées acoustiques ?), quels sont les outils utilisés pour apprendre les associations entre ces paramètres et les unités phonétiques calculées en sortie, et enfin à quels écarts de performances conduit l'utilisation de ces connaissances motrices et dans quelles conditions.

Castellini *et al.* (2011) cherchent à reproduire avec un modèle les résultats expérimentaux de D'Ausilio *et al.* (2009) qui ont montré que la discrimination dans le bruit des plosives dentales (/d/ et /t/) et bilabiales (/b/ et /p/) était impactée par la simulation magnétique transcranienne des zones motrices associées au contrôle de la langue ou des lèvres. À partir de données d'articulographie électromagnétique (Grimaldi *et al.*, 2008), ils apprennent un réseau de neurones qui à partir d'entrées acoustiques calcule les paramètres articulatoires correspondants. Ils obtiennent des résultats qui montrent que, dans ce cadre, adjoindre aux paramètres acoustiques

des paramètres articulatoires inférés grâce à leur réseau de neurones améliore les performances en présence de bruit.

Badino *et al.* (2012) utilisent des données issues du corpus *MOCHA-TIMIT* (Wrench, 1999) pour apprendre les paramètres d'un *Multi-Layer Perceptron* (*MLP*) servant à réaliser le passage de l'espace acoustique vers l'espace articulatoire. Un système de reconnaissance, implémenté sous forme de DBN-HMM (*Deep Belief Network* et *Hidden Markov Model*), et qui prend en entrée la combinaison des paramètres acoustiques et des paramètres articulatoires inférés par le *MLP*, montre une réduction relative de 16.6% du taux d'erreur de reconnaissance des phonèmes par rapport à un système similaire n'utilisant que les paramètres acoustiques. Il faut tout de même souligner que l'apprentissage et le test se font sur des données produites par un seul locuteur.

Avec les mêmes données, Canevari *et al.* (2013c) apprennent l'inversion des paramètres acoustiques vers l'espace articulatoire sous forme de *Deep Neural Network*. Ils montrent une diminution relative du taux d'erreur de reconnaissance de phonèmes de 8.4% en combinant les paramètres moteurs reconstitués aux paramètres acoustiques, qui passe à 10.9% lorsque le processus de reconstruction attribue des poids différents aux paramètres articulatoires en fonction d'une mesure de leur pertinence.

Canevari *et al.* (2013b) montrent des résultats plus mitigés que ceux de Badino *et al.* (2012) ; Canevari *et al.* (2013c). En plus du corpus utilisé par Badino *et al.* (2012), ils utilisent les données du corpus *mngu0* (Richmond *et al.*, 2011) et celles du « Lecce corpus » (Grimaldi *et al.*, 2008) (parmi lesquelles ils n'utilisent que cinq locutrices italiennes). L'ajout des connaissances motrices n'apporte pas un gain systématique sur l'ensemble des corpus. Ils proposent donc une analyse des erreurs au niveau de chaque catégorie phonémique pour voir lesquelles sont le mieux impactées.

Avec des outils similaires, Canevari *et al.* (2013a) s'intéressent à l'hypothèse selon laquelle les représentations motrices permettraient de faire de la normalisation du locuteur. Ils apprennent donc la correspondance entre l'acoustique et l'articulatoire à partir des données de l'auditeur et d'un locuteur servant à l'apprentissage, et font le test sur des données acoustiques produites par un locuteur différent. Ils observent alors une faible baisse de l'erreur relative (à peine quelques pourcents) liée à l'utilisation de connaissances motrices.

Finalement, il ressort de ces travaux que l'apport de l'utilisation de connaissances motrices en condition de parole normale (au sens de test dans des conditions identiques à celles de l'apprentissage) n'est pas toujours clair, mais qu'il le devient en conditions dégradées (présence de bruit, ou test sur un locuteur différent).

3.2.2 Le modèle neurocomputationnel *ACT* proposé par Bernd Kröger

Kröger *et al.* (2012) décrivent un modèle neurocomputationnel du traitement sensori-moteur de la parole sur lequel ils travaillent depuis une dizaine d'années. Ils le baptisent *ACT* pour souligner que ce sont les ACTions du conduit vocal – niveau auquel se fait la planification motrice – qui sont les unités centrales de ce modèle. La figure 2.9 présente la structure du modèle *ACT*.

Comme le montre la figure 2.9, la structure du modèle *ACT* est assez proche de celle du modèle *DIVA* que nous avons présenté à la section 2.1.1. En effet, les composants principaux des deux modèles sont : des représentations phonémiques, un système de contrôle direct, un

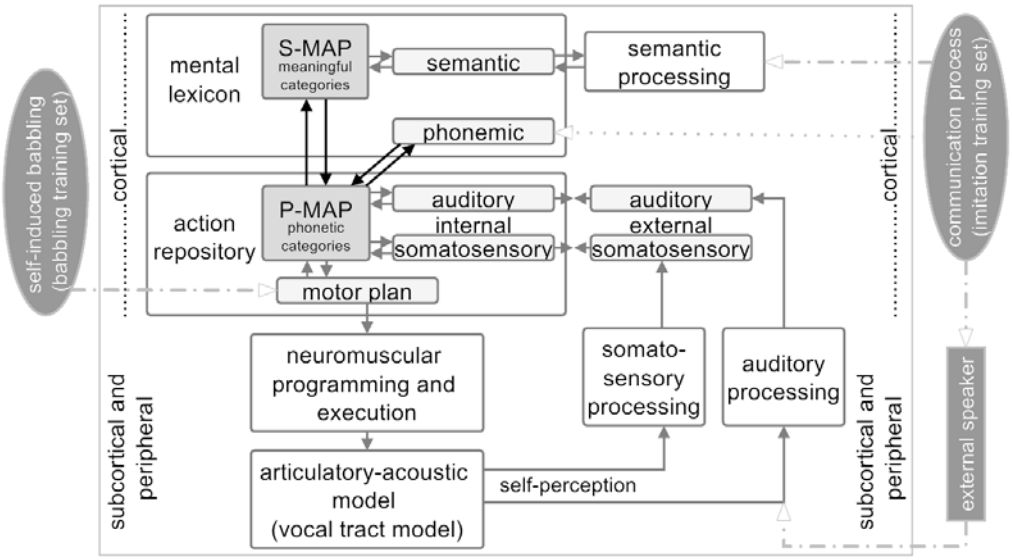


Figure 2.9: ***ACT* : un modèle neurocomputationnel du contrôle sensori-moteur de la parole.** Les cadres blancs représentent des modules de traitement, les cadres gris représentent des cartes auto-adaptatives (*S-MAP* et *P-MAP*) ou des cartes d'états mentaux, au sein desquelles les activations correspondent à des représentations sémantiques, phonémiques, auditives, somatosensorielles, ou à un plan moteur. Les ellipses grises représentent la manière dont les connaissances sont acquises par le modèle, par interaction avec l'extérieur. (Cette figure est tirée de Eckers et Kröger (2012), qui ont eux-mêmes adapté Kröger *et al.* (2011)).

système de rétroaction, et un simulateur de conduit vocal. Ainsi, les deux modèles sont capables, à partir de représentations phonémiques, de générer des trajectoires articulatoires cohérentes ainsi que les signaux acoustiques correspondants.

Le modèl *ACT* repose sur un modèle articulatoire de conduit vocal à trois dimensions (Birkholz et Jackel, 2004 ; Birkholz et Kröger, 2006, 2007 ; Birkholz *et al.*, 2006, 2007 ; Kröger et Birkholz, 2007). C'est grâce aux productions de ce modèle de conduit vocal que les paramètres du modèle *ACT* peuvent être appris. Kröger *et al.* (2009) s'appuient sur les travux de Oller *et al.* (1999) pour justifier que cet apprentissage se fasse en deux temps : une phase de babillage (Kröger *et al.*, 2006), qui est implémentée sous forme d'exploration aléatoire de l'espace des paramètres articulatoires du modèle de conduit vocal, et une phase d'imitation, durant laquelle les connaissances qui ont émergé durant la phase de babillage sont affinées.

Les apprentissages reposent sur des méthodes classiques d'apprentissage Hebbien (Hebb, 2002) dans les réseaux de neurones, et utilisent des cartes auto-adaptatives (aussi connues sous le noms de cartes de Kohonen, Kohonen (1982)), pour faire émerger de la méta-structure dans les co-activations apprises.

Un certain nombre de travaux ont montré ce que le modèle *ACT* est capable d'apprendre dans des configurations de plus en plus complexes :

- le système de cinq voyelles /i-e-a-o-u/ (Kröger *et al.*, 2009) ;
- des syllabes de type Consonne-Voyelle en combinant ces cinq voyelles avec les plosives /b-d-g/ (Kröger *et al.*, 2009) ;

- un petit modèle de langage en ajoutant à ce qui précède les consonnes /p-t-k-m-n-l/ pour former des syllabes de type Voyelle, Consonne-Voyelle ou Consonne-Consonne-Voyelle (Kroger *et al.*, 2011) ;

- les 200 syllabes les plus fréquentes de l'allemand adressé à un enfant de six ans (Kröger *et al.*, 2011) ;

- sur ce même ensemble de données, Kröger *et al.* (2014) parviennent à faire émerger des caractéristiques phonétiques comme l'accentuation des syllabes, le lieu et la manière d'articulation des consonnes, le caractère antérieur/postérieur des voyelles, etc.

Bien que, comme le modèle *DIVA*, le modèle *ACT* ait été conçu comme un modèle de production, il permet donc d'étudier l'acquisition de la parole, et également de rendre compte de certains phénomènes liés à la perception. Le modèle supporte ainsi l'idée que la perception catégorielle soit plus marquée pour les plosives que pour les voyelles (Kröger *et al.*, 2009), et une version du modèle étendu pour prendre en compte l'information visuelle du signal de parole permet également d'observer une certaine forme d'effet McGurk grâce à un mécanisme d'inhibition (Kröger et Kannampuzha, 2008).

3.2.3 Le modèle *PRESENCE* proposé par Roger Moore : un cadre de réflexion générale qui transcende la recherche en parole

Le modèle *PRESENCE* (PREdictive SEnsorimotor Control and Emulation, Moore (2007)) est le fruit de la volonté de son auteur de s'inspirer d'un grand nombre de domaines de recherche, dont certains sont assez éloignés de celui de la parole, pour intégrer un ensemble de principes au sein d'un modèle unique dans le but de proposer une vision globale qui pourrait aider à guider la recherche en traitement de la parole, par la machine ou par l'humain.

Moore (2013) résume quatre idées clés sur lesquelles est basé le modèle *PRESENCE* [12] :

- Il existe un lien intime entre les comportements sensoriels et les comportements moteurs chez les organismes vivants, ce qui est attesté par des preuves de plus en plus nombreuses (voir la section 2.2).

- L'adaptation à des perturbations dans des environnements réels peut se faire grâce au principe de contre-réaction (en anglais, *negative feedback control*, voir Powers (1973) ; Lindblom (1990a)) qui permet d'adapter dynamiquement l'action d'un système en fonction de l'écart à l'objectif perçu pour compenser les perturbations et maintenir un état stable.

- L'apprentissage et la modélisation pourraient être facilités par l'utilisation de mécanismes d'imitation ou d'imagerie mentale (Iacoboni, 2005 ; Wilson et Knoblich, 2005).

[12] « four areas could have significant implications for the future of spoken language processing (Moore, 2007) : the growing evidence for an intimate relationship between sensor and motor behaviour in living organisms (Pulvermüller, 2005 ; Rizzolatti et Arbib, 1998), the power of negative feedback control to accommodate unpredictable disturbances in real-world environments (Powers, 1973 ; Lindblom, 1990a), mechanisms for imitation and mental imagery for learning and modelling (Iacoboni, 2005 ; Wilson et Knoblich, 2005), and hierarchical models of temporal memory for predicting future behaviour and anticipating the outcome of events (Hawkins et Blakeslee, 2004). »(Moore, 2013, pages 124–125)

- La prédiction de comportements futurs ou l'anticipation de l'issue d'évènements peut être implémentée par des modèles hiérarchiques de mémoire temporelle (Hawkins et Blakeslee, 2004).

La figure 2.10 montre l'architecture du modèle *PRESENCE*.

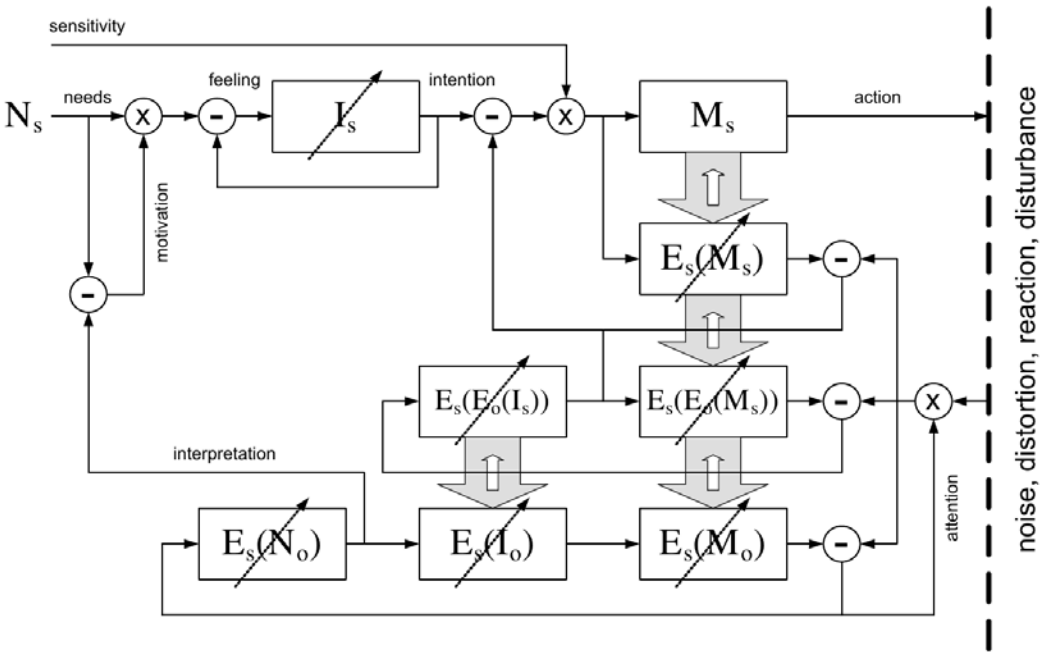


Figure 2.10: **Le modèle PRESENCE** (*PREdictive SENsorimotor Control and Emulation*). Les indices S et O réfèrent respectivement à soi (*Self*) et à l'autre (*Other*). N représente les besoins (*Needs*), I les intentions (*Intentions*), M l'activité motrice (*Motor activity*) et la notation $E()$ désigne un processus d'émulation (*Emulation*). Les flèches diagonales en pointillés indiquent un processus de recherche. Par exemple, le module I_S en haut à gauche permet de représenter la recherche d'une intention permettant de satisfaire un besoin. Cette figure est tirée de Moore (2007).

Nous ne décrirons pas chaque couche de ce modèle présenté figure 2.10 (le lecteur intéressé pourra se référer à Moore (2007)) mais nous nous contenterons de souligner deux points que cette figure met clairement en avant, et qui seront importants pour la suite. Premièrement, le cœur du modèle *PRESENCE* réside dans un mécanisme d'émulation (E), qui est à rapprocher de la notion de modèle interne, que nous avons présentée à la section 2.3.1. Deuxièmement, ce modèle fait intervenir de manière tout aussi centrale la distinction entre ses états mentaux propres (indice S) et ceux de l'autre (indice O), ce qui est à rapprocher de la théorie de l'esprit, que nous avons introduite à la section 2.3.2.

Finalement, le modèle *PRESENCE* peut être vu comme un modèle général décrivant un agent en interaction, mais si l'on considère les implications pour le domaine de la communication parlée voici ce qu'il en ressort d'après Moore (2013) :

- Un système de production de parole devrait choisir des caractéristiques adaptées aux besoins de l'auditeur, surveiller l'effet de ses propres productions, et modifier son comportement selon ses modèles internes de l'auditeur.

- Un modèle de reconnaissance de parole pourrait capturer les intentions du locuteur dans un modèle interne générateur (direct) et adapter ce modèle à la voix du locuteur en se basant sur la connaissance de sa propre voix.

4 Notre chemin de modélisation

Malgré de nombreuses années de débat entre les défenseurs des théories motrices et ceux des théories auditives, aucun des arguments proposés de part et d'autre n'est décisif, et il semble que le débat théorique stagne. Les études sur la production et la perception de la parole se font la plupart du temps de manière indépendante, voire même cloisonnée. Pourtant, chaque camp fait des hypothèses, hélas parfois implicites, sur ce qui se passe de l'autre côté. Même dans le domaine des modèles cognitifs, que nous venons de passer en revue, il n'y a pas de comparaison claire des approches perceptives, motrices et perceptuo-motrices – et, du coup, pas d'analyse des raisons computationnelles pour lesquelles on pourrait être amené à sélectionner telle ou telle architecture – ou, dans un cadre plus large, de ce qui pourrait, par un gain de performance, avoir fourni un éventuel mécanisme phylogénétique ayant conduit à sélectionner dans l'évolution tel mécanisme plutôt que tel autre.

L'approche adoptée dans cette thèse consiste à se donner un cadre formel permettant d'étudier conjointement les aspects liés à la production et à la perception de la parole. Pour cela, nous utilisons des outils mathématiques pour construire un cadre intégrateur dans lequel formaliser les hypothèses sur lesquelles reposent les différentes théories issues de la littérature, afin de pouvoir les comparer de manière quantitative.

Plus précisément, le cadre formel que nous choisissons d'utiliser est celui de la programmation bayésienne (Bessière *et al.*, 2013). Ce cadre est inspiré des travaux de Jaynes (2003), qui propose d'adopter une approche subjectiviste des probabilités : représenter des états de connaissance sous forme de distributions de probabilité construites à partir d'observations ponctuelles permet de traduire la variabilité et l'incertitude inhérentes aux phénomènes observés.

Il s'agit d'une approche particulièrement adaptée aux problèmes de robotique (Lebeltel *et al.*, 2004) qui permet aux robots d'adopter, à partir d'observations partielles, un comportement approprié en environnement non-contrôlé. Par ailleurs, les approches bayésiennes sont également de plus en plus répandues dans le domaine de la modélisation cognitive (Jones et Love, 2011 ; Griffiths *et al.*, 2010 ; Perfors *et al.*, 2011 ; Colas *et al.*, 2010). Le choix de la modélisation bayésienne qui est fait dans cette thèse comme cadre formel s'appuie sur de nombreuses motivations, parmi lesquelles on peut en citer au moins deux.

- Les outils bayésiens permettent de construire des modèles plausibles et facilement interprétables de systèmes sensori-moteurs complexes (Bessière *et al.*, 2013, 2008 ; Lebeltel *et al.*, 2004). Par exemple, Gilet *et al.* (2011) proposent un modèle de boucle perception-action qui permet d'étudier l'interaction entre connaissances motrices et perceptives dans le cas de l'écriture. Dans cette thèse, c'est une approche similaire qui est adoptée, afin d'étudier les interactions perceptuo-motrices dans le cas de la parole.

- Nous défendons l'idée qu'il est nécessaire de se placer au niveau de l'étude de la situation de communication, où chacun est tour à tour en situation de production ou de perception, pour bien étudier le rôle des connaissances motrices et des connaissances perceptives en

parole. Dans ce cadre, l'utilisation des probabilités pour représenter des états de connaissances, comme le propose Jaynes (2003), permet de construire des modèles subjectifs d'agents bayésiens, capables à la fois de produire et de percevoir la parole, en décrivant de manière explicite comment ces agents internalisent des propriétés de leur environnement ou de leur interlocuteur. Un tel modèle d'agent a été proposé dans la thèse de Moulin-Frier (2011) qui étudie l'émergence de systèmes phonologiques au sein d'une société d'agents.

Chapitre 3

COSMO : Un modèle formel et générique d'agent communicant pour l'étude quantitative des interactions sensori-motrices lors de la communication parlée

1 Modélisation conceptuelle de la situation de communication parlée **37**
2 *COSMO* : Un modèle formel probabiliste d'agent communicant **40**
3 Inférences probabilistes dans *COSMO* **43**
4 Théorème d'indistinguabilité des théories motrice et auditive en perception de la parole **51**
5 Conclusion **56**

Ce chapitre présente l'élaboration de *COSMO*, un modèle probabiliste d'agent communicant qui, construit sur une hypothèse d'internalisation de la situation de communication, offre un cadre computationnel intégrateur permettant d'étudier les interactions sensori-motrices en production et en perception de la parole.

1 Modélisation conceptuelle de la situation de communication parlée

Le *Trésor de la Langue Française* définit la parole comme étant la « faculté d'exprimer et de communiquer la pensée au moyen du système des sons du langage articulé émis par les organes phonateurs ». Pour ce qui est de la communication, toujours d'après le Trésor de la Langue Française, il s'agit d'un « processus par lequel une personne (ou un groupe de personnes) émet un message et le transmet à une autre personne (ou groupe de personnes) qui le reçoit, avec une marge d'erreurs possibles (due, d'une part, au codage de la langue parlée ou écrite,

langage gestuel ou autres signes et symboles, par l'émetteur, puis au décodage du message par le récepteur, d'autre part au véhicule ou canal de communication emprunté). »

Si l'on regroupe ces deux définitions, il vient que la communication parlée est un processus par lequel un locuteur (ou émetteur) encode un message sous forme de sons du langage articulé, puis que ce message, transmis par l'environnement, est décodé par l'auditeur (ou récepteur). Ces trois phases de la communication parlée sont schématisées figure 3.1.

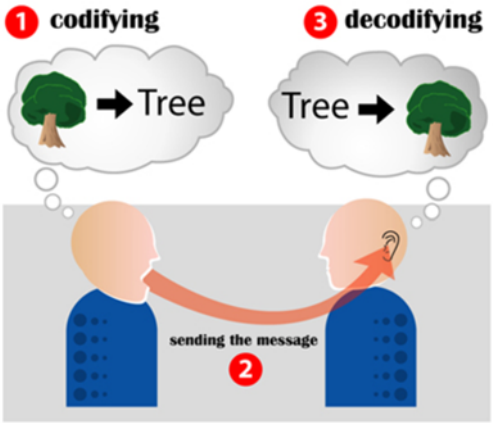


Figure 3.1: **Les trois phases de la communication**: (1) encodage d'un message par le locuteur, (2) transmission par l'environnement, (3) décodage du message par l'auditeur (d'après `http://en.wikipedia.org/wiki/Communication`).

Nous proposons maintenant d'étudier en détail les différentes étapes de la communication parlée, représentées figure 3.1, afin d'en extraire les variables permettant de décrire la situation de communication, dans le but d'aller vers un modèle formel.

Tout d'abord, le locuteur a une représentation interne du message qu'il souhaite transmettre. Dans tout notre travail nous considérerons des messages extrêmement simplifiés, constitués exclusivement d'un seul item que nous dénommerons « objet ». Ainsi, nous supposons que la communication entre émetteur et récepteur s'établit autour d'un tel « objet », dont l'émetteur souhaite transmettre vocalement l'identité au récepteur. Nous appelons cet objet O_S, pour « objet considéré du point de vue du locuteur » (*speaker*[13]). Nous utilisons le terme d'objet de communication au sens large, en compactant les différents niveaux d'analyses (phonétique, phonologique, lexical, morphologique, syntaxique, sémantique...). L'objet peut ainsi être aussi bien un mot qu'un phonème ou un concept.

Ensuite, pour communiquer, le locuteur utilise ses organes phonatoires. Des commandes musculaires sont envoyées pour contrôler l'action de différents articulateurs, le larynx, les cordes vocales, le pharynx, le velum, la mandibule, la langue et les lèvres, afin d'articuler les sons du langage. Le système moteur comprend donc des représentations[14] des commandes articulatoires et des processus cognitifs permettant le contrôle du conduit vocal. Nous appelons M la représentation interne de ce système moteur[15].

[13]Par souci de consistance, nous choisissons de conserver dans cette thèse les notations utilisées dans les articles publiés dans les revues et conférences internationales, qui sont rédigés en anglais.

[14]« Représentation » est ici pris dans un sens très général, correspondant typiquement à un ensemble de valeurs accessibles au locuteur pour mettre en place des processus inférentiels.

[15]Bien évidemment, il existe très certainement une hiérarchie de représentations motrices, des programmes

Le conduit vocal du locuteur produit alors une onde sonore, qui est transmise par l'environnement jusqu'au système sensoriel de l'auditeur. Ce système sensoriel comprend l'oreille, qui permet de percevoir le son, et l'ensemble des représentations et processus cognitifs qui assurent le traitement auditif dans le cerveau. Nous appelons S la représentation interne de ce système sensoriel[16].

L'auditeur a également une représentation interne de l'objet communiqué, qu'il reconnaît à partir du percept sensoriel S. Nous l'appelons O_L, pour « objet considéré du point de vue de l'auditeur » (*listener*). Il est possible que l'objet O_S que le locuteur veut transmettre et l'objet O_L que l'auditeur reconnaît soient différent, à cause d'erreurs de codage ou de décodage, ou à cause de dégradations dues à l'environnement. Nous supposons donc qu'en plus du signal de parole, les agents disposent d'une autre manière de vérifier la cohérence entre O_S et O_L. Ceci peut par exemple prendre la forme d'un mécanisme d'attention partagée, tel que la deixis[17]. Nous introduisons donc une variable supplémentaire C_{Env}, qui décrit le succès de la communication (c'est-à-dire le fait que O_S et O_L soient identiques).

La figure 3.2 présente un modèle de l'interaction entre un agent locuteur et un agent auditeur en situation de communication.

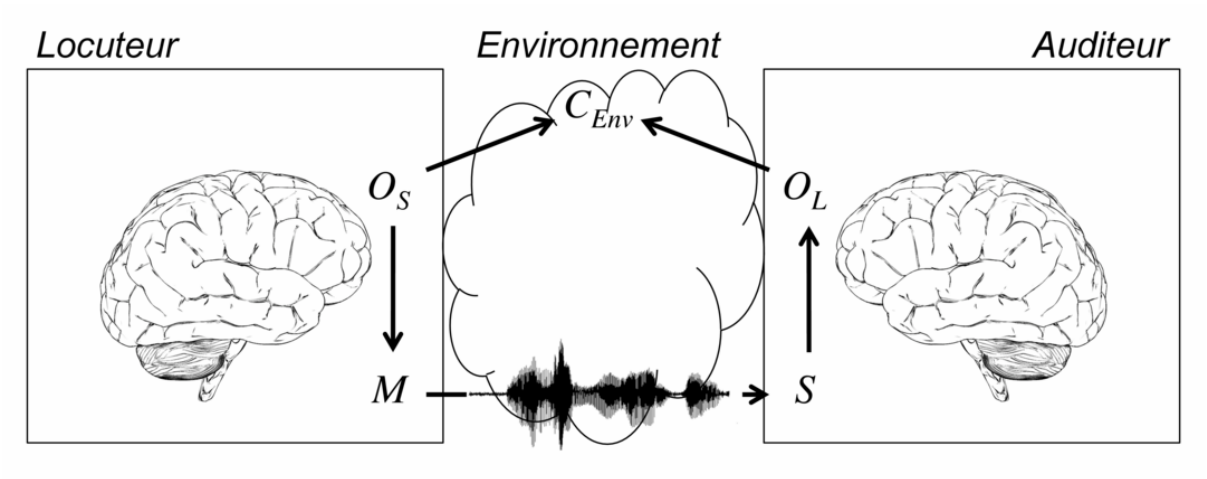


Figure 3.2: **La situation de communication**: un locuteur communique un objet O_S à un auditeur en produisant un geste moteur M qui, propagé par l'environnement, arrive à l'auditeur sous la forme d'un percept sensoriel S, ce qui lui permet de reconnaître l'objet O_L. La variable C_{Env} décrit le succès de la communication, ce qui est le cas lorsque les objets O_S et O_L sont identiques. (figure tirée de Laurent *et al.* (2012), adaptée de Moulin-Frier *et al.* (2012).)

moteurs de haut niveau jusqu'aux commandes motrices proches de la réalisation effective (voir par exemple Rapin (2011)). Sans rentrer dans cette complexité potentielle, nous considérons M comme le point d'entrée de bas niveau de cette hiérarchie, au plus près des sorties motrices.

[16]De même, il existe très certainement une hiérarchie de représentations sensorielles et perceptives, de la première analyse au niveau des capteurs primaires du système auditif (les "cellules ciliées" de la membrane basilaire dans la cochlée) jusqu'à des représentations élaborées dans le cortex, du cortex auditif primaire au cortex auditif secondaire, aux aires associatives et jusqu'à des réseaux corticaux complexes (voir par exemple Grabski (2012)). Sans rentrer dans cette complexité potentielle, nous considérons S comme le point d'entrée de bas niveau de cette hiérarchie, au plus près des entrées sensorielles.

[17] La deixis désigne la capacité de partager une référence commune en montrant des choses du geste ou du regard. Cette notion est présentée plus en détails dans la section 2.3.2 du chapitre 2.

2 *COSMO* : Un modèle formel probabiliste d'agent communicant

Le modèle proposé figure 3.2 est un modèle asymétrique, dans lequel chaque agent est enfermé dans un rôle fixe : l'agent locuteur en production de parole, et l'agent auditeur en perception. Or nous avons défendu dans le chapitre précédent que l'étude des systèmes cognitifs en production et en perception ne pouvait pas se limiter à des systèmes respectivement purement moteurs et purement perceptifs. Cette section présente un modèle probabiliste d'agent communicant qui, ayant internalisé la boucle de communication dans sa totalité, est capable de se comporter tour à tour comme un locuteur et comme un auditeur, et de mettre en relation ses processus moteurs et perceptifs.

2.1 Hypothèse d'internalisation de la situation de communication dans l'architecture cognitive de l'agent communicant

Construire un modèle d'agent communicant, c'est proposer une architecture reliant les différentes représentations cognitives que cet agent sollicite pour la communication.

Même si à un moment donné cet agent se comporte soit en tant que locuteur soir en tant qu'auditeur, il est clair qu'il remplit alternativement les deux rôles au cours de son existence. Il faut donc qu'il dispose à la fois d'un système moteur reliant des objets à des représentations motrices, et d'un système perceptif, qui relie des représentations sensorielles aux objets.

Mais production et perception de la parole sont-elles des tâches complètement indépendantes ? Les sytèmes moteurs et perceptifs sont-ils tout à fait cloisonnés ? Les données expérimentales présentées dans la section 2.2 du chapitre précédent répondent négativement à ces deux questions en mettant en évidence l'existence et le rôle fonctionnel possible d'un lien sensori-moteur. Le modèle d'agent que l'on est en train de construire comprendra donc un lien explicite entre variables motrices et variables perceptives.

Par ailleurs, les données expérimentales présentées par Jacquemot *et al.* (2007) mènent à la conclusion qu'il existe une séparation fonctionnelle entre les codes phonologiques en production et en perception, mais que ces codes phonologiques différents en entrée et en sortie sont reliés par des mécanismes de conversion. C'est pourquoi nous choisissons dans notre modèle de faire figurer séparément les représentations des objets du point de vue de la production et de la perception, en nous donnant un lien supplémentaire pour assurer la cohérence entre ces représentations.

Finalement, alors que la situation de communication représentée figure 3.2 était décrite par les liens entre les cinq variables O_S, M, S, O_L et C, nous montrons figure 3.3 que l'architecture cognitive que l'on retient pour notre agent communicant consiste en une internalisation de ces cinq variables et des liens qui les relient.

L'agent communicant que l'on modélise a donc internalisé la totalité de la boucle de communication présentée figure 3.2 qui, partant de l'objet phonologique à communiquer O_S, passe par un ensemble de représentations des commandes motrices M, dont il sait prédire les signaux perceptifs S qui en résultent, à partir desquels l'objet O_L est reconnu, puis comparé à l'objet O_S pour déterminer le succès de la communication exprimé par la variable C. Cette internalisation d'une boucle de communication complète, permettant à un agent de se représenter son partenaire de communication et d'adapter ses stratégies en conséquence, est justifiée pour nous

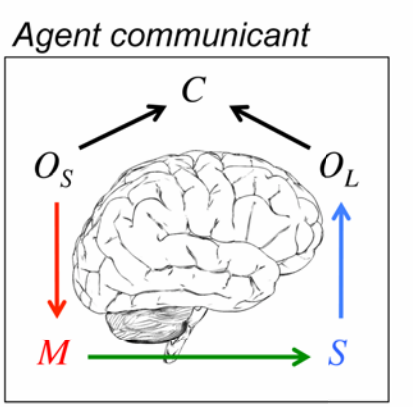


Figure 3.3: **L'agent communicant**, ayant internalisé la situation de communication, peut se comporter à la fois comme locuteur et comme auditeur.

comme étant une composante possible et vraisemblable d'une « théorie de l'esprit » telle que nous l'avons introduite dans la section 2.3.2 du chapitre précédent. Nous verrons que cette hypothèse d'internalisation permet des inférences puissantes fournissant un cadre intégrateur potentiel aux principales théories de la communication parlée.

Nous baptisons ce modèle cognitif *COSMO*, ce qui constitue un acronyme reprenant les noms des différentes variables apparaissant dans le modèle, et qui signifie également *Communicating Objects using Sensori-Motor Operations*.

2.2 Dépendances probabilistes du modèle *COSMO*

Conformément au cadre de la Programmation Bayésienne des Robots (Lebeltel *et al.*, 2004 ; Bessière *et al.*, 2008 ; Colas *et al.*, 2010 ; Bessière *et al.*, 2013), nous adoptons une approche subjectiviste des probabilités selon laquelle nos modèles encodent des états de connaissances, susceptibles d'évoluer au cours du temps, et qui comprennent une incertitude quantifiée sous forme de distributions de probabilité. Le modèle *COSMO* est alors décrit dans sa totalité par un unique objet mathématique : la distribution de probabilité conjointe sur l'ensemble des variables d'intérêt $P(O_S\ M\ S\ O_L\ C)$.

Puisque les objets que nous manipulons avec les modèles bayésiens utilisés dans cette thèse sont des distributions de probabilité, rappelons ici quelques règles élémentaires de calcul probabiliste, que nous utiliserons dans toute la suite.

La règle de normalisation fait partie de la définition des distributions de probabilité :

$$\sum_A P(A) = 1 \ . \tag{3.1}$$

La règle du produit est liée à l'une des définitions de la probabilité conditionnelle :

$$P(A\ B) = P(A)P(B \mid A) = P(B)P(A \mid B) \ . \tag{3.2}$$

La règle de marginalisation est une combinaison des deux précédentes :

$$P(A) = \sum_B P(A\ B) \ . \tag{3.3}$$

Le théorème de Bayes s'obtient à partir de la règle du produit :

$$P(A \mid B) = \frac{P(A)P(B \mid A)}{P(B)} \ , \quad P(B) \neq 0 \ . \tag{3.4}$$

La distribution de probabilité conjointe définissant notre modèle *COSMO*, par utilisations successives de la règle du produit, peut se réécrire ainsi :

$$P(O_S\ M\ S\ O_L\ C) = P(O_S)P(M \mid O_S)P(S \mid M\ O_S)P(O_L \mid S\ M\ O_S)P(C \mid O_L\ S\ M\ O_S) \tag{3.5}$$

En tant que modélisateur, nous faisons maintenant une série d'hypothèses simplificatrices pour réduire la complexité du modèle et des calculs d'inférences futurs, ce qui a aussi pour effet de rendre le modèle plus facile à interpréter.

1. S est indépendant de O_S conditionnellement à M. Cela revient à dire que la connaissance du geste moteur M produit par le locuteur est suffisante pour déterminer le son produit S.

$$P(S \mid M\ O_S) = P(S \mid M) \ . \tag{3.6}$$

2. O_L est indépendant de O_S et de M conditionnellement à S. Cela revient à supposer que l'objet O_L reconnu par l'auditeur ne dépend que de l'entrée perceptive S qui lui parvient.

$$P(O_L \mid S\ M\ O_S) = P(O_L \mid S) \ . \tag{3.7}$$

3. C est indépendant de M et de S conditionnellement à O_S et à O_L. Cela revient à dire que le succès de la communication ne dépend que des objets O_S et O_L considérés par le locuteur et l'auditeur.

$$P(C \mid O_L\ S\ M\ O_S) = P(C \mid O_L\ O_S) \ . \tag{3.8}$$

Ces hypothèses d'indépendances conditionnelles conduisent à une nouvelle expression de la distribution de probabilité conjointe, que l'on appelle « décomposition » :

$$P(O_S\ M\ S\ O_L\ C) = P(O_S)P(M \mid O_S)P(S \mid M)P(O_L \mid S)P(C \mid O_S\ O_L) \ . \tag{3.9}$$

Au vu de cette équation, il est maintenant possible d'interpréter la figure 3.3 comme un réseau bayésien où les flèches représentent les distributions de probabilité conditionnelles qui interviennent dans la décomposition de la distribution de probabilité conjointe.

L'égalité dans l'équation 3.9 fournit également une implémentation d'une méthode de calcul de la distribution de probabilité conjointe dans le cadre de notre modèle. Elle repose sur des choix simplificateurs de modélisation (les hypothèses d'indépendances conditionnelles qui viennent d'être présentées) qui permettent également d'expliciter différents sous-systèmes :

Le prior sur les objets $P(O_S)$ encode la connaissance a priori qu'a l'agent sur les objets. $P(O_S)$ prend la forme d'une distribution de probabilité uniforme pour traduire le fait que les différents objets ont la même fréquence d'apparition dans l'environnement (ou l'ignorance de cette fréquence d'apparition).

Le système moteur $P(M \mid O_S)$ décrit la connaissance qu'a l'agent sur les gestes moteurs associés aux objets. Ce terme encode le répertoire moteur de l'agent.

Le système sensori-moteur $P(S \mid M)$ décrit la connaissance qu'a l'agent sur la relation entre les gestes articulatoires et leur conséquences sensorielles. Ce terme encode un modèle interne[18] (éventuellement approximatif, incomplet, ou partiellement erroné) de la transformation articulatori-acoustique.

Le système auditif $P(O_L \mid S)$ décrit la connaissance qu'a l'agent sur la relation entre les entrées perceptives et les objets. Ce terme encode le classifieur auditif de l'agent.

Le système de validation de la communication $P(C \mid O_S\ O_L)$ décrit la connaissance qu'a l'agent sur le succès de la communication. C est une variable booléenne qui est vraie avec une probabilité 1 lorsque les objets considérés du point de vue du locuteur et de l'auditeur sont les mêmes ($O_S = O_L$), et qui est fausse sinon.

3 Inférences probabilistes dans *COSMO*

3.1 Principe de l'inférence bayésienne

Notre modèle est entièrement décrit par la distribution de probabilité conjointe sur l'ensemble de ses variables, car en effet on peut en inférer par le calcul n'importe quelle distribution de probabilité faisant intervenir un sous-ensemble de variables. Il s'agit là d'un résultat général d'inférence bayésienne, dont nous rappelons la démonstration. Supposons que l'on se donne une partition de l'ensemble des variables d'un modèle en trois ensembles : *Cherchées*, *Connues*, et *Libres*, et que l'on cherche à calculer la distribution suivante :

$$P(Cherchées \mid Connues)\ .$$

Si l'ensemble *Connues* est vide, marginaliser par rapport aux variables de l'ensemble *Libres* permet d'obtenir :

$$P(Cherchées) = \sum_{Libres} P(Cherchées\ Libres)\ , \tag{3.10}$$

ce qui montre comment calculer notre distribution à partir de la conjointe. Si maintenant l'ensemble *Connues* n'est pas vide, d'après le théorème de Bayes, on a :

$$P(Cherchées \mid Connues) = \frac{P(Cherchées\ Connues)}{P(Connues)}\ . \tag{3.11}$$

Cette égalité n'a de sens que lorsque $P(Connues)$ est non nul. En effet, dans le cas contraire, le terme $P(Cherchées \mid Connues)$ n'est pas défini : cela reviendrait à raisonner sachant qu'un événement impossible (de probabilité nulle) s'est réalisé, ce qui est absurde. Comme par ailleurs nous calculons une distribution sur les variables de l'ensemble *Cherchées*, le terme $1/P(Connues)$ est une constante dont la valeur ne nous intéresse pas. Le calcul de la distribution de probabilité $P(Cherchées \mid Connues)$ se fait à une constante de normalisation près (pouvant

[18] La notion de modèle interne est présentée en détails dans la Section 2.3.1 du Chapitre 2, page 22.

être calculée a posteriori), ce que nous indiquons en utilisant l'opérateur de proportionnalité $\propto$:

$$P(Cherchées \mid Connues) \propto P(Cherchées\ Connues)\ . \tag{3.12}$$

Alors, en marginalisant par rapport aux variables de l'ensemble *Libres*, on obtient :

$$P(Cherchées \mid Connues) \propto \sum_{Libres} P(Cherchées\ Libres\ Connues)\ , \tag{3.13}$$

ce qui achève de montrer comment on peut calculer toute distribution de probabilité à partir de la distribution de probabilité conjointe.

Il s'agit là d'un résultat très fort : à partir du choix des variables d'un modèle, et de la distribution de probabilité conjointe sur l'ensemble de ces variables, on peut calculer toutes les distributions de probabilité reliant les différentes variables entre elles[19]. Ensuite, afin de réduire la complexité des calculs d'inférence, le modélisateur ajoute des hypothèses judicieuses d'indépendance de certaines variables conditionnellement à d'autres, ce qui a pour effet de décomposer l'expression de la distribution de probabilité conjointe en un produit de termes plus simples, qui sont également plus expressifs, c'est-à-dire qu'ils ont du sens par rapport au phénomène modélisé.

Dans le cadre de notre modèle *COSMO*, grâce au choix de nos variables et au choix de la décomposition de la distribution de probabilité conjointe (voir l'équation 3.9) décrivant notre agent communicant en un prior sur les objets, un système moteur, un système perceptif, un système sensori-moteur et un système de validation, nous pouvons décrire tous les liens possibles entre objets (O_S et O_L), gestes moteurs (M), percepts sensoriels (S), et succès de la communication (C). En particulier, nous allons nous intéresser aux tâches de production et de perception de la parole, et montrer comment ces tâches peuvent toutes les deux être réalisées en faisant intervenir soit des représentations purement perceptives, soit des représentations purement motrices, soit les deux à la fois.

3.2 Implémentation probabiliste des tâches de production et de perception de la parole

Une tâche de production de la parole consiste à réaliser des gestes moteurs permettant de communiquer un objet. Cela revient à calculer un terme de la forme $P(M \mid O)$, c'est-à-dire répondre à la question « Quelle est la distribution de probabilité sur les gestes moteurs M permettant de produire l'objet O ? » De même, une tâche de perception de la parole consiste, à partir d'une entrée perceptive, à reconnaître l'objet correspondant. Cela revient à calculer un terme de la forme $P(O \mid S)$, c'est-à-dire répondre à la question « Quelle est la distribution de probabilité sur les objets O susceptibles de correspondre à une entrée perceptive S ? »

Dans le cadre de *COSMO*, il y a deux représentations distinctes des objets de la communication : l'une, O_S, du point de vue du locuteur, et l'autre, O_L, du point de vue de l'auditeur. Nous proposons donc trois manières différentes d'aborder les tâches de production et de perception :

[19] L'équation 3.13 donne une méthode générale de calcul exact, mais qui en pratique n'est pas calculable dans un temps raisonnable : Cooper (1990) a montré que le problème de l'inférence probabiliste dans les réseaux bayésiens est un problème $\mathcal{NP} - hard$. On ne peut donc pas le traiter sans faire des hypothèses de modélisation simplificatrices, ou/et utiliser des techniques sophistiquées d'optimisation ou d'approximation du calcul, comme les méthodes de Monte-Carlo par exemple (Metropolis et Ulam, 1949).

1. Adopter un point de vue centré sur le rôle de locuteur revient à se focaliser sur l'objet O_S, ce qui est l'approche suivie par les théories motrices.

2. Adopter un point de vue centré sur le rôle d'auditeur revient à se focaliser sur l'objet O_L, ce qui est l'approche suivie par les théories perceptives.

3. Adopter un point de vue centré sur les rôles de locuteur autant que d'auditeur, revient à se focaliser autant sur O_S que sur O_L, ce qui est l'approche suivie par les théories perceptuo-motrices. Dans le cadre de *COSMO*, cela est fait en utilisant la variable C : conditionner une distribution de probabilité par $[C = 1]$ impose l'égalité entre O_S et O_L pendant l'inférence, et matérialise le fait que l'on se donne comme objectif explicite le succès de la communication.

Ainsi, effectuer avec *COSMO* une tâche de production dans le cadre d'une théorie motrice revient à calculer la distribution de probabilité conditionnelle $P(M \mid O_S)$. Dans le cadre d'une théorie perceptive, cela s'instancie $P(M \mid O_L)$, et dans le cadre d'une théorie perceptuo-motrice, on peut calculer indifféremment $P(M \mid O_S\ [C{=}1])$ ou $P(M \mid O_L\ [C{=}1])$ puisque l'on peut démontrer que ces deux distributions de probabilité sont identiques.

De même, effectuer avec *COSMO* une tâche de perception dans le cadre d'une théorie motrice revient à calculer la distribution de probabilité conditionnelle $P(O_S \mid S)$. Dans le cadre d'une théorie perceptive, cela s'instancie $P(O_L \mid S)$, et dans le cadre d'une théorie perceptuo-motrice, on peut calculer indifféremment $P(O_S \mid S\ [C{=}1])$ ou $P(O_L \mid S\ [C{=}1])$ puisque l'on peut démontrer que ces deux distributions de probabilité sont identiques.

3.3 Inférences probabilistes pour les tâches de perception

Comme les travaux présentés dans cette thèse se sont concentrés sur l'étude des interactions sensori-motrices en perception, nous nous contentons de détailler ici les calculs d'inférence pour les tâches de perception, bien qu'une démarche analogue permette d'obtenir les inférences correspondant aux tâches de production, pour lesquelles nous ne présenterons que le résultat des calculs.

3.3.1 Implémentation dans le cadre d'une théorie motrice

Réaliser une tâche de perception dans le cadre d'une théorie motrice, c'est calculer la distribution de probabilité conditionnelle $P(O_S \mid S)$. D'après le théorème de Bayes, on a :

$$P(O_S \mid S) = \frac{P(O_S\ S)}{P(S)}\ . \tag{3.14}$$

Il s'agit d'un calcul de distribution sur O_S, dans lequel le terme $P(S)$ est une constante. Ce terme permet d'assurer la normalisation de la distribution de probabilité conditionnelle $P(O_S \mid S)$, et on peut se contenter de le calculer a posteriori. On peut donc écrire, en utilisant l'opérateur de proportionalité $\propto$:

$$P(O_S \mid S) \propto P(O_S\ S)\ . \tag{3.15}$$

En marginalisant par rapport aux variables libres, on obtient :

$$P(O_S \mid S) \propto \sum_{M\ O_L\ C} P(O_S\ M\ S\ O_L\ C)\ . \tag{3.16}$$

En utilisant maintenant la décomposition de la distribution de probabilité conjointe, cela donne :

$$P(O_S \mid S) \propto \sum_M \sum_{O_L} \sum_C P(O_S)P(M \mid O_S)P(S \mid M)P(O_L \mid S)P(C \mid O_S\ O_L)\ . \tag{3.17}$$

En factorisant ce qui peut l'être, il appert :

$$P(O_S \mid S) \propto P(O_S) \sum_M \left(P(M \mid O_S)P(S \mid M) \sum_{O_L} \left(P(O_L \mid S) \sum_C P(C \mid O_S\ O_L) \right) \right)\ . \tag{3.18}$$

Il est possible de simplifier cette expression en utilisant la règle de normalisation deux fois, sur C puis sur O_L :

$$P(O_S \mid S) \propto P(O_S) \sum_M \Big(P(M \mid O_S)P(S \mid M) \Big)\ . \tag{3.19}$$

Comme les objets suivent une distribution de probabilité uniforme, le terme $P(O_S)$ est constant, et peut donc être absorbé par l'opérateur de proportionalité :

$$P(O_S \mid S) \propto \sum_M \Big(P(M \mid O_S)P(S \mid M) \Big)\ . \tag{3.20}$$

Ainsi, nous avons mis en évidence le fait qu'une théorie motrice de la perception fait intervenir deux parties de notre modèle : le système moteur $P(M \mid O_S)$ et le lien sensori-moteur $P(S \mid M)$. En sommant sur tous les gestes moteurs M susceptibles d'avoir produit l'entrée perceptive S, notre implémentation d'une théorie motrice de la perception fait de l'analyse par la synthèse en combinant un modèle direct $P(S \mid M)$ de la transformation articulatori-acoustique avec un décodeur articulatoire[20] $P(M \mid O_S)$.

De plus, nous tenons à souligner que, dans le cadre de *COSMO*, le problème de l'inversion motrice se résoud assez naturellement. Rappelons que, le système moteur étant par nature redondant, plusieurs gestes moteurs différents produisent les mêmes conséquences perceptives, et qu'ainsi certains travaux de modélisation implémentent un processus élaboré de sélection du « bon » geste. Dans *COSMO*, l'équation (3.20) montre que tous les gestes moteurs sont considérés, pondérés par leur probabilité de correspondre à l'objet d'une part, et au percept sensoriel d'autre part.

3.3.2 Implémentation dans le cadre d'une théorie auditive

Réaliser une tâche de perception dans le cadre d'une théorie auditive, c'est calculer la distribution de probabilité conditionnelle $P(O_L \mid S)$. En marginalisant par rapport aux variables libres, on obtient :

$$P(O_L \mid S) = \sum_{O_S\ M\ C} P(O_S\ M\ O_L\ C \mid S)\ . \tag{3.21}$$

D'après le théorème de Bayes, on en déduit :

$$P(O_L \mid S) = \frac{\sum_{O_S\ M\ C} P(O_S\ M\ S\ O_L\ C)}{P(S)}\ . \tag{3.22}$$

[20] Il est bon de souligner qu'ici le terme de *décodeur articulatoire* décrit la fonction remplie par le terme $P(M \mid O_S)$ lors d'une tâche de perception, même si le fait que le lien entre objets et gestes moteurs soit exprimé dans le modèle sous forme d'une distribution de probabilité conditionnelle de la forme $P(M \mid O_S)$ plutôt que $P(O_S \mid M)$ serait sans doute plus justement traduit par le terme d'*encodeur articulatoire*.

En applicant la règle du produit deux fois au numérateur, il vient :

$$P(O_L \mid S) = \frac{\sum_{O_S\ M\ C} P(O_S\ M\ S)P(O_L \mid S\ M\ O_S)P(C \mid O_L\ S\ M\ O_S)}{P(S)} \ . \tag{3.23}$$

En utilisant maintenant deux des hypothèses qui ont permis de simplifier l'expression de la distribution de probabilité conjointe (voir équations (3.7) et (3.8)), cela donne :

$$P(O_L \mid S) = \frac{\sum_{O_S\ M\ C} P(O_S\ M\ S)P(O_L \mid S)P(C \mid O_S\ O_L)}{P(S)} \ . \tag{3.24}$$

En factorisant ce qui peut l'être, il appert :

$$P(O_L \mid S) = \frac{P(O_L \mid S) \sum_{O_S\ M} \left(P(O_S\ M\ S) \sum_{C} P(C \mid O_S\ O_L) \right)}{P(S)} \ . \tag{3.25}$$

Il est possible de simplifier cette expression en utilisant la règle de normalisation au numérateur :

$$P(O_L \mid S) = \frac{P(O_L \mid S) \sum_{O_S\ M} P(O_S\ M\ S)}{P(S)} \ . \tag{3.26}$$

On peut alors d'appliquer la règle de marginalisation au numérateur :

$$P(O_L \mid S) = \frac{P(O_L \mid S)P(S)}{P(S)} \ . \tag{3.27}$$

Ceci peut alors se simplifier puisque le numérateur et le dénominateur se compensent :

$$P(O_L \mid S) = P(O_L \mid S) \ . \tag{3.28}$$

Il est légitime de se demander si il est nécessaire de passer par toutes ces étapes de calcul pour en arriver à ce qui semble être une tautologie. Il faut donc noter que le = est ici un opérateur d'implémentation, dont le sens est que la tâche en partie gauche est calculée en utilisant le terme en partie droite. Cette égalité qui semble évidente est en fait un résultat général d'inférence bayésienne : lorsque l'on calcule une tâche dont l'expression est un terme qui intervient dans la décompositon de la distribution de probabilité conjointe, l'inférence conduit à une égalité du type x=x.

Ce que nous venons de montrer avec l'équation (3.28), c'est que la tâche de perception de parole dans le cadre d'une théorie purement auditive est directement calculée par l'un des termes apparaissant dans la décomposition de la distribution de probabilité conjointe du modèle : le système perceptif $P(O_L \mid S)$. Une théorie auditive de la perception se limite donc à l'exploitation d'un classifieur auditif : elle ne fait intervenir que le lien entre les représentations sensorielles S et les objets O_L.

3.3.3 Implémentation dans le cadre d'une théorie perceptuo-motrice

Réaliser une tâche de perception dans le cadre d'une théorie perceptuo-motrice, c'est calculer indifféremment $P(O_L \mid S\ [C{=}1])$ ou $P(O_S \mid S\ [C{=}1])$.

D'après le théorème de Bayes, on a :

$$P(O_S \mid S\ \ [C{=}1]) = \frac{P(O_S\ S\ \ [C{=}1])}{P(S\ \ [C{=}1])}\ . \tag{3.29}$$

Il s'agit d'un calcul de distribution sur O_S, dans lequel le terme $P(S\ \ [C{=}1])$ est une constante. Ce terme permet d'assurer la normalisation de la distribution de probabilité conditionnelle $P(O_S \mid S\ \ [C{=}1])$, et on peut se contenter de le calculer a posteriori. On peut donc écrire, en utilisant l'opérateur de proportionalité $\propto$:

$$P(O_S \mid S\ \ [C{=}1]) \propto P(O_S\ S\ \ [C{=}1])\ . \tag{3.30}$$

En marginalisant par rapport aux variables libres, on obtient :

$$P(O_S \mid S\ \ [C{=}1]) \propto \sum_{M\ O_L} P(O_S\ M\ S\ O_L\ \ [C{=}1])\ . \tag{3.31}$$

En utilisant maintenant la décomposition de la distribution de probabilité conjointe, cela donne :

$$P(O_S \mid S\ \ [C{=}1]) \propto \sum_{M}\sum_{O_L} \Big(P(O_S)P(M \mid O_S)P(S \mid M)P(O_L \mid S)P([C{=}1]\ \mid O_S\ O_L)\Big)\ . \tag{3.32}$$

Puisque la distribution de probabilité conditionnelle $P([C{=}1]\ \mid O_S\ O_L)$ vaut 1 pour $O_S{=}O_L$ et vaut 0 partout ailleurs, avoir la contrainte $C{=}1$ impose l'égalité $P(O_S|\,[C{=}1]) = P(O_L|\,[C{=}1])$. On note $[O_L{=}O_S]$ pour référer indifféremment à l'un ou l'autre des objets, qui ont la même distribution de probabilité. On peut alors simplifier la somme sur les objets O_L en retirant tous les termes nuls :

$$P(O_S \mid S\ \ [C{=}1]) \propto \sum_{M} \Big(P(O_S)P(M \mid O_S)P(S \mid M)P([O_L{=}O_S]\ \mid S)\Big)\ . \tag{3.33}$$

En factorisant ce qui ne dépend pas de M, il appert :

$$P(O_S \mid S\ \ [C{=}1]) \propto P(O_S)P([O_L{=}O_S]\ \mid S)\sum_{M} \Big(P(M \mid O_S)P(S \mid M)\Big)\ . \tag{3.34}$$

Comme les objets suivent une distribution de probabilité uniforme, le terme $P(O_S)$ est constant, et peut donc être absorbé par l'opérateur de proportionalité :

$$P(O_S \mid S\ \ [C{=}1]) \propto P([O_L{=}O_S]\ \mid S)\sum_{M} \Big(P(M \mid O_S)P(S \mid M)\Big)\ . \tag{3.35}$$

Ce que nous venons de montrer, c'est que le calcul de la tâche de perception de la parole dans le cadre d'une théorie perceptuo-motrice fait intervenir trois parties de notre modèle : le système perceptif $P(O_L \mid S)$, le système moteur $P(M \mid O_S)$ et le lien sensori-moteur $P(S \mid M)$. De plus, la partie droite de l'équation (3.35) n'est autre que le produit des parties droites des équations (3.28) et (3.20), c'est-à-dire que la perception perceptuo-motrice apparaît comme étant le produit d'une perception motrice et d'une perception auditive.

Ainsi, l'utilisation de notre variable de cohérence C, qui traduisait au départ l'internalisation du succès de la communication, permet également d'exprimer la tâche de perception perceptuo-motrice sous forme de fusion bayésienne de capteurs : on tire parti à la fois de l'information apportée par le système moteur et par le système perceptif.

3.4 Inférences probabilistes pour les tâches de production

Les calculs d'inférence pour les tâches de production se font de la même manière, *mutatis mutandis*, que ceux qui sont présentés section 3.3 pour les tâches de perception et ne sont pas détaillés ici. Nous présentons en revanche les résultats de ces calculs et leur interprétation.

3.4.1 Implémentation dans le cadre d'une théorie motrice

Réaliser une tâche de production dans le cadre d'une théorie motrice, c'est calculer la distribution de probabilité conditionnelle $P(M \mid O_S)$. De manière similaire au calcul de la section 3.3.2, le terme $P(M \mid O_S)$ apparaissant déjà dans la décomposition de la distribution de probabilité conjointe, le calcul conduira de manière rassurante à l'égalité suivante :

$$P(M \mid O_S) = P(M \mid O_S) \ . \tag{3.36}$$

Ainsi, la tâche de production de parole dans le cadre d'une théorie purement motrice ne fait intervenir que le système moteur $P(M \mid O_S)$. Une théorie motrice de la production se limite donc à l'exploitation d'un répertoire moteur : elle ne fait intervenir que le lien entre les objets à produire O_S et les gestes M correspondants.

3.4.2 Implémentation dans le cadre d'une théorie auditive

Réaliser une tâche de production dans le cadre d'une théorie auditive, c'est calculer la distribution de probabilité conditionnelle $P(M \mid O_L)$. Par inférence bayésienne sur la distribution de probabilité conjointe, on obtient le résultat suivant :

$$P(M \mid O_L) \propto P(M) \sum_S \Big(P(S \mid M) P(O_L \mid S) \Big) \ . \tag{3.37}$$

Ainsi, la tâche de production dans le cadre d'une théorie auditive fait intervenir deux parties de notre modèle : le système perceptif $P(O_L \mid S)$ et le lien sensori-motreur $P(S \mid M)$, ainsi qu'une distribution a priori $P(M)$ sur les gestes moteurs. En sommant sur tous les percepts sensoriels S susceptibles de correspondre à l'objet à produire O_L, notre théorie auditive de la production fait de l'analyse par la synthèse en combinant des cibles acoustiques $P(O_L \mid S)$ avec un modèle direct $P(S \mid M)$ de la transformation articulatori-acoustique.

3.4.3 Implémentation dans le cadre d'une théorie perceptuo-motrice

Réaliser une tâche de production dans le cadre d'une théorie perceptuo-motrice, c'est calculer indifféremment $P(M \mid O_S\ [C{=}1])$ ou $P(M \mid O_L\ [C{=}1])$. Par inférence bayésienne sur la distribution de probabilité conjointe, il vient le résultat suivant :

$$P(M \mid O_L\ [C{=}1]) \propto P(M \mid\ [O_S{=}O_L]) \sum_S \Big(P(S \mid M) P(O_L \mid S) \Big) \ . \tag{3.38}$$

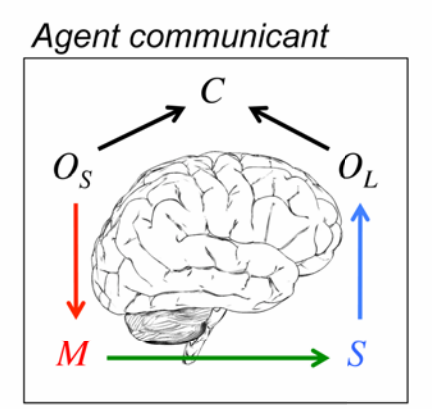
Ainsi, la tâche de production de la parole dans le cadre d'une théorie perceptuo-motrice fait intervenir trois parties de notre modèle : le système moteur $P(M \mid O_S)$, le système perceptif $P(O_L \mid S)$ et le lien sensori-moteur $P(S \mid M)$. De plus, la partie droite de l'équation (3.38) n'est autre que le produit des parties droites des équations (3.36) et (3.37), c'est-à-dire que

la production perceptuo-motrice apparaît comme étant le produit d'une production faisant intervenir des représentations purement motrices et d'une production faisant intervenir des représentations purement auditives.

3.5 Résumé des calculs d'inférence

Ainsi, notre modèle *COSMO* est un modèle probabiliste d'agent communicant qui, construit sur une hypothèse d'internalisation de la situation de communication, offre un cadre computationnel intégrateur permettant d'étudier les interactions sensori-motrices en production et en perception de la parole.

Dans un souci de synthèse, voici sur des pages voisines le modèle *COSMO* décrit figure 3.4 par un réseau bayésien et par la décomposition de sa distribution de probabilité conjointe, et la figure 3.5 qui résume tous les calculs d'inférence précédents.



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

Figure 3.4: **Le modèle *COSMO*** d'agent communicant, décrit par un réseau bayésien et par la décomposition de sa distribution de probabilité conjointe.

Cette figure est à mettre en parallèle avec la figure 3.6, déjà présentée dans le chapitre précédant à la section 2.1.3, qui propose une taxonomie des différentes théories de production et de perception de la parole, illlustrée par des travaux emblématique de chaque entrée.

C'est une propriété majeure de *COSMO* que de proposer de décrire, dans un même cadre théorique et dans un même formalisme probabiliste, l'ensemble des théories auditives, motrices et perceptuo-motrices de la communication parlée, dans des enjeux de production comme de perception. C'est cette propriété majeure que nous allons exploiter tout au long de cette thèse, pour comparer ces différentes théories, notamment dans le contexte de la perception de la parole. Dans le reste de ce document, nous utiliserons le terme (singulier) de théorie auditive, théorie motrice et théorie perceptuo-motrice au sens d'unicité que confèrent les différentes équations pour *une* théorie donnée – sans perdre de vue qu'il existe, indépendamment de notre formalisme unificateur, *des* implémentations variées dans chaque cadre proposé.

	Tâche de production inférence de la forme $P(M \mid O)$	Tâche de perception inférence de la forme $P(O \mid S)$
Théorie motrice focalisation sur O_S	$\underbrace{P(M \mid O_S)}_{\text{répertoire moteur}}$	$\propto \sum_M \Big(\underbrace{P(M \mid O_S)}_{\text{décodeur articulatoire}} \times \underbrace{P(S \mid M)}_{\text{modèle inverse}} \Big)$
Théorie auditive focalisation sur O_L	$\propto P(M) \sum_S \Big(\underbrace{P(S \mid M)}_{\text{modèle direct}} \times \underbrace{P(O_L \mid S)}_{\text{cibles acoustiques}} \Big)$	$\underbrace{P(O_L \mid S)}_{\text{classifieur auditif}}$
Théorie perceptuo-motrice C=1, i.e. O_S=O_L	$\propto \underbrace{P(M \mid [O_S{=}O_L])}_{\text{production motrice}} \underbrace{\sum_S \big(P(S \mid M)P(O_L \mid S)\big)}_{\text{production auditive}}$	$\propto \underbrace{P([O_L{=}O_S] \mid S)}_{\text{perception auditive}} \underbrace{\sum_M \big(P(M \mid O_S)P(S \mid M)\big)}_{\text{perception motrice}}$

Figure 3.5: **Inférences probabilistes dans *COSMO***, pour les tâches de production et de peception prédites par les théories motrice, perceptive et perceptuo-motrices.

	Tâche	
Théorie	Production	Perception
Motrice	Articulatory Phonology, Browman et Goldstein (1989)	Motor Theory, Liberman et Mattingly (1985)
Auditive	Auditory reference frames for speech planning, Guenther *et al.* (1998)	Auditory theories, Diehl *et al.* (2004)
Perceptuo-motrice	DIVA model, Guenther (2006)	Perception for Action Control Theory, Schwartz *et al.* (2012a)

Figure 3.6: **Taxonomie des modèles de production et de perception de la parole**, adaptée de Moulin-Frier *et al.* (2010). L'objectif de cette table n'est bien sûr pas l'exhaustivité, mais il s'agit simplement d'illustrer la manière dont les tâches de production et de perception sont vues par des travaux représentatifs des différentes familles de théories.

4 Théorème d'indistinguabilité des théories motrice et auditive en perception de la parole

Dans cette partie nous allons montrer que, bien que les théories motrice et auditive de la perception de la parole soient régies par des équations différentes (obtenues par les calculs d'inférence de la section précédente), il existe des cas où ces théories sont indistinguables[21].

Nous donnons ici une description informelle de notre théorème d'indistinguabilité, qui sera démontré rigoureusement plus loin. Ce théorème pose un cadre, constitué de conditions idéales, sous lesquelles il n'est pas possible de distinguer les prédictions des théories motrice et auditive de la perception car elles sont identiques. Ces conditions idéales dont la conjonction suffit à assurer l'indistinguabilité sont les suivantes :

- apprentissage parfait du classifieur auditif à partir des productions d'un maître ;

[21] Nous qualifions d'indistinguables des théories qui ne peuvent pas être distinguées, au sens où elles font rigoureusement les mêmes prédictions expérimentales, ce qui garantit l'impossibilité d'observer des différences.

- identité motrice parfaite avec le maître ;
- connaissance parfaite de la transformation articulatori-acoustique.

Lorsque ces hypothèses caractérisant des conditions idéales d'apprentissage sont vérifiées, les théories motrice et auditive font les mêmes prédictions quelle que soit la tâche de perception demandée ensuite à notre modèle d'agent (différents locuteurs, différents niveaux de bruit, etc).

La démonstration du théorème d'indistinguabilité et l'explicitation des conditions parfaites sur lesquelles il repose nécessitent au préalable d'expliquer comment sont apprises les différentes distributions de probabilité intervenant dans la décomposition de la distribution de probabilité conjointe (voir par exemple figure 3.4) de notre modèle *COSMO*.

4.1 Apprentissage par interactions avec un agent maître

On considère un scénario d'apprentissage supervisé faisant intervenir deux agents : un agent maître et un agent apprenant qui interagissent au sein d'un environnement.

La programmation bayésienne fournit un cadre dans lequel il est possible d'utiliser une variable de modèle (Colas *et al.*, 2010), généralement notée π, pour représenter formellement les hypothèses de modélisation conduisant à des états de connaissance différents. Puisque le modèle est par nature dépendant des choix du modélisateur, dans le cadre de la programmation bayésienne (Lebeltel *et al.*, 2004) toute distribution de probabilité exprimée dans le cadre d'un modèle M doit être conditionnée par une variable π_M, qui représente les choix de modélisation.

Par souci de lisibilité, nous n'avons pas noté systématiquement de variable π dans les calculs d'inférence de la section 3.5 car il n'y avait pas d'ambiguïté. En revanche, dans le scénario d'apprentissage supervisé que nous proposons maintenant, notre modèle *COSMO* est instancié dans deux agents différents : un agent maître $\pi_{Maître}$ et un agent apprenant π_{Ag}, qui interagissent dans un environnement π_{Env}.

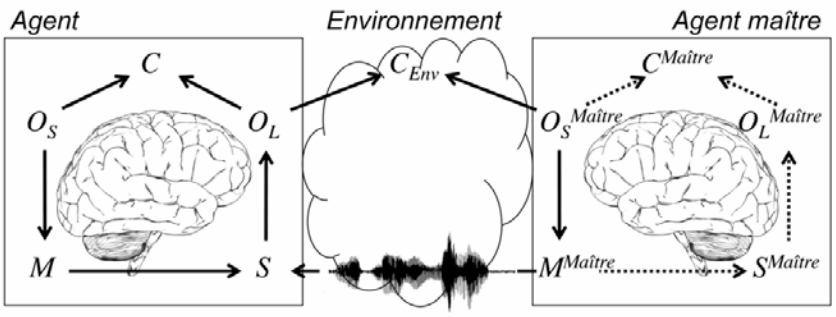


Figure 3.7: **Scénario d'apprentissage supervisé** : L'agent π_{Ag} interagit avec l'agent maître $\pi_{Maître}$ au sein de l'environnement π_{Env}.

Dans le scénario d'apprentissage supervisé présenté figure 3.7, l'agent π_{Ag} apprend son classifieur auditif $P(O_L \mid S\ \pi_{Ag})$ à partir des productions d'un maître. Le principe de l'apprentissage est que l'agent apprend le lien entre des stimuli s, générés par le maître (dans des conditions décrites par le modèle du maître $\pi_{Maître}$ et par le modèle de l'environnement π_{Env}), et les objets o_l qui, grâce à un mécanisme d'attention partagée, tel que la deixis par exemple, sont identiques

à ceux produits par le maître (c'est-à-dire $O_S^{Maître}$=O_L). L'algorithme d'apprentissage consiste alors à itérer les étapes suivantes.

- L'agent maître choisit aléatoirement des objets o_s selon la distribution de probabilité uniforme $P(O_S^{Maître} \mid \pi_{Maître})$.
- Pour chacune de ces valeurs de o_s, l'agent maître choisit une consigne motrice m selon la distribution de probabilité $P\left(M^{Maître} \mid \left[O_S^{Maître}{=}o_s\right] \ \pi_{Maître}\right)$ correspondante.
- Ces consignes motrices m sont transformées par l'environnement et parviennent à l'agent apprenant sous forme d'entrées perceptives s. Cette transformation est décrite par la distribution de probabilité conditionnelle $P\left(S \mid \left[M^{Maître}{=}m\right] \ \pi_{Env}\right)$ qui encode deux choses : le passage d'une consigne motrice vers un signal acoustique (on considère donc que le conduit vocal du maître, transformant une commande motrice en son, fait partie de l'environnement) et des perturbations éventuelles de ce signal (par exemple du bruit de communication).
- L'agent maître transmet également à l'agent apprenant la valeur de l'objet o_s grâce à un mécanisme d'attention partagée, tel que la deixis par exemple. Dans le modèle de l'environnement, c'est la variable C_{Env} qui permet d'assurer la cohérence entre le o_s fourni par le maître et le o_l de l'agent apprenant.
- Ainsi, à chaque itération, l'agent reçoit un couple de valeurs $< o_l, \ s >$ qu'il utilise pour mettre à jour la distribution de probabilité conditionnelle $P\left(O_L \mid [S{=}s] \ \pi_{Ag}\right)$ de son classifieur auditif avec l'observation que le stimulus s peut correspondre à l'objet o_l.

L'algorithme que nous venons de présenter pour l'apprentissage du classifieur auditif de l'agent à partir des productions du maître procède par une succession de tirages aléatoires selon les distributions de probabilité du modèle du maître $\pi_{Maître}$ ou du modèle de l'environnement π_{Env}. Plus précisément, ce que l'agent apprend comme classifieur auditif est une approximation asymptotique par des tirages successifs (c'est-à-dire une approximation de Monte-Carlo) d'une distribution de probabilité dont on peut par des calculs d'inférence obtenir la forme exacte :

$$P(O_L \mid S \ \pi_{Ag}) \ \approx \sum_{M^{Maître}} \Big(P(M^{Maître} \mid O_S^{Maître} \ \pi_{Maître}) \times P(S \mid M^{Maître} \ \pi_{Env})\Big) \ . \quad (3.39)$$

C'est à partir de ce scénario d'apprentissage supervisé, selon lequel le classifieur auditif de l'agent est appris à partir des productions motrices du maître, et en ajoutant des hypothèses de conditions idéales d'apprentissage et de similitude parfaite des systèmes moteurs du maître et de l'agent, que nous allons prouver l'indistinguabilité des théories motrice et auditive de la perception de la parole.

4.2 Conditions suffisantes d'indistinguabilité

Nous décrivons ici trois hypothèses **H1**, **H2** et **H3**, correspondant aux « conditions idéales » introduites précédemment, et qui, prises ensemble, suffisent à garantir l'indistinguabilité des théories motrice et auditive de la perception de la parole implémentées dans notre modèle *COSMO*.

- Nous supposons que le système perceptif est parfaitement appris à partir des productions motrices d'un maître.

  L'algorithme d'apprentissage supervisé présenté à la section 4.1 permet d'apprendre le classifeur auditif grâce à des tirages aléatoires qui réalisent une approximation présentée à l'équation 3.39. La première hypothèse consiste alors uniquement à remplacer cette approximation par une égalité :

  $$\textbf{H1} \ : \ P(O_L \mid S\, \pi_{Ag}) = \sum_{M^{Maître}} \Big(P(M^{Maître} \mid O_S^{Maître}\ \pi_{Maître}) \times P(S \mid M^{Maître}\ \pi_{Env}) \Big) \ . \tag{3.40}$$

  C'est-à-dire qu'au lieu de procéder à des tirages aléatoires intermédiaires, on définit le terme $P(O_L \mid S\ \pi_{Ag})$ comme encodant parfaitement toute la richesse d'information exprimée par les distributions de probabilité spécifiant la manière dont le maître choisit ses gestes moteurs, et la façon dont ceux-ci sont transformés par l'environnement. $H1$ est donc une hypothèse qui formalise le caractère exhaustif et parfait de l'apprentissage.

- Nous supposons que l'agent et le maître ont des répertoires moteurs identiques.

  $$\textbf{H2} \ : \ P(M \mid O_S\ \pi_{Ag}) = P(M^{Maître} \mid O_S^{Maître}\ \pi_{Maître}) \ . \tag{3.41}$$

- Nous supposons que l'agent a parfaitement capturé les propriétés de l'environnement au moment de l'apprentissage dans sa représentation interne de la transformation articulatori-acoustique.

  $$\textbf{H3} \ : \ P(S \mid M\ \pi_{Ag}) = P(S \mid M^{Maître}\ \pi_{Env}) \ . \tag{3.42}$$

Théorème : *Si les hypothèses **H1**, **H2** et **H3** sont vérifiées, c'est-à-dire si le système perceptif est parfaitement appris à partir des productions motrices d'un maître, si de plus l'agent apprenant et l'agent maître ont des répertoires moteurs identiques, et si enfin l'agent a parfaitement capturé dans sa représentation interne de la transformation articulatori-acoustique les propriétés de l'environnement au moment de l'apprentissage, alors les instanciations motrice et auditive de la tâche de perception sont indistinguables.*

Preuve du théorème d'indistinguabilité :
D'après l'équation (3.28), la tâche de perception exprimée dans le cadre d'une théorie purement auditive se limite à l'exploitation du système auditif $P(O_L \mid S\ \pi_{Ag})$. D'après l'hypothèse **H1** formalisée par l'équation 3.40, la tâche de perception se réécrit ainsi :

$$P(O_L \mid S\ \pi_{Ag}) = \sum_{M^{Maître}} \Big(P(M^{Maître} \mid O_S^{Maître}\ \pi_{Maître}) \times P(S \mid M^{Maître}\ \pi_{Env}) \Big) \ . \tag{3.43}$$

En remplaçant les deux termes en partie droite de cette équation grâce aux hypothèses **H2** et **H3** formalisées par les équations 3.41 et 3.42, il appert :

$$P(O_L \mid S\ \pi_{Ag}) = \sum_{M} \left(P(M \mid O_S\ \pi_{Ag}) \times P(S \mid M\ \pi_{Ag}) \right) \ . \tag{3.44}$$

Dans cette équation 3.44, le terme de gauche est l'expression d'une théorie purement auditive de la perception (voir l'équation 3.28) et le terme de droite est l'expression d'une théorie purement

motrice de la perception (voir l'équation 3.20). Les hypothèses **H1**, **H2** et **H3** sont donc suffisantes pour garantir que les inférences motrice et auditive de la tâche de perception conduisent au même résultat.

□

Nous venons de prouver que sous des hypothèses d'apprentissage parfait, d'internalisation parfaite de la transformation articulatori-acoustique et des propriétés de l'environnement, et d'identité motrice entre le maître et l'agent, les théories motrices et auditives de la perception sont alors régies par les mêmes équations. Elles font donc les mêmes prédictions, et sont de ce fait indistinguables sous ces conditions.

4.3 Discussion

Les théories motrice et auditive de la perception sont-elles pour autant toujours indistinguables ? Évidemment non. Elles le sont lorsque les trois hypothèses sont vérifiées, ce qui n'est *a priori* pas vrai en pratique dans la plupart des cas. Pour pouvoir distinguer les théories motrices des théories auditives, il est nécessaire qu'au moins l'une des trois hypothèses ne soit pas vérifiée. Considérons ces hypothèses afin d'étudier leur plausibilité en pratique.

H1 est une hypothèse qui porte sur la manière dont l'agent apprend son système perceptif. Remplacer le signe $\approx$ de l'équation 3.39 par le signe $=$ pour obtenir **H1** revient à effectuer une infinité de tirages. L'hypothèse **H1** suppose donc que l'agent a reçu une infinité d'exemplaires d'apprentissage et les conserve, ce qui lui permet de capturer parfaitement le lien entre les objets et les signaux produits par le maître. De manière implicite, **H1** suppose également que la représentation interne de l'agent est suffisamment expressive pour capturer toute la richesse de l'information que lui fournit le maître.

H2 est une hypothèse qui porte sur le système moteur de l'agent. **H2** suppose qu'il y ait une identité motrice parfaite entre le maître et l'agent. Dans notre description de l'apprentissage, nous avons dit que le maître produisait selon une théorie motrice pure de la production. Il n'est en fait pas nécessaire de se limiter à ce cas, et le théorème d'indistinguabilité est également vérifié lorsque, quelle que soit la manière de produire $P(M^{Maître} \mid O_S^{Maître}\ \pi_{Maître})$ du maître, l'agent encode exactement la même information dans son système moteur $P(M \mid O_S\ \pi_{Ag})$.

En pratique deux individus ne sont jamais exactement identiques : chacun a un conduit vocal qui lui est propre, et de plus chacun produit à sa manière, avec ses gestes idiosyncrasiques.

H3 est une hypothèse qui porte sur la manière dont l'agent capture dans son modèle interne une représentation du phénomène physique qui relie l'articulatoire à l'acoustique et des perturbations éventuelles dues à la propagation du signal par l'environnement. **H3** suppose que cette internalisation est parfaite. En pratique ce n'est pas le cas : il n'y a aucune raison de penser que les modèles internes soient en mesure de capturer toute la complexité des phénomènes physiques en jeu, ce qui nécessiterait en particulier une connaissance parfaite de l'évolution au cours du temps de l'état exact de chacun des muscles du conduit vocal et de l'ensemble des propriétés de l'environnement.

Finalement, ces trois hypothèses sur lesquelles reposent le théorème d'indistinguabilité ne font que fournir formellement une idéalisation des conditions d'apprentissage, et permettent de mieux comprendre la nature de l'information apprise.

Le théorème d'indistinguabilité formalise de manière rigoureuse un argument avancé par les tenants des théories auditives (Diehl *et al.*, 2004, page 168) : « L'auditeur ne recouvre pas les gestes, mais il perçoit les conséquences acoustiques des gestes. Toute régularité de la production de la parole (comme par exemple les effets de contexte) se reflète dans le signal acoustique, et, par des mécanismes généraux d'apprentissage perceptif, l'auditeur en vient à savoir tirer profit des corrélats acoustiques de ces régularités de production pour déterminer le contenu phonétique du signal de parole. » [22]

L'argument est que, comme la perception auditive est apprise à partir de productions motrices, on sait reconnaître toutes les conséquences acoustiques des spécificités de la production. Dit autrement, toute l'information est présente dans le signal de parole et peut être apprise. Notre travail avec le modèle *COSMO* décrit de manière formelle de quelle information il s'agit, et comment elle peut être apprise.

Diehl *et al.* poussent le raisonnement plus loin : « En soi, la forte corrélation existant entre la production et la perception n'apporte rien au débat opposant la théorie motrice, la théorie du réalisme direct et les approches auditives générales de la perception de la parole. Elles prédisent toutes les trois l'existence d'une telle corrélation. Les distinguer empiriquement nécessite l'utilisation d'autres types de données [...] » (Diehl *et al.*, 2004, pages 168). [23]

Diehl *et al.* mettent en avant la difficulté de distinguer théories motrices et théories auditives. C'est également ce que l'on montre dans le cadre de *COSMO* avec le théorème d'indistinguabilité qui met en évidence des conditions parfaites dans lesquelles les deux ne peuvent être séparées.

En revanche, alors que les « autres types de données » dont parlent Diehl *et al.* ne leur ont permis que de réfuter l'hypothèse selon laquelle la parole serait spéciale et sur laquelle repose la théorie motrice de la perception, notre approche, ayant formalisé les conditions d'indistinguabilité, nous permet également de savoir comment sortir de ces conditions pour comparer efficacement les implémentations purement motrice et purement perceptive de la perception de la parole.

5 Conclusion

Dans ce chapitre, nous avons présenté l'élaboration, dans le cadre formel de la programmation bayésienne, de *COSMO*, notre modèle d'agent communicant qui constitue un cadre computationnel intégrateur permettant d'étudier et de comparer quantitativement les théories motrice, auditive et perceptuo-motrice de la parole sur des tâches de production ou de perception. Plus précisément, c'est la formalisation d'hypothèses cognitives (sur l'internalisation, les notions de

[22] Le texte original est en anglais : "Listeners do not recover gestures, but they do perceive the acoustic consequences of gestures. Any regularities of speech production (e.g., context dependencies) will be reflected in the acoustic signal, and, through general mechanisms of perceptual learning, listeners come to make use of the acoustic correlates of these production regularities in judging the phonemic content of speech signals".

[23] "By itself, the high correlation between speech production and perception is uninformative with respect to the debate between MT, DRT, and GA. All three predict that such a correlation must exist. Distinguishing them empirically requires other kinds of data [...]".

modèles internes, les conditions de communication) qui permet d'obtenir un modèle explicite et précis : *COSMO*.

Dans ce chapitre notre approche formelle nous a conduit à un premier résultat fort : un théorème d'indistinguabilité des théories motrice et auditive en perception de la parole. Cela explique qu'il soit si difficile de démêler ces théories : il existe des cas où les données expérimentales ne peuvent pas trancher, des conditions dans lesquelles ces théories sont indistinguables et font les mêmes prédictions.

Cependant, les hypothèses **H1**, **H2** et **H3** du théorème d'indistinguabilité sont très restrictives, et nous proposons au chapitre suivant un algorithme d'apprentissage réaliste qui, en permettant de s'éloigner des conditions idéales du théorème, donne du sens à la comparaison des prédictions des théories motrice et auditive.

Chapitre 4

Exploitation du modèle *COSMO*, dans un cadre théorique simplifié, pour comparer les approches motrice et auditive de la perception

1 Prise de position **60**

2 Instantiation du modèle *COSMO* pour comparer quantitativement les approches auditives et motrices de la perception **61**

3 Algorithmes d'apprentissage **64**

4 Comparaisons au sein de *COSMO* des prédictions des théories motrice, auditive et perceptuo-motrice de la perception **76**

5 Conclusion **83**

Dans la section 3.3 du chapitre 3, nous avons montré comment notre modèle intégrateur *COSMO* permet de réaliser des tâches de perception en utilisant uniquement des connaissances auditives, uniquement des connaissances motrices, ou en réalisant une fusion perceptuo-motrice. Nous avons montré que, dans certains cas, les théories purement motrices et purement auditives de la perception sont indistinguables, mais ce théorème d'indistinguabilité repose sur des hypothèses fortes d'apprentissage parfait, qui ne sont vraisemblablement pas vérifiées en pratique. Dans la présente section, nous proposons des algorithmes d'apprentissage plus réalistes qui permettent de sortir des conditions idéales garantissant l'indistinguabilité. Plus précisément, les algorithmes considérés permettent d'apprendre à la fois le système perceptif, le lien sensori-moteur, et le système moteur à partir d'entrées perceptives uniquement, les mêmes dans les trois cas.

Notre objectif est donc double. D'une part, il s'agit de proposer des principes d'apprentissage montrant que le système peut effectivement acquérir toutes les composantes de la distribution de probabilité conjointe qui le définit (c'est-à-dire que l'on propose des principes de construction du système perceptif, des répertoires moteurs, et du modèle interne de la transformation articulatori-acoustique). D'autre part, il s'agit de montrer comment ceci fournit des chemins

d'apprentissages imparfaits qui permettent de sortir du cadre d'indistinguabilité, ce qui rend possible des comparaisons expérimentales entre les théories motrice, auditive, et perceptuo-motrice.

Dans ce contexte, nous nous appuierons sur un résultat expérimental qui se dégage de plus en plus clairement des données récentes de neurosciences cognitives (voir la section 2.2 du chapitre 2) : le rôle des connaissances motrices semble apparaître dans les données comportementales lorsque la communication s'effectue en conditions dégradées, particulièrement en présence de bruit acoustique. C'est donc autour du rôle du bruit que nous attendons de voir apparaître des capacités de distinguabilité des différents modèles.

1 Prise de position

La simulation informatique de ces apprentissages dans un cadre théorique très simplifié nous permet de défendre dans ce chapitre les propositions suivantes.

La robustesse au bruit du modèle moteur. Les observations expérimentales dans les résultats de nos simulations feront effectivement apparaître une plus grande robustesse au bruit du modèle moteur par rapport au modèle auditif, qui s'explique par le fait que le modèle moteur, plus difficile à apprendre, est moins bien connu, et c'est cette imprécision qui se traduit par de la robustesse au bruit. En termes probabilistes, le modèle auditif, qui est un modèle facile à apprendre, aura des distributions très piquées sur les zones de l'espace correspondant exactement à son apprentissage, mais ne sera pas compétent ailleurs ; alors que le modèle moteur, plus difficile à apprendre, fera intervenir des distributions de probabilité plus plates, qui contiennent des connaissances imprécises, mais lui permettent d'être performant en généralisant mieux sur de plus grandes portions de l'espace.

La voie motrice : précise en production, imprécise en perception. Le modèle moteur de la perception fait intervenir deux termes : des prototypes moteurs de gestes fréquemment associés aux objets, et un modèle interne qui encode des connaissances sur la transformation articulatori-acoustique. Alors que le premier terme peut être très précis (il est facile à apprendre, et correspond à des distributions piquées, ce qui explique également la consistance observée sur des tâches de production) le second terme est bien plus difficile à apprendre, et c'est là que réside l'imprécision inhérente au modèle moteur qui lui apporte de la robustesse au bruit en perception.

Le rôle de la non-linéarité. La non-linéarité de la transformation articulatori-acoustique permet de développer des idiosyncrasies motrices qui ne seraient pas possibles sinon. La non-linéarité se caractérise par l'existence de plateaux séparés par une zone de forte instabilité et sur lesquels la variation des paramètres articulatoires ne cause que peu ou pas de variation des paramètres acoustiques. Cette non-linéarité a été théorisée par Stevens (1972, 1989) sous le nom de « théorie quantique de la parole ». Stevens propose dans cette théorie que les non-linéarités structurent l'espace perceptuo-moteur et font émerger naturellement, de part et d'autre de la frontière, des contrastes utilisés par les langues pour établir leurs systèmes de traits phonétiques. Dans le contexte de nos travaux, nous

utiliserons avant tout le fait qu'une telle transformation articulatori-acoustique est, dans un espace discrétisé, non-injective (*many-to-one*), ce qui implique qu'une même valeur des paramètres acoustiques puisse avoir plusieurs antécédents moteurs. Une telle non-linéarité rend alors possible des choix moteurs idiosyncrasiques (c'est-à-dire caractéristiques d'un individu) qui n'ont pas d'incidence perceptives.[24]

Le rôle de la fusion perceptuo-motrice. La fusion audio-motrice en perception permet de tirer parti avantageusement de la combinaison d'un système auditif précis mais peu robuste aux dégradations avec un modèle moteur auquel son imprécision apporte de la robustesse.

Ces propositions ayant été posées, la suite du chapitre décrit maintenant le cadre d'une étude théorique dans lequel nous allons instancier notre modèle *COSMO*.

2 Instantiation du modèle *COSMO* pour comparer quantitativement les approches auditives et motrices de la perception

Dans le chapitre 3, nous avons présenté les différentes étapes de construction d'une version générique de notre modèle *COSMO*, qui était suffisante pour aboutir à la démonstration de notre théorème d'indistinguabilité. Nous nous sommes donnés un ensemble de variables abstraites ainsi qu'une méthode de calcul de la distribution de probabilité conjointe sur ces variables, à partir du choix d'une décomposition en termes plus simples basée sur un ensemble d'hypothèses d'indépendance conditionnelle.

Pour pouvoir faire les premières comparaisons quantitatives entre les modèles auditif et moteur, il est nécessaire de davantage spécifier le modèle. Plus précisément, et conformément au cadre de la programmation bayésienne (Bessière *et al.*, 2013), la conception de notre modèle se poursuit avec les étapes suivantes.

- Donner le domaine de définition de chaque variable.
- Choisir les formes paramétriques à utiliser pour décrire chacun des termes intervenant dans la décomposition de la distribution de probabilité conjointe de notre modèle probabiliste *COSMO*. Le choix qui est fait à cette étape consiste simplement à sélectionner des fonctions mathématiques adaptées pour encoder les distributions de probabilité qui modélisent les états de connaissance de notre agent communicant sur les liens entre les variables du modèle.
- Spécifier et implémenter les algorithmes d'apprentissage à partir desquels les paramètres de ces formes paramétriques sont appris. Par exemple, si la forme paramétrique que l'on a choisie à l'étape précédente pour représenter une distribution $P\left(A \mid [B{=}b]\right)$ est une distribution de probabilité gaussienne $Gauss(\mu_b^A, \sigma_b^A)$, il faut ici donner un algorithme selon lequel sont appris la moyenne μ_b^A et l'écart-type σ_b^A de chaque gaussienne (il y en a une pour chaque valeur b de la variable B).

[24] Le *Trésor de la Langue Française* définie l'idiosyncrasie comme étent une prédisposition particulière d'un organisme qui fait que l'individu réagit d'une manière personnelle à l'influence des agents extérieurs.

- Préciser les jeux de données à partir desquels les apprentissages se font.

2.1 Domaines de définition des variables du modèle

Le modèle *COSMO* est un modèle générique qui peut être instancié de différentes manières. Alors que la démonstration du théorème d'indistinguabilité faite au chapitre précédent est indépendante de la définition précise des variables, dans le présent chapitre on instancie le modèle *COSMO* pour étudier un cas théorique simple, pour lequel les variables motrices et perceptives sont monodimensionnelles. Dans le chapitre suivant nous en proposerons une instanciation plus réaliste, pour pouvoir manipuler des syllabes.

Pour chacune des variables de notre modèle *COSMO* (voir par exemple figure 4.1) nous précisons maintenant son ensemble de définition.

Les objets O_S et O_L sont des variables discrètes pouvant prendre deux valeurs possibles : o^+ et o^-, qui correspondent à deux catégories (fictives) de phonèmes distincts.

La consigne motrice M est définie dans un espace moteur à une seule dimension : l'intervalle $[-14, 14]$ choisi arbitrairement. Cet intervalle est discrétisé uniformément en 281 valeurs possibles.

La représentation sensorielle S des conséquences perceptives des consignes motrices est définie dans un espace sensoriel à une dimension : l'intervalle $[-14, 14]$ choisi pour que les définitions de M et de S soient symétriques. De même, cet intervalle est discrétisé uniformément en 281 valeurs possibles.

Le succès de la communication est décrit par la variable booléenne C qui a donc deux valeurs possibles : 1 et 0. Il s'agit d'une variable de cohérence (Bessière *et al.*, 2013) dont nous avons vu au chapitre 3 qu'elle permet de réaliser de la fusion sensori-motrice afin d'implémenter les tâches de production et de perception dans le cadre des théories perceptuo-motrices de la parole.

Il est à noter que toutes ces variables sont discrètes. C, O_S et O_L sont des variables discrètes par nature : la première est une variable booléenne, et les deux suivantes dénotent des catégories distinctes de phonèmes. En revanche, les variables M et S sont discrètes par choix[25].

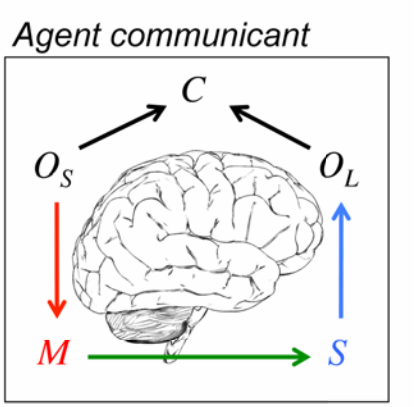
2.2 Formes paramétriques

Dans la démarche de la programmation bayésienne, le choix des formes paramétriques est une étape dans laquelle le modélisateur fait des hypothèses sur la forme des distributions de probabilité représentant les états de connaissance de l'agent, et choisit des outils mathématiques pour les encoder.

Nous rappelons figure 4.1 la structure de dépendances probabilistes du modèle *COSMO* pour mettre en évidence les différents sous-systèmes à spécifier dans le modèle d'agent π_{Ag}.

Le prior sur les objets $P(O_S \mid \pi_{Ag})$ encode la connaissance a priori qu'a l'agent sur les objets. $P(O_S \mid \pi_{Ag})$ prend la forme d'une distribution de probabilité uniforme pour traduire

[25] On peut trouver une discussion à ce sujet dans Bessière *et al.* (2013).



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

Figure 4.1: **Le modèle *COSMO*** est décrit par sa distribution de probabilité conjointe, qui se décompose en un produit de sous-systèmes : le système moteur, le système auditif et le système sensori-moteur.

le fait que les différents objets ont la même fréquence d'apparition dans l'environnement (ou l'ignorance de cette fréquence d'apparition). On a donc $P\left([O_S{=}o^+] \mid \pi_{Ag}\right) = 1/2$ et $P\left([O_S{=}o^-] \mid \pi_{Ag}\right) = 1/2$.

Le système moteur $P(M \mid O_S\ \pi_{Ag})$ décrit la connaissance qu'a l'agent sur les gestes moteurs associés aux objets. On choisit d'utiliser des distributions de probabilité gaussiennes : $P\left(M \mid [O_S{=}o]\ \pi_{Ag}\right) = Gauss(\mu_o^M, \sigma_o^M)$. Puisque O_S a deux valeurs possibles, o^- et o^+, la connaissance qu'a l'agent sur son système moteur est donc encodée par deux distributions de probabilité gaussiennes : une pour chacun des prototypes moteurs $P\left(M \mid [O_S{=}o^-]\ \pi_{Ag}\right)$ et $P\left(M \mid [O_S{=}o^+]\ \pi_{Ag}\right)$.

Le système sensori-moteur $P(S \mid M\ \pi_{Ag})$ décrit la connaissance qu'a l'agent sur la relation entre les consignes motrices (gestes articulatoires) et leurs conséquences perceptives. Ce terme encode un modèle interne de la transformation articulatori-acoustique. On choisit d'utiliser des distributions de probabilité gaussiennes : pour chacune des valeurs possibles de la variable M, les connaissances de l'agent sur les conséquences perceptives de cette consigne articulatoire m sont encodées sous forme d'une distribution de probabilité gaussienne, c'est-à-dire $P\left(S \mid [M{=}m]\ \pi_{Ag}\right) = Gauss(\mu_m^S, \sigma_m^S)$, ce qui fait 281 gaussiennes.

Le système auditif $P(O_L \mid S\ \pi_{Ag})$ décrit la connaissance qu'a l'agent sur la relation entre les entrées perceptives et les objets. Ce terme encode le classifieur auditif de l'agent. On choisit d'utiliser un classifieur gaussien. Plus précisément, le terme $P(O_L \mid S\ \pi_{Ag})$ est fixé comme étant l'inversion probabiliste des prototypes auditifs gaussiens correspondant à chaque objet, qui sont de la forme $P(S \mid O_L) = Gauss(\mu_o^S, \sigma_o^S)$.

Le système de validation de la communication $P(C \mid O_S\, O_L\, \pi_{Ag})$ décrit la connaissance qu'a l'agent sur le succès de la communication. C est une variable booléenne qui est vraie avec

une probabilité 1 lorsque les objets considérés du point de vue du locuteur et de l'auditeur sont les mêmes (O_S=O_L), et qui est fausse sinon. Mathématiquement, cela se traduit par l'utilisation d'un Dirac fontionnel, ce qui peut s'écrire $P\left([C{=}1]\,|O_S\ O_L\ \pi_{Ag}\right) = \delta_{O_S=O_L}$.

Les connaissances encodées dans notre modèle dépendent donc de la valeur de plusieurs paramètres : μ_o^M et σ_o^M (quatre paramètres) qui contrôlent les prototypes moteurs correspondant à chaque objet o, μ_m^S et σ_m^S (2×281 paramètres) qui contrôlent la représentation dans le modèle interne de l'agent des conséquences perceptives du geste moteur m, et μ_o^S et σ_o^S (quatre paramètres) qui contrôlent les prototypes acoustiques correspondant à chaque objet. Ces paramètres peuvent prendre des valeurs différentes dans chaque instance de notre modèle, et leur valeur peut évoluer au cours de phases d'apprentissage lors desquelles l'agent va mettre à jour ses connaissances à partir des observations qu'il fait sur son environnement.

3 Algorithmes d'apprentissage

Alors que le théorème d'indistinguabilité repose sur trois hypothèses d'apprentissage parfait du classifieur auditif, d'internalisation parfaite de la tranformation articulatori-acoustique et d'identité motrice parfaite entre le maître et l'agent, dans cette section nous proposons des algorithmes d'apprentissage qui s'écartent de ces trois hypothèses, ce qui permettra d'observer des différences entre les réponses des systèmes moteur et perceptif en perception.

3.1 Interactions avec un agent maître

L'apprentissage des paramètres de notre modèle d'agent communicant se fait au travers d'interactions avec un agent maître. L'agent π_{Ag} et l'agent maître $\pi_{Maître}$ sont deux instances de notre modèle *COSMO*. Cela veut dire en particulier que les modèles π_{Ag} et $\pi_{Maître}$ sont définis par des distributions de probabilité conjointes décomposées de la même manière, sur les mêmes variables, et que leurs distributions de probabilité conditionnelles sont décrites par les mêmes formes paramétriques. En revanche, les valeurs des paramètres de ces formes paramétriques peuvent différer.

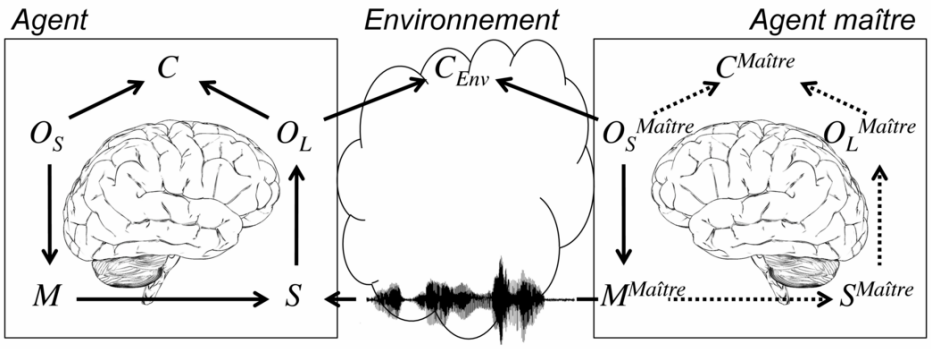


Figure 4.2: **Scénario d'interactions avec un agent maître** : L'agent π_{Ag} interagit avec l'agent maître $\pi_{Maître}$ au sein de l'environnement π_{Env}. L'agent reçoit des stimuli étiquetés $< s,\ o >$ lui permettant d'apprendre les liens entre les objets o communiqués par le maître et les entrées perceptives s correspondantes.

Le paradigme d'interaction avec l'agent maître qui est la base de notre scénario d'apprentissage supervisé est le même que celui qui est présenté à la section 4.1 du chapitre précédent, que l'on rappelle figure 4.2 : l'agent π_{Ag} reçoit des stimuli étiquetés $< s,\ o >$ lui permettant d'apprendre les liens entre les objets o communiqués par le maître $\pi_{Maître}$ et les entrées perceptives s correspondantes.

3.1.1 Les propriétés du modèle de l'environnement

Rappelons ici les deux objectifs de ce chapitre : comparer les prédictions des approches purement motrices et purement auditives de la perception dans le bruit d'une part, et analyser l'impact de la non-linéarité de la transformation articulatori-acoustique sur ces prédictions. Alors que pour s'attaquer au premier objectif il suffit d'ajouter au moment du test un bruit de communication qui n'était pas présent lors de l'apprentissage, le second nécessite de faire des hypothèses sur la nature de la transformation articulatori-acoustique réalisée par l'environnement.

Ainsi, dans cette section et la suivante, nous spécifions les modèles de l'environnement π_{Env} et du maître $\pi_{Maître}$ afin d'étudier un cas d'école inspiré de l'idée selon laquelle les non-linéarités de la transformation articulatori-acoustique décrites par Stevens fournissent des frontières catégorielles naturelles. On choisit donc de décrire le lien physique entre nos variables monodimensionnelles M et S (qui représentent respectivement les gestes articulatoires et leurs conséquences perceptives) par une fonction sigmoïde.

$$S = sigmoide(M, a, b) = b \times \frac{Arctan(a \times M)}{Arctan(a \times b)} \ . \tag{4.1}$$

Le comportement de cette fonction est défini grâce à deux paramètres a et b, qui permettent respectivement de contrôler la pente et l'amplitude de la sigmoïde. De plus, il s'agit d'une fonction impaire, qui est donc symétrique par rapport à l'origine.

La figure 4.3 met en parallèle la vision de Stevens (1989) des non-linéarités et notre fonction sigmoïde avec plusieurs valeurs de pente utilisées pour modéliser différents degrés de non-linéarité dans la relation articulatori-acoustique.

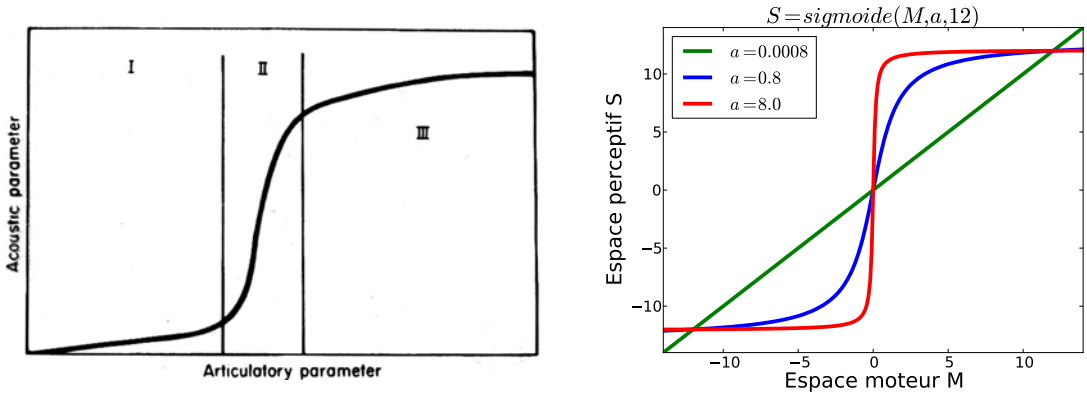


Figure 4.3: **Modélisation des non-linéarités** : À gauche, les non-linéarités vues par Stevens (1989) (deux plateaux sont séparés par une zone de forte instabilité) ; à droite trois fonctions sigmoïdes pour lesquelles $b = 12$, avec trois valeurs différentes de a pour montrer comment ce paramètre permet de contrôler le degré de linéarité.

Comme le montre la figure 4.3, une faible valeur de a correspond à une transformation $M \mapsto S$ quasiment linéaire, alors que des valeurs plus élevées de a correspondent à une transformation plus ou moins fortement non-linéaire. Cette fonction sigmoïde modélise de manière déterministe la manière dont on obtient le signal physique à partir des consignes motrices du maître. À cela vient s'ajouter un modèle de perturbation gaussien qui encode un bruit d'environnement. La distribution sur les entrées perceptives reçues par l'agent lorsque le maître choisit une consigne motrice m s'écrit alors

$$P\left(S \mid \left[M^{Maître}{=}m\right] \ \pi_{Env}\right) = Gauss(sigmoide(m, a, b),\ \sigma_{Env}) \ ,$$

c'est-à-dire une distribution de probabilité gaussienne dans laquelle la moyenne est donnée par la fonction sigmoïde, et où l'écart-type σ_{Env} quantifie la manière dont le modèle de perturbation (essentiellement le bruit de l'environnement) vient dégrader le S physique qui aurait été perçu sinon. La valeur de σ_{Env} est fixée à 1 pour l'apprentissage, et variera entre 1 et 10 au cours de l'évaluation pour tester la robustesse au bruit de l'environnement.

3.1.2 Les propriétés du modèle du maître

Pour les besoins de l'apprentissage, il n'y a que deux termes de la décomposition de la distribution de probabilité conjointe du modèle du maître qui nous intéressent : la manière de choisir les objets à communiquer selon le prior $P\left(O_S^{Maître} \mid \pi_{Maître}\right)$, et la manière de choisir des consignes motrices pour chaque objet o selon la distribution de probabilité conditionnelle $P\left(M^{Maître} \mid \left[O_S^{Maître}{=}o\right] \ \pi_{Maître}\right)$.

Le premier terme étant encodé par une distribution de probabilité uniforme, il n'y a pas de paramètre à fixer : on a $P\left(\left[O_S^{Maître}{=}o^-\right] \mid \pi_{Maître}\right) = P\left(\left[O_S^{Maître}{=}o^+\right] \mid \pi_{Maître}\right) = 1/2$. En revanche, il faut fixer la valeur des paramètres des gaussiennes encodant les répertoires moteurs $P\left(M^{Maître} \mid O_S^{Maître} \ \pi_{Maître}\right)$.

Puisque la non-linéarité de la transformation articulatori-acoustique (schématisée figure 4.3) vient structurer l'espace perceptif (celui de la variable S), on tire parti de cette structure en positionnant alors les prototypes moteurs du maître $P(M^{Maître} \mid O_S^{Maître} \ \pi_{Maître})$ de part et d'autre de la zone de forte instabilité, attribuant ainsi à chaque objet (o^+ et o^-) une zone de plateau. Puisque les formes paramétriques qui ont été choisies pour représenter les prototypes moteurs sont $P\left(M^{Maître} \mid \left[O_S^{Maître}{=}o\right] \ \pi_{Maître}\right) = Gauss(\mu_o, \sigma_o)$, pour finir de spécifier le système moteur de l'agent maître, il suffit de préciser que l'on fixe $\mu_o^- = -5$, $\mu_o^+ = 5$, et $\sigma_o^- = \sigma_o^+ = 1$.

Pour fixer les idées, la figure 4.4 montre les prototypes moteurs du maître, trois différentes transformations articulatori-acoustique (avec $\sigma_{Env} = 1$) correspondant à différents niveaux de non-linéarité, et les distributions $P(S \mid O_S^{Maître})$ sur les entrées perceptives qui arrivent à l'agent.

Lorsque la transformation articulatori-acoustique est parfaitememt linéaire, et en l'absence de bruit de communication, les prototypes moteurs $P\left(M^{Maître} \mid O_S \ \pi_{Maître}\right)$ du maître et leur image $P\left(S \mid O_S^{Maître}\right)$ dans l'espace acoustique sont parfaitement similaires : ces distributions de probabilités ont la même moyenne et le même écart-type. En revanche, augmenter le degré de non-linéarité de la transformation articulatori-acoustique a trois conséquences : écarter la moyenne des distributions de probabilité $P\left(S \mid O_S^{Maître}\right)$, réduire leur écart-type, et induire une légère asymétrie des distributions de probabilité sur S. Finalement, la non-linéarité a pour effet de séparer davantage dans l'espace acoustique ce qui l'est moins dans l'espace moteur.

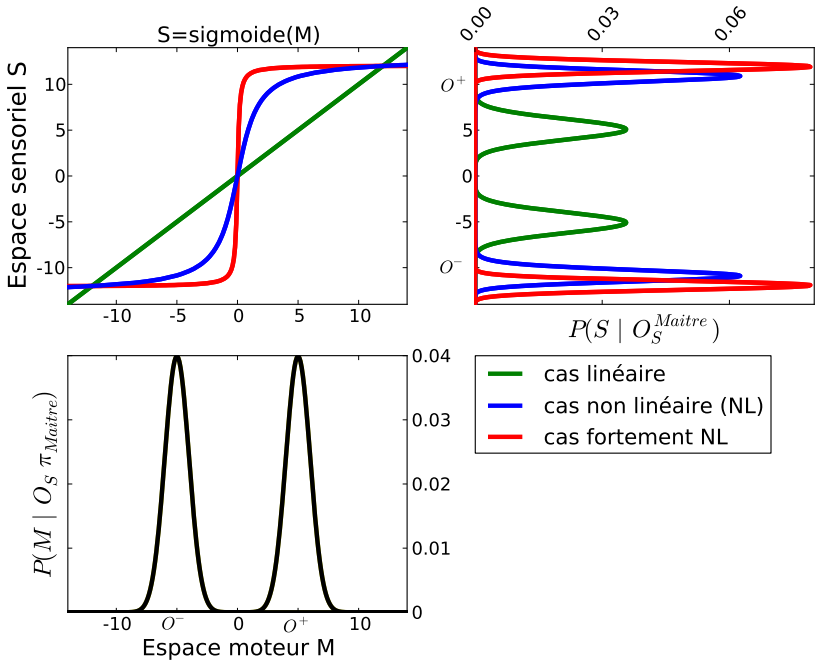


Figure 4.4: **Les distributions de probabilité sur les stimuli reçus par l'agent** $P\left(S \mid O_S^{Maître}\right)$ (en haut à droite) dépendent des prototypes moteurs du maître $P(M^{Maître} \mid O_S^{Maître}\ \pi_{Maître})$ (en bas à gauche), et de la transformation articulatori-acoustique $P\left(S \mid M^{Maître}\ \pi_{Env}\right)$ réalisée par l'environnement (en haut à gauche).

3.1.3 Données d'apprentissage

Dans ce qui va suivre, l'agent π_{Ag} apprend son classifieur auditif puis son modèle interne et ses répertoires moteurs à partir des mêmes données, qui sont obtenues en combinant le modèle du maître $\pi_{Maître}$ et le modèle de l'environnement π_{Env}. Les interactions entre le maître et l'agent apprenant se déroulent de la manière décrite à la section 4.1 du chapitre 3 : l'agent reçoit du maître des stimuli s ainsi que les catégories d'objets o correspondantes. Ces couples de valeurs $< s,\ o >$ sont obtenus par des tirages successifs d'un objet o selon le prior $P\left(O_S^{Maître} \mid \pi_{Maître}\right)$, d'une consigne motrice m selon le prototype moteur $P(M^{Maître} \mid [O_S^{Maître}{=}o]\ \pi_{Maître})$, et d'un stimulus s selon le modèle de l'environnement $P\left(S \mid [M^{Maître}{=}m]\ \pi_{Env}\right)$.

Par souci d'efficacité, dans l'implémentation des algorithmes d'apprentissage qui vont suivre, il est possible de faire l'économie des tirages intermédiaires en tirant directement des stimuli s selon la distribution de probabilité $P\left(S \mid O_S^{Maître}\right)$ que l'on peut précalculer une fois pour toutes de la manière suivante :

$$P\left(S \mid O_S^{Maître}\right) \propto \sum_{M^{Maître}} P\left(M^{Maître} \mid O_S^{Maître}\ \pi_{Maître}\right) P\left(S \mid M^{Maître}\ \pi_{Env}\right) \tag{4.2}$$

Ainsi, la branche auditive (le classifieur auditif) et la branche motrice (le modèle interne de la transformation articulatori-acoustique et les répertoires moteurs) sont apprises à partir des mêmes données, au sens où les entrées $< s, o >$ de l'apprentissage sont obtenues de la même

manière : en tirant des stimuli s selon la même distribution de probabilité $P(S \mid [O_S^{Maître}{=}o])$ qui a été calculée au préalable suivant l'équation 4.2.

3.2 Apprentissage supervisé du classifieur auditif

Le système auditif $P(O_L \mid S\ \pi_{Ag})$ de l'agent est défini comme étant un classifieur gaussien. Au cours de l'apprentissage, l'agent se construit des prototypes auditifs gaussiens de la forme $P(S \mid O_L) = Gauss(\mu_o^S, \sigma_o^S)$, qui encodent la connaissance que l'agent se construit progressivement sur les zones de l'espace perceptif qui sont associées aux objets, et dont l'inversion probabiliste donne le terme $P(O_L \mid S\ \pi_{Ag})$.

Plus précisément, pour chaque couple de valeurs $< s,\ o >$ transmis par le maître, l'agent met à jour ses connaissances avec l'observation selon laquelle l'objet o peut correspondre à un signal perceptif s. Concrètement, pour chacun de ces couples $< s,\ o >$, l'agent utilise la valeur de s pour mettre à jour les valeurs μ_o^S et σ_o^S de la moyenne et de l'écart-type de la gaussienne sur l'espace perceptif correspondant à l'objet o.

Les valeurs initiales des paramètres μ_o^S et σ_o^S sont choisies pour que ces gaussiennes se comportent comme des distributions de probabilité uniformes : on met la moyenne μ_o^S au centre de l'espace perceptif, et l'écart-type initial σ_o^S reçoit une très grande valeur comparativement à la taille de l'espace perceptif des S. Ensuite, au cours de l'apprentissage, l'agent affine progressivement sa représentation du lien entre objets et percepts, et ses prototypes gaussiens de la forme $P(S \mid O_L)$ convergent petit à petit.

La figure 4.5 montre les prototypes auditifs que l'agent se construit lors de l'apprentissage pour différents degrés de non-linéarité de la transformation articulatori-acoustique.

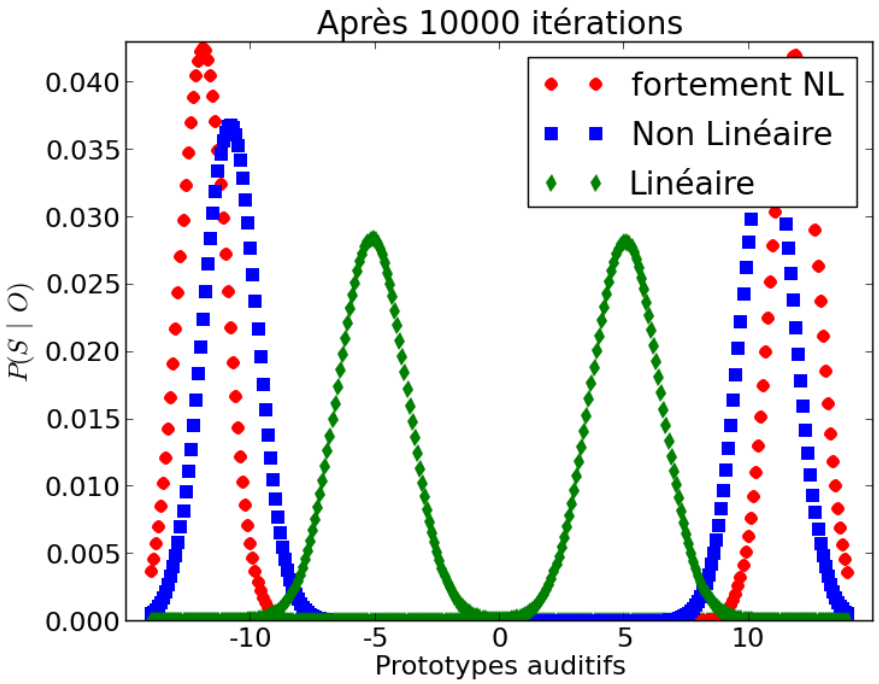


Figure 4.5: **Les prototypes auditifs** dont dispose l'agent après convergence de l'apprentissage, pour trois degrés de non-linéarité de la transformation articulatori-acoustique.

Cette figure 4.5 montre que plus la transformation articulatori-acoustique est fortement non-linéaire, plus les prototypes auditifs gaussiens correspondant aux deux objets ont des moyennes écartées, et plus leur variance est petite. Si l'on compare la figure 4.5 avec la figure 4.4, il en ressort que les prototypes $P(S \mid O_L)$ appris par l'agent sont quasiment identiques aux distributions $P\left(S \mid O_S^{Maître}\right)$ des stimuli qui lui parviennent, à ceci près que les distributions $P(S \mid O_L)$, étant gaussiennes, sont symétriques, alors que ce n'est pas le cas des distributions $P\left(S \mid O_S^{Maître}\right)$. En tous cas, la comparaison des figures 4.5 et 4.4 montre que les prototypes auditifs que l'agent se construit au cours de l'apprentissage capturent convenablement l'information qui lui parvient du maître.

On définit alors le système auditif de l'agent comme étant l'inversion bayésienne de ces prototypes gaussiens. Ainsi, le système perceptif de l'agent $P(O_L \mid S\ \pi_{Ag})$ consiste simplement en un classifieur gaussien dont les paramètres sont appris de manière supervisée.

3.3 Apprentissage de compétences motrices par accomodation

Dans cette section nous présentons un algorithme qui permet à l'agent, à partir des mêmes paires $< o,\ s >$ données par le maître que dans la partie précédente, d'apprendre par accomodation son modèle interne de la transformation articulatori-acoustique $P(S \mid M\ \pi_{Ag})$, et ses répertoires moteurs $P(M \mid O_S\ \pi_{Ag})$ de consignes articulatoires à associer aux objets. Cet algorithme permet donc à l'agent d'apprendre les paramètres de son système moteur et de son système sensori-moteur à partir d'entrées perceptives uniquement.

3.3.1 Paradigme d'imitation

L'approche qui est souvent adoptée dans des travaux similaires (comme ceux de Guenther (1995) avec le modèle *DIVA*) pour l'apprentissage de compétences motrices consiste à réaliser successivement deux étapes distinctes. Une première étape de babillage aléatoire, souvent exhaustif, permet d'apprendre le lien sensori-moteur en échantillonant grossièrement sur l'ensemble de l'espace moteur. Une fois que l'on dispose de ce premier maillage sommaire, une seconde étape d'apprentissage par imitation permet d'affiner progressivement le choix des consignes motrices visant à reproduire des cibles acoustiques. En effet, un mécanisme de feedback auditif permet de comparer les cibles à atteindre avec ce qui a été effectivement produit, et lorsque la cible n'est pas bien réalisée l'écart avec l'objectif permet de déterminer la direction dans laquelle faire évoluer la consigne motrice afin de réduire cette erreur[26].

Au lieu d'apprendre le lien sensori-moteur en commençant par explorer aléatoirement l'espace des actuateurs, une autre approche issue de la robotique, connue sous le nom de « goal babbling », (Rolf *et al.*, 2010) propose au contraire de se donner directement des objectifs dans l'espace des buts (c'est-à-dire dans notre cas l'espace acoustique). Le choix de ces objectifs peut également être aléatoire, ou être guidé par la notion de motivation intrinsèque : une certaine forme de curiosité guide l'apprentissage de l'agent qui organise son exploration en cherchant à maximiser les progrès réalisés vis-à-vis de la tâche à accomplir (voir par exemple Baranes et Oudeyer (2013), ou encore Moulin-Frier et Oudeyer (2013a) pour une application dans le domaine de la parole).

[26] Une implémentation typique de ce mécanisme est l'algorithme de rétropropagation du gradient classiquement utilisé pour l'apprentissage des paramètres des réseaux de neurones.

Ce que nous proposons de décrire et d'implémenter dans la suite de ce chapitre, c'est un mécanisme original d'apprentissage de compétences motrices par accomodation, qui se distingue de ce qui vient d'être présenté par plusieurs aspects :

- il ne nécessite pas de faire intervenir de boucle de feedback, de se donner une mesure d'erreur, ou de définir une notion de progrès ;
- il se fait en une unique étape et ne présuppose pas de première phase de babillage aléatoire, sans buts ;
- à l'inverse, dans le paradigme d'apprentissage que nous proposons ce sont les cibles qui guident l'exploration de l'espace sensori-moteur ;
- et, contrairement au cas du « goal babbling » où le choix des cibles est simplement limité par la taille de l'espace des buts, notre paradigme d'apprentissage par accomodation intègre une dimension sociale dans la mesure où les cibles à imiter sont des cibles écologiques fournies par un agent maître guidant l'apprentissage. (Autrement dit, au lieu d'avoir des cibles couvrant tout l'espace possible, l'agent n'apprend que les portions *utiles* de l'espace : celles utilisées par le maître.)

Le point clé de cet algorithme est la manière d'implémenter la tâche d'imitation, c'est-à-dire la manière dont l'agent choisit ses gestes moteurs pour essayer de reproduire les cibles acoustiques fournies par le maître.

3.3.2 Inférence probabiliste pour la tâche d'imitation

Nous formalisons la tâche d'imitation de la manière suivante : étant donnée une paire $< s,\ o >$ fournie par le maître, associant un stimulus s à un objet o, l'agent apprenant π_{Ag} doit choisir une consigne motrice m dans le but d'atteindre la cible s. Cela revient à effectuer un tirage aléatoire selon la distribution de probabilité $P\,(M \mid [S{=}s]\ [O_S{=}o]\ \pi_{Ag})$. Nous montrons maintenant par inférence probabiliste comment se calcule cette distribution de probabilité qui encode la tâche d'imitation. D'après le théorème de Bayes, on a :

$$P\,(M \mid [S{=}s]\ [O_S{=}o]\ \pi_{Ag}) = \frac{P\,(M\ [S{=}s]\ [O_S{=}o] \mid \pi_{Ag})}{P\,([S{=}s]\ [O_S{=}o] \mid \pi_{Ag})}\ . \tag{4.3}$$

Il s'agit du calcul d'une distribution de probabilité sur la variable M, dans lequel le dénominateur $P\,([S{=}s]\ [O_S{=}o] \mid \pi_{Ag})$ est une constante. Ce terme permet d'assurer la normalisation de la distribution de probabilité conditionnelle $P\,(M\ [S{=}s]\ [O_S{=}o] \mid \pi_{Ag})$, et on peut se contenter de le calculer a posteriori. On peut donc écrire, en utilisant l'opérateur de proportionalité $\propto$:

$$P\,(M \mid [S{=}s]\ [O_S{=}o]\ \pi_{Ag}) \propto P\,(M\ [S{=}s]\ [O_S{=}o] \mid \pi_{Ag})\ . \tag{4.4}$$

En marginalisant par rapport aux variables libres et en réordonnant les variables, on obtient :

$$P\,(M \mid [S{=}s]\ [O_S{=}o]\ \pi_{Ag}) \propto \sum_{O_L\ C} P\,([O_S{=}o]\ M\ [S{=}s]\ O_L\ C \mid \pi_{Ag})\ . \tag{4.5}$$

En utilisant maintenant la décomposition de la distribution de probabilité conjointe de notre modèle *COSMO*, cela donne :

$$P\left(M \mid [S{=}s]\ [O_S{=}o]\ \pi_{Ag}\right) \propto \sum_{O_L\ C} \Big(\begin{array}{l} P\left([O_S{=}o] \mid \pi_{Ag}\right) \times P(M \mid [O_S{=}o]\ \pi_{Ag}) \times \\ P\left([S{=}s] \mid M\ \pi_{Ag}\right) \times \\ P\left(O_L \mid [S{=}s]\ \pi_{Ag}\right) \times P\left(C \mid O_L\ [O_S{=}o]\ \pi_{Ag}\right) \Big) . \end{array} \tag{4.6}$$

Le terme $P\left([O_S{=}o] \mid \pi_{Ag}\right)$ est constant puisque la distribution de probabilité $P(O_S \mid \pi_{Ag})$ est uniforme. Il est donc absorbé par le symbole de proportionalité. En factorisant ce qui peut l'être, il appert alors :

$$\begin{array}{rl} P\left(M \mid [S{=}s]\ [O_S{=}o]\ \pi_{Ag}\right) & \propto\ P\left(M \mid [O_S{=}o]\ \pi_{Ag}\right) \times P\left([S{=}s] \mid M\ \pi_{Ag}\right) \times \\ & \sum_{O_L} \Bigg(P\left(O_L \mid [S{=}s]\ \pi_{Ag}\right) \times \sum_{C} P\left(C \mid O_L\ [O_S{=}o]\ \pi_{Ag}\right) \Bigg). \end{array} \tag{4.7}$$

D'après la règle de normalisation, la somme sur C, puis la somme sur O_L sont toutes deux égales à 1. Le calcul se simplifie, et il reste donc :

$$P\left(M \mid [S{=}s]\ [O_S{=}o]\ \pi_{Ag}\right) \propto P\left(M \mid [O_S{=}o]\ \pi_{Ag}\right) \times P\left([S{=}s] \mid M\ \pi_{Ag}\right) . \tag{4.8}$$

Le calcul de la tâche d'imitation combine donc deux termes : le premier, $P(M \mid [O_S{=}o]\ \pi_{Ag})$, fait intervenir le système moteur de l'agent, et le second, $P\left([S{=}s] \mid M\ \pi_{Ag}\right)$, fait intervenir le système sensori-moteur de l'agent. C'est-à-dire que, dans le choix des gestes moteurs susceptibles d'atteindre une cible s associée à l'objet o, interviennent à la fois l'état de connaissance de la transformation articulatori-acoustique stocké dans le modèle interne de l'agent, et les connaissances encodées dans les répertoires moteurs de l'agent sur les gestes qu'il est habitué à associer à l'objet o. Ce sont ces états de connaissance qui vont évoluer au cours de l'apprentissage, au fur et à mesure que l'agent π_{Ag} accumule de nouvelles observations des conséquences de ses actions sur sa perception.

3.3.3 L'algorithme d'apprentissage de compétences motrices par accomodation : du babillage orienté vers des cibles

Initialement, les moyennes μ_m^S des gaussiennes $P\left(S \mid [M{=}m]\ \pi_{Ag}\right)$ sont fixées à 0, et leurs écarts-types σ_m^S à une valeur très grande par rapport à la taille de l'espace des S. De cette manière, les gaussiennes sont dégénérées et se comportent comme des distributions de probabilité uniformes, ce qui permet d'encoder le fait qu'initialement l'agent ne dispose d'aucune connaissance sur le lien qui existe entre les consignes motrices et leurs conséquences perceptives. De même, les moyennes μ_o^M des gaussiennes $P\left(M \mid [O_S{=}o]\ \pi_{Ag}\right)$ sont fixées à 0, et leurs écarts-types σ_o^M à une valeur très grande. Ensuite, l'algorithme itère les étapes suivantes.

La production d'une consigne motrice par le maître se fait en deux étapes. Le maître tire tout d'abord un objet o à communiquer selon la distribution de probabilité uniforme

$P\left(O_S^{Maître} \mid \pi_{Maître}\right)$. Il tire ensuite un geste moteur m selon la distribution de probabilité gaussienne $P\left(M^{Maître} \mid \left[O_S^{Maître}{=}o\right] \ \pi_{Maître}\right)$ de son système de production qui correspond à l'objet o.

La transmission par l'environnement : l'agent apprenant reçoit alors un stimulus perceptif s qui correspond à l'image de ce m par la transformation articulatori-acoustique après perturbation lors de son passage par l'environnement. Ce s est donc obtenu grâce à un tirage selon la distribution de probabilité gaussienne $P\left(S \mid \left[M^{Maître}{=}m\right] \ \pi_{Env}\right) = Gauss(sigmoide(m, a, b), \ \sigma_{Env})$.

La communication de l'objet : en plus de la valeur de s, le maître fournit également à l'agent la valeur o de l'objet, grâce au mécanisme d'attention partagée implémenté par le biais de la variable de cohérence C_{Env}.

Le choix d'un geste moteur pour imiter le maître : l'agent essaye d'imiter le maître ; c'est-à-dire qu'il interprète le stimulus perceptif s qui lui arrive comme une cible à reproduire. L'agent va donc calculer un geste moteur m' susceptible de lui permettre d'atteindre la cible s sachant qu'elle correspond à l'objet o également fourni par le maître. Cette tâche d'imitation combine deux types de connaissances : des connaissances sur la transformation réalisée par l'environnement qui sont stockées dans le modèle interne de l'agent $P\left([S{=}s] \mid M \ \pi_{Ag}\right)$; et des connaissances sur les gestes moteurs fréquemment associés aux objets qui sont stockées dans les répertoires moteurs de l'agent $P\left(M \mid [O_S{=}o] \ \pi_{Ag}\right)$.

La réalisation de la consigne motrice : l'agent envoie cette consigne motrice m' à son système moteur, ce qui a pour effet de produire une conséquence perceptive s'. Cette valeur de s' est obtenue en utilisant directement un modèle physique externe, qui est bien distinct du modèle interne de l'agent, qui lui est cognitif. Alors que le premier est déterminé par la physique de la transformation articulatori-acoustique, le second est un modèle cognitif qui encode un état de connaissance à un moment donné, que l'agent se construit progressivement à partir de l'observation des conséquences de ses actions sur sa perception.

La mise à jour des représentations internes : l'agent met à jour son modèle interne de la transformation articulatori-acoustique en prenant en compte l'observation que la consigne m' a la conséquence s' pour mettre à jour la moyenne $\mu_{m'}^S$ et l'écart-type $\sigma_{m'}^S$ de la gaussienne $P\left(S \mid [M{=}m'] \ \pi_{Ag}\right)$. De plus, l'agent fait également évoluer ses prototypes moteurs en utilisant l'association du geste moteur m' à l'objet o que le maître lui a transmis pour mettre à jour les paramètres μ_o^M et σ_o^M de la gaussienne $P\left(M \mid [O_S{=}o] \ \pi_{Ag}\right)$.

3.3.4 Dynamique de l'apprentissage

Il est tout d'abord intéressant de remarquer que, bien qu'il s'agisse d'apprentissage supervisé dans la mesure où le maître fournit des stimuli s étiquetés par leur catégorie d'objet o, la tâche d'imitation se fait sans retour de la part du maître, et sans boucle de feedback perceptif. Ainsi, au début de l'apprentissage, les gestes moteurs m' choisis par l'agent ont peu de chance de correspondre à la cible s à atteindre. En effet, initialement l'agent ne dispose d'aucune connaissance

dans aucun des deux termes $P\,(M \mid [O_S{=}o]\ \pi_{Ag})$ et $P\,([S{=}s] \mid M\ \pi_{Ag})$ qui interviennent dans la tâche d'imitation :

$$P(M \mid [S{=}s]\ [O_S{=}o]\ \pi_{Ag}) \propto P(M \mid [O_S{=}o]\ \pi_{Ag}) \times P\,([S{=}s] \mid M\ \pi_{Ag})\ .$$

Au début de l'apprentissage, l'agent choisit donc les gestes moteurs à réaliser au hasard, selon des distributions de probabilité qui sont presque uniformes. Par la suite, en tirant parti des nouvelles observations effectuées à chaque étape, notre algorithme d'apprentissage fait que l'exploration de l'espace sensori-moteur est conduite par l'évolution de ces deux termes $P\,(M \mid [O_S{=}o]\ \pi_{Ag})$ et $P\,([S{=}s] \mid M\ \pi_{Ag})$, que l'on peut résumer par les deux principes suivants.

L'affinage progressif du modèle interne de la transformation articulatori-acoustique
Si à une étape de l'algorithme la conséquence perceptive s' de la réalisation du geste m' est éloignée de la cible s à imiter, l'agent dispose tout de même d'une nouvelle information – le fait que le geste m' peut avoir la conséquence s' – et met à jour ses connaissances en modifiant en conséquence les valeurs de la moyenne $\mu^S_{m'}$ et de l'écart-type $\sigma^S_{m'}$ de la distribution de probabilité gaussienne $P\,(S \mid [M{=}m']\ \pi_{Ag})$, ce qui a pour effet de produire une distribution un peu plus piquée autour de la valeur de s'. Grâce à cette observation, lors de la prochaine itération de l'algorithme, pour reproduire la cible s l'agent choisira le geste m' avec une probabilité inférieure à celle de l'itération courante ; et pour reproduire la cible s' l'agent choisira le geste m' avec une probabilité supérieure.

L'ancrage dans les répertoires moteurs de choix idiosyncrasiques
À chaque étape de l'algorithme d'apprentissage, l'agent utilise également la valeur m' du geste moteur choisi pour atteindre la cible s associée à l'objet o fourni par le maître pour mettre à jour ses répertoires moteurs. L'agent modifie donc les valeurs de la moyenne μ^M_o et de l'écart-type σ^M_o de la distribution de probabilité gaussienne $P\,(M \mid [O_S{=}o]\ \pi_{Ag})$, ce qui a pour effet de produire une distribution un peu plus piquée autour de la valeur de m'. Grâce à cette observation, lors de la prochaine itération de l'algorithme, pour reproduire une cible associée à l'objet o, l'agent choisira le geste m' avec une probabilité supérieure à celle de l'itération courante. Il s'agit là d'un mécanisme d'ancrage grâce auquel, pour atteindre une cible l'agent aura tendance à priviliégier, parmi les gestes moteurs susceptibles de permettre d'atteindre la cible, ceux qu'il a déjà utilisés par le passé. C'est ce mécanisme qui permet également de développer des idiosyncrasies en ancrant des choix moteurs au cours de l'apprentissage.

La combinaison de ces deux principes fait que, au cours de l'apprentissage, le choix des gestes moteurs pour la tâche d'imitation se concentre sur des zones de plus en plus réduites de l'espace moteur qui permettent de bien atteindre les cibles. Ainsi, au fur et à mesure que l'agent accumule de nouvelles observations, notre algorithme passe progressivement d'un stade initial dépourvu de connaissances pouvant ressembler à de l'exploration uniforme, à un second stade d'apprentissage de plus en plus fin des zones de l'espace sensori-moteur correspondant aux cibles à reproduire, jusqu'à convergence de l'apprentissage.

3.3.5 Propriétés des modèles appris

Dans cette section nous proposons d'illustrer ce qui a été appris par l'agent au travers de ses interactions avec le maître et de ses tentatives d'imitation. Nous présentons les modèles appris au cours d'une réalisation de notre algorithme d'apprentissage. Puisque cet algorithme fait intervenir des tirages aléatoires, des initialisations différentes du générateur aléatoire conduiront à des choix différents de stimuli pour le maître, et de gestes moteurs différents pour l'agent qui l'imite. Cependant, les données que nous allons présenter permettent d'illustrer qualitativement des principes qui ont été reproduits dans toutes les simulations qui ont été faites dans le cadre des travaux de cette thèse.

Le modèle interne appris

La figure 4.6 montre le modèle interne de la transformation articulatori-acoustique que l'agent s'est constitué au cours d'une réalisation de l'algorithme d'apprentissage, dans le cas non-linéaire (la pente de la sigmoïde est à $0,8$), et après 100 000 itérations de l'algorithme.

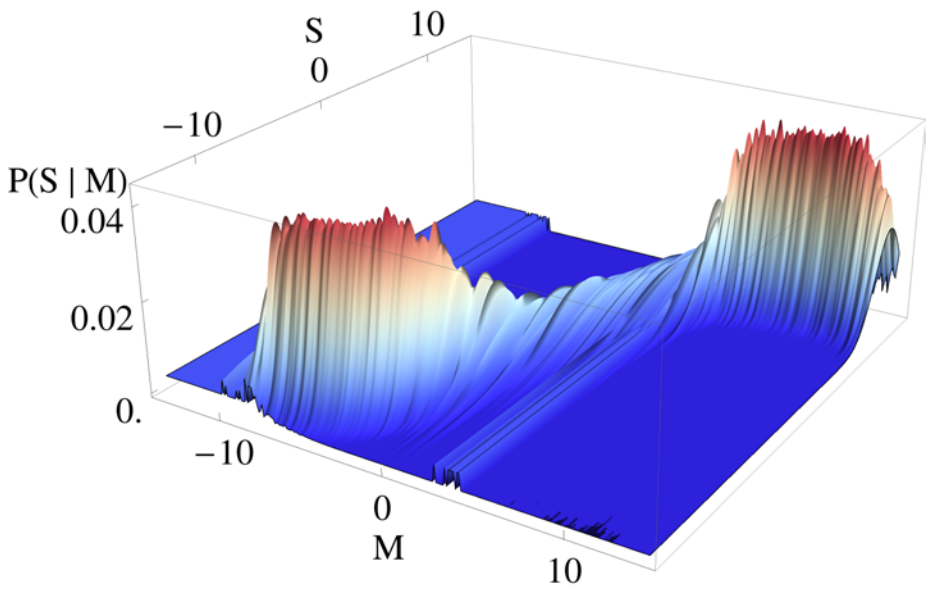


Figure 4.6: **Le modèle interne appris par l'agent** pour la transformation articulatori-acoustique après convergence de l'apprentissage par accomodation du système sensori-moteur. Cette figure montre, pour chaque valeur m de la variable M, la distribution de probabilité $P\left(S \mid [M{=}m]\ \pi_{Ag}\right)$ qui décrit la connaissance que l'agent a accumulée dans son modèle interne sur les conséquences perceptives S de la consigne motrice m.

Comme le montre la figure 4.6, l'agent capture dans son modèle interne une représentation approximative de la fonction sigmoïde reliant M et S (qui est présentée figure 4.3). Les distributions $P\left(S \mid [M{=}m]\ \pi_{Ag}\right)$ apprises par l'agent dans son modèle interne se répartissent en trois groupes :

- des distributions très plates, quasiment uniformes, comme par exemple dans la région des valeurs très négatives de M, pour lesquelles l'agent n'a rien appris du tout ;

- des distributions très piquées, autour des zones des S donnés à imiter par le maître, qui montrent que l'agent a bien appris la transformation articulatori-acoustique dans ces régions-là ;

- des distributions intermédiaires, qui montrent qu'ailleurs la représentation que l'agent s'est construit dans son modèle interne est encore relativement imprécise.

En résumé, l'agent connaît bien les zones de l'espace sensori-moteur qui correspondent aux stimuli fournis par le maître, mais dans le reste de l'espace il n'a que des connaissances très partielles, qui sont issues du début de l'apprentissage.

Les répertoires moteurs appris

La figure 4.7 montre les prototypes de gestes moteurs vers lesquels l'agent a convergé après apprentissage par imitation du maître.

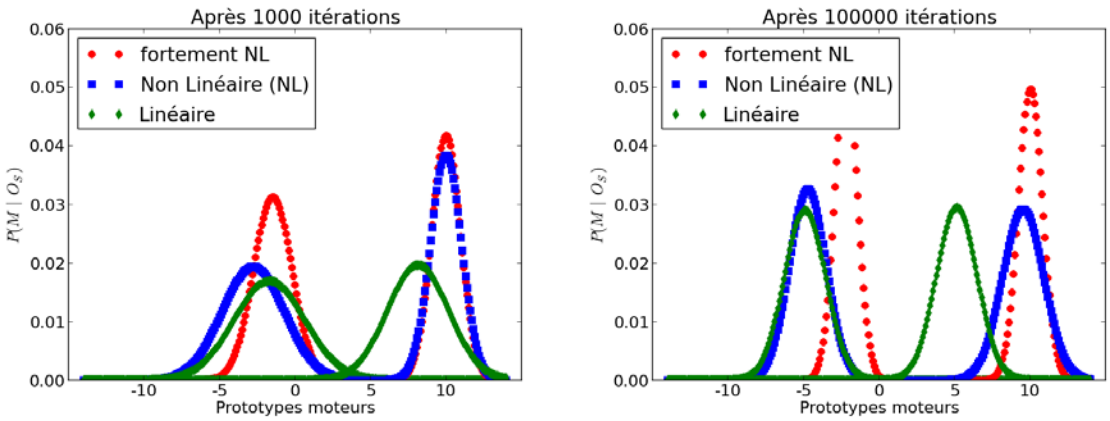


Figure 4.7: **Les répertoires moteurs appris par l'agent** lors de l'apprentissage du système moteur par accomodation. À gauche, après un petit nombre d'itérations ; à droite après convergence de l'apprentissage.

Dans le cas où la transformation articulatori-acoustique est linéaire, l'agent aligne progressivement ses productions sur celles du maître : on passe, dans l'exemple présenté sur la figure 4.7 de $\mu_{o-}^M \approx -2.5$ et $\mu_{o+}^M \approx 8$ au début de l'apprentissage à $\mu_{o-}^M = -5$ et $\mu_{o+}^M = 5$ après convergence. Ce sont en effet les seules valeurs qui peuvent permettre de bien atteindre les cibles données par le maître au cours de l'apprentissage. On observe donc que, même si l'agent semble initialement avoir ancré des choix moteurs, la nécessité de devoir imiter le maître va petit-à-petit l'amener à corriger ces choix.

Considérons maintenant les cas non-linéaires. Une transformation articulatori-acoustique non-linéaire est caractérisée par l'existence de plateaux pour lesquels une variation de la consigne motrice ne provoque que peu ou pas de variation de la conséquence acoustique. Dans ce cas l'agent a donc un grand choix de gestes moteurs permettant d'atteindre une cible. En ce sens, une transformation articulatori-acoustique non-linéaire gomme dans l'espace acoustique des différences présentes dans l'espace moteur.

Dans ce cadre, le mécanisme d'ancrage dans les répertoires moteurs des premiers choix de gestes à associer aux objets au début de l'apprentissage permet à l'agent de développer des idiosyncrasies. En effet, alors que pour atteindre une cible correspondant à l'objet o l'agent a initialement un grand nombre de gestes moteurs parmi lesquels il peut choisir, à chaque fois que l'agent choisit un geste m, il met à jour le prototype moteur correspondant à l'objet o en conséquence, ce qui a pour effet d'augmenter la probabilité que ce geste m soit associé à nouveau à l'objet o lors d'une étape future de l'apprentissage.

C'est ce qu'illustre par exemple le cas fortement non-linéaire présenté sur la figure 4.7 : au début de l'apprentissage on a les mêmes valeurs $\mu^M_{o-} \approx -2$ et $\mu^M_{o+} \approx 10$ qu'à la fin de l'apprentissage. Les choix faits initialement, qui permettaient d'atteindre les cibles à imiter, se sont renforcés, si bien que l'agent s'est construit des prototypes de gestes moteurs très différents de ceux du maître (dont les moyennes sont $\mu^M_{o-} = -5$ et $\mu^M_{o+} = 5$).

En revanche, dans le cas non-linéaire de la figure 4.7, les moyennes des prototypes moteurs de l'agent passent de $\mu^M_{o-} \approx -3$ et $\mu^M_{o+} \approx 10$ au début de l'aprentissage à $\mu^M_{o-} \approx -5$ et $\mu^M_{o+} \approx 10$ après convergence. Alors que le prototype moteur que l'agent associe à l'objet o^+ a la même moyenne au début et à la fin de l'apprentissage, celui correspondant à l'objet o^- voit sa moyenne changer. La raison en est qu'au début de l'apprentissage l'agent associe dans ses répertoires moteurs $P\left(M \mid [O_S{=}o^-]\ \pi_{Ag}\right)$ une probabilité non négligeable à des gestes moteurs $M > 0$. De tels gestes, dont l'image par la fonction sigmoïde non-linéaire sera très éloignée de la cible donnée par le maître, ont donc une probabilité très élevée de ne pas être choisis pour la tâche d'imitation. Du coup, le mécanisme d'ancrage aura pour effet de diminuer progressivement le poids relatif de ces gestes qui sont de mauvais candidats en augmentant le poids des autres, ce qui a pour effet de décaler progressivement la moyenne μ^M_{o-} de -3 vers -5. Cela montre que l'ancrage des idosyncrasies grâce au mécanisme de renforcement que nous avons proposé ne se fait pas au détriment de la capacité à atteindre les cibles proposées par le maître.

On peut donc résumer le comportement de notre algorithme d'apprentissage par accomodation à deux idées simples :

- L'accumulation de connaissances grâce au paradigme d'imitation permet de choisir des gestes permettant d'atteindre les cibles proposées par le maître.

- Si, comme c'est le cas lorsque la transformation articulatori-acoustique est non-linéaire, il y a plusieurs candidats de gestes moteurs permettant d'atteindre aussi bien les cibles, il n'y a alors rien qui pousse l'agent à converger vers les mêmes prototypes de gestes moteurs que ceux du maître. Dans ce cas, l'agent privilégie les gestes qu'il a déjà choisis par le passé, et ancre ainsi des choix idiosyncrasiques.

4 Comparaisons au sein de *COSMO* des prédictions des théories motrice, auditive et perceptuo-motrice de la perception

Dans la section précédente nous avons décrit des algorithmes permettant d'apprendre grâce aux mêmes données (les couples $< s, o >$ associant objets et entrées perceptives fournis par l'agent maître) les paramètres du système perceptif, ceux du sytème sensori-moteur, et ceux

du système moteur. Il s'agit maintenant d'évaluer comparativement les modèles auditif, moteur et perceptuo-moteur de la perception, et de tester leur robustesse aux conditions adverses en comparant leurs performances respectives en fonction du niveau de dégradation du signal dans l'environnement au moment du test. Pour cela, nous commençons par nous assurer que l'algorithme d'apprentissage par accomodation que nous avons mis en place nous a bien permis de nous écarter des hypothèses de notre théorème d'indistinguabilité.

4.1 Apprentissage et indistinguabilité

Nous avons prouvé dans la section 4 du chapitre précédent un théorème d'indistinguabilité mettant en évidence trois conditions suffisantes pour garantir l'indistinguabilité des instanciations de la tâche de perception de la parole dans le cadre des théories purement motrice et purement auditive. Il s'agit maintenant de montrer en quoi les algorithmes d'apprentissage que nous venons de proposer pour les systèmes perceptif, moteur et perceptuo-moteur permettent de s'écarter des trois hypothèses qui garantissent l'indistinguabilité.

- L'hypothèse **H1** d'apprentissage parfait suppose deux choses : une infinité d'exemplaires au cours de l'apprentissage, et une représentation interne suffisamment expressive pour capturer dans le système perceptif toute la richesse de l'information fournie par le maître. Dans l'algorithme d'apprentissage proposé, nous limitons le nombre d'itérations, et le choix d'implémenter le système perceptif comme un classifieur gaussien implique qu'il y ait une perte d'information puisque l'agent résume le lien entre les objets et les entrées perceptives sous forme de distributions de probabilité gaussiennes.

  Il est à noter que dans la branche de perception motrice a lieu une perte d'information du même ordre puisqu'il y a là aussi des choix de représentation sous forme de distributions de probabilités gaussiennes. Cependant, même si la perte d'information qui a lieu des deux côtés est du même ordre (ce qui permettra dans la suite de comparer les modèles à égalité), elle n'est pas identique car la perception motrice fait intervenir deux termes $P\left(M^{Ag} \mid O_S^{Ag}\ \pi_{Ag}\right)$ et $P\left(S^{Ag} \mid M^{Ag}\ \pi_{Ag}\right)$ et les résumés gaussiens se font aux deux niveaux, alors que la perception auditive n'en fait intervenir qu'un seul : $P\left(O_L^{Ag} \mid S^{Ag}\ \pi_{Ag}\right)$.

- L'hypothèse **H3** d'internalisation parfaite de la transformation articulatori-acoustique réalisée par l'environnement n'est pas vérifiée par nos algorithmes d'apprentissage. Puisque nous limitons le nombre d'itérations de l'algorithme d'apprentissage, il reste en effet des zones de l'espace sensori-moteur pour lesquelles l'agent n'a aucune connaissance, et d'autres portions de cet espace pour lesquelles il ne dispose que d'une information imprécise (c'est-à-dire une distribution de probabilité encore assez plate). Finalement, après apprentissage, l'agent ne connaît bien la transformation articulatori-acoustique que pour certains gestes moteurs, qui ont été choisis un nombre de fois suffisant car ils permettent de bien atteindre les cibles proposées par le maître.

- L'hypothèse **H2** d'identité motrice parfaite entre le maître et l'agent n'a aucune raison d'être vérifiée. En effet, l'agent apprend son système moteur en associant aux objets les gestes moteurs qu'il produit dans une tentative de reproduire les cibles proposées par le maître, et l'inférence probabiliste correspondant à ce processus d'imitiation fait intervenir

l'inversion probabiliste du modèle interne de la transformation articulatori-acoustique qui est lui même très imprécis. Il n'y a donc aucune raison pour que l'agent choisisse exactement les mêmes gestes moteurs que ceux utilisés par le maître, sauf si ceux-ci sont les seuls candidats possibles, comme c'est le cas lorsque la transformation articulatori-acoustique est linéaire. En revanche, dans le cas d'une transformation articulatori-acoustique non-linéaire, l'algorithme d'apprentissage que nous proposons laisse, comme nous l'avons vu, la possibilité à l'agent de réaliser des choix idiosyncrasiques qui lui sont propres et par lesquels il va en général se distinguer du maître.

4.2 Méthode de comparaison

La section précédente montre qu'il y a de bonnes raisons de s'attendre à observer des différences de prédictions entre nos implémentations auditive et motrice de la tâche de perception. Nous décrivons dans cette section la méthode que nous avons choisie pour les comparer.

4.2.1 Données d'évaluation

Pour évaluer les différentes instances de notre modèle sur les tâches de perception, nous suivons là encore notre paradigme d'interactions entre un agent π_{Ag} et un agent maître $\pi_{Maître}$ au sein d'un environnement π_{Env}. Les données qui vont servir à l'évaluation sont générées selon les étapes suivantes, en faisant varier un terme de bruit d'environnement afin de comparer la robustesse des différents modèles à des conditions dégradées.

- Le maître tire des objets o selon la distribution de probabilité uniforme $P\left(O_S^{Maître} \mid \pi_{Maître}\right)$.
- Pour chacun de ces objets o, le maître tire un geste moteur m correspondant à l'objet selon la distribution $P\left(M^{Maître} \mid \left[O_S^{Maître}{=}o\right] \ \pi_{Maître}\right)$. Nous utilisons les mêmes valeurs de paramètres que celles qui ont servi pour l'apprentissage : $\mu_o^- = -5$, $\mu_o^+ = 5$, et $\sigma_o^- = \sigma_o^+ = 1$.
- L'agent reçoit alors des stimuli perceptifs s qui correspondent aux images des m par la transformation articulatori-acoustique après perturbation lors de leur passage par l'environnement. Chacun de ces s est donc obtenu grâce à un tirage selon une distribution de probabilité gaussienne $P\left(S \mid \left[M^{Maître}{=}m\right] \ \pi_{Env}\right) = Gauss(sigmoide(m, a, b), \sigma_{Env})$. Les valeurs des paramètres a et b de pente et d'amplitude de la sigmoïde sont les mêmes que celles qui ont servi pour l'apprentissage, c'est-à-dire $b = 12$, et $a = 0.0008$, 0.8 ou 8 selon que l'on considère une transformation articulatori-acoustique linéaire, non-linéaire, ou fortement non-linéaire. En revanche, pour tester la robustesse de nos modèles à des conditions adverses, nous constituons différents *corpora* d'évaluation pour différentes valeurs de σ_{Env}, qui correspondent à différents niveaux de bruit d'environnement.

C'est ainsi que nous obtenons l'ensemble des couples $< s,\ o >$ qui serviront pour l'évaluation.

4.2.2 Implémentations de la tâche de perception

La tâche de perception consiste pour l'agent à essayer de retrouver la catégorie o du percept s qu'il reçoit du maître. Cette tâche de perception du stimulus s est réalisée au sein du modèle

COSMO en inférant grâce aux équations présentées dans la section 3.3 du chapitre 3 la distribution de probabilité $P\left(O_S \mid [S{=}s]\ \pi_{Ag}\right)$, $P\left(O_L \mid [S{=}s]\ \pi_{Ag}\right)$ ou $P\left(O_L \mid [S{=}s]\ [C{=}1]\ \pi_{Ag}\right)$ selon que l'on s'intéresse aux prédictions des théories auditive, motrice, ou perceptuo-motrice.

Notre implémentation d'une théorie auditive de la perception consiste donc à solliciter uniquement le système perceptif $P(O_L \mid S\ \pi_{Ag})$, qui est défini comme l'inversion probabiliste des prototypes auditifs gaussiens correspondant à chaque objet[27].

Notre implémentation d'une théorie motrice de la perception consiste d'après l'équation 3.20 à calculer[28] la distribution de probabilité conditionnelle

$$P(O_S \mid S\ \pi_{Ag}) \propto \sum_M \left(P(M \mid O_S\ \pi_{Ag}) P(S \mid M\ \pi_{Ag})\right)\ ,$$

ce qui peut s'interpréter simplement comme la combinaison d'un modèle inverse de la transformation articulatori-acoustique avec un décodeur articulatoire. Notre implémentation d'une théorie perceptuo-motrice de la perception consiste d'après l'équation 3.35 à calculer la distribution de probabilité

$$P\left(O_L \mid S\ [C{=}1]\ \pi_{Ag}\right) \propto P\left([O_L{=}O_S]\ \mid S\ \pi_{Ag}\right) \sum_M \Big(P(M \mid O_S\ \pi_{Ag}) P(S \mid M\ \pi_{Ag})\Big)\ ,$$

ce qui peut s'interpréter comme la fusion bayésienne des deux termes précédents. On obtient donc la distribution $P\left(O_L \mid S\ [C{=}1]\ \pi_{Ag}\right)$ en calculant le produit normalisé des termes $P(O_L \mid S\ \pi_{Ag})$ et $P(O_S \mid S\ \pi_{Ag})$ calculés précédemment.

Pour résumer, le maître produit une série de stimuli s à partir d'objets o, et à partir de ces s l'agent calcule des distributions de probabilité de la forme $P\left(O \mid [S{=}s]\ \pi_{Ag}\right)$ sur les objets reconnus. À partir de ces distributions de probabilité, nous calculons la probabilité moyenne de bonne réponse des trois modèles : auditif, moteur et perceptuo-moteur.

4.2.3 Un score global pour décrire les performances des modèles

Évaluer un modèle (qu'il soit moteur, auditif ou perceptuo-moteur) sur une tâche de perception c'est comparer les objets o communiqués par le maître avec les distributions de probabilité sur les objets reconnus par l'agent selon l'inférence correspondant à ce modèle. En effet, si l'on prend un peu de recul sur le scénario d'évaluation proposé, il s'agit d'un processus qui à chaque étape prend en entrée un objet o, tire aléatoirement un s intermédiaire selon la distribution de probabilité $P\left(S \mid \left[O_S^{Maître}{=}o\right]\right)$, et produit en sortie une distribution de probabilité sur les objets reconnus par l'agent.

Il est donc possible de résumer les performances des modèles par des distributions de probabilité de la forme $P\left(O^{Ag} \mid O_S^{Maître}\right)$, c'est-à-dire sachant les objets communiqués par le maître, quelles sont les distributions de probabilité sur les objets reconnus par l'agent. On peut alors décrire les performances d'un modèle pour la tâche de perception sous forme de matrices de confusion, telle que celle de la figure 4.8, où l'on trouve en ligne les objets communiqués par le maître, et en colonne les objets reconnus par l'agent.

[27] Pour modéliser le fait qu'un modèle gaussien n'est pas compétent pour répondre loin des moyennes de ses prototypes, avant de procéder à l'inversion bayésienne on normalise les prototypes auditifs dans l'espace des S en ajoutant une ligne de base à une valeur 10 fois inférieure à celle de la loi uniforme.

[28] De manière symétrique au cas des théories auditives, notre implémentation de la tâche de perception dans le cadre des théories motrices fait également intervenir une normalisation dans l'espace des S en ajoutant une ligne de base de même valeur que précédemment.

	o^-	o^+
o^-	$P\left(\left[O^{Ag}{=}o^-\right] \mid \left[O_S^{Maître}{=}o^-\right]\right)$	$P\left(\left[O^{Ag}{=}o^+\right] \mid \left[O_S^{Maître}{=}o^-\right]\right)$
o^+	$P\left(\left[O^{Ag}{=}o^-\right] \mid \left[O_S^{Maître}{=}o^+\right]\right)$	$P\left(\left[O^{Ag}{=}o^+\right] \mid \left[O_S^{Maître}{=}o^+\right]\right)$

Figure 4.8: **Résumé des performance des modèles** en classification sous forme de matrice de confusion. Chaque ligne correspond, pour une catégorie d'objet communiqué par le maître, à la distribution de probabilité moyenne sur les objets reconnus par l'agent.

Il y a au moins deux manières de calculer le contenu de ces matrices de confusion. Une première façon de faire consiste à suivre l'algorithme d'apprentissage en réalisant tous les tirages aléatoires successifs de valeurs pour les variables $O_S^{Maître}$, $M^{Maître}$, S^{Ag} puis O^{Ag}. Cela permet d'obtenir une approximation d'autant plus précise que le nombre de tirages est grand. Une autre façon de faire consiste à tirer parti du fait que chaque distribution de probabilité intervenant est entièrement spécifiée pour faire de l'inférence exacte. On obtient ainsi la valeur du score théorique mesurant les performances du modèle considéré sur les tâches de perception.

En terme d'implémentation, il suffit alors de calculer deux termes : la distribution de probabilité $P(S^{Ag} \mid O_S^{Maître})$ sur les stimuli produits par le maître qui parviennent à l'agent d'une part, et les réponses $P\left(O^{Ag} \mid \left[S^{Ag}{=}s\right] \ \pi_{Ag}\right)$ des trois modèles à chaque stimuli s possible d'autre part. Les valeurs des coefficients des matrices de confusion vers lesquelles converge l'évaluation de chacun des trois modèles après suffisamment d'itérations sont alors obtenues en combinant ces deux termes ainsi que le montre l'équation suivante :

$$P\left(O^{Ag} \mid O_S^{Maître}\right) = \sum_{s \in S} P\left([S{=}s] \mid O_S^{Maître}\right) \times P\left(O^{Ag} \mid \left[S^{Ag}{=}s\right] \ \pi_{Ag}\right) \ . \tag{4.9}$$

À partir de ces matrices de confusion, on définit un indicateur global : le taux de reconnaissance, qui se calcule comme étant la moyenne des coefficients diagonaux de la matrice de confusion. Ce taux de reconaissance décrit la précision en reconnaissance du modèle considéré, moyennée sur les différentes catégories d'objets.

4.3 Résultats : comparaisons des performances des différents modèles

Les résultats de simulation que nous allons présenter maintenant montrent l'évolution de ce taux de performance en fonction d'un certain niveau de dégradation du signal de parole. Les paramètres des modèles auditif et moteur ont été appris par des interactions entre l'agent et le maître au sein d'un environnement qui perturbait la transmission du signal acoustique du maître vers l'agent en ajoutant un bruit gaussien d'écart-type σ_{Env}. Il s'agit maintenant de faire varier au moment de l'évaluation des modèles ce paramètre σ_{Env} qui contrôle le niveau de dégradation du signal.

4.3.1 Des dynamiques d'évolution différentes dans la branche auditive et dans la branche motrice.

La figure 4.9 permet de comparer l'évolution des performances des modèles moteur et auditif de la perception sur des tests en conditions dégradées au cours de l'apprentissage.

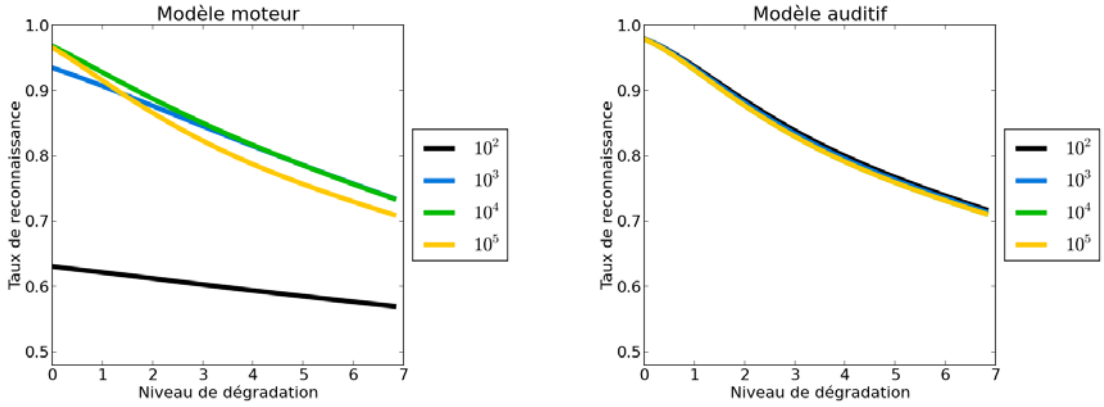


Figure 4.9: **Évolution de la robustesse aux dégradations au cours de l'apprentissage**, pour les modèles moteur (à gauche) et auditif (à droite), dans le cas non-linéaire. Chaque courbe correspond à un modèle pour lequel l'apprentissage à été interrompu après un certain nombre d'itérations. On compare ainsi quatre ordres de grandeurs différents : de 10^2 à 10^5. Le niveau de dégradation 0 revient à faire du test dans les conditions d'apprentissage, et le niveau de bruit 7 revient à tirer des perturbations selon une distribution de probabilité gaussienne dont l'écart-type est égal à environ un quart de la taille de l'espace perceptif. Un taux de reconnaissance de 1.0 correspond à un modèle parfait, et un taux de reconnaissance de 0.5 correspond à un modèle qui répond au hasard (uniformément).

Les données de la figure 4.9 montrent que les modèles auditifs et moteurs ont des dynamiques d'évolution différentes : le modèle audio apprend rapidement et arrive très vite à saturation (100 exemplaires lui suffisent pour arriver à de bonnes performances), alors que le modèle moteur évolue plus lentement. En effet, les performances du modèle auditif n'évoluent pas lorsque le nombre d'itérations de l'algorithme d'apprentissage dépasse la valeur de 100 (toutes les courbes sont quasiment superposées), alors que les performances du modèle moteur continuent de s'améliorer en conditions normales, au détriment d'une baisse de robustesse en conditions dégradées. Il convient de faire un parallèle avec le phénomène de surapprentissage classique dans le domaine du *machine learning*. Plus l'apprentissage de nos modèles auditif et moteur dure, plus ils se spécialisent sur leur corpus d'apprentissage et plus leurs performances lors de tests en conditions normales s'améliorent, mais cela se fait au détriment de leurs capacités de généralisation, ce qui se traduit par une baisse de performances en conditions dégradées.

On peut également interpréter ces résultats à la lumière de notre théorème d'indistinguabilité et de son hypothèse **H1** d'apprentissage parfait. Même si le choix de distributions de probabilité gaussiennes pour encoder le classifieur auditif de l'agent ne permet pas de capturer dans toute sa finesse l'information contenue dans le modèle du maître et dans celui de l'environnement, après un grand nombre d'itérations des algorithmes d'apprentissage, les modèles moteur et auditif de l'agent montrent des performances très similaires.

De cette figure 4.9 nous retiendrons principalement l'idée que, que ce soit pour de l'apprentissage ou pour de l'adaptation en ligne, le modèle auditif a une dynamique d'évolution beaucoup plus rapide que celle du modèle moteur. En effet, cela s'explique par le fait que dans la voie auditive il suffit d'apprendre des associations directes entre les stimuli et les objets, alors que l'apprentissage de la voie motrice nécessite d'apprendre à la fois un modèle interne de la transformation articulatori-acoustique et les associations entre les objets et les gestes moteurs.

4.3.2 Comparaison de la robustesse aux dégradations des différents modèles

La figure 4.10 montre comment évolue le taux de reconnaissance des trois modèles en fonction du niveau de dégradation au moment du test, dans le cas non-linéaire (pour une valeur de pente $a = 0.8$), et lorsque l'apprentissage des modèles a été interrompu après 1000 itérations (on se place donc dans un cas où l'apprentissage du modèle auditif a déjà convergé, alors que celui du modèle moteur pourrait encore continuer).

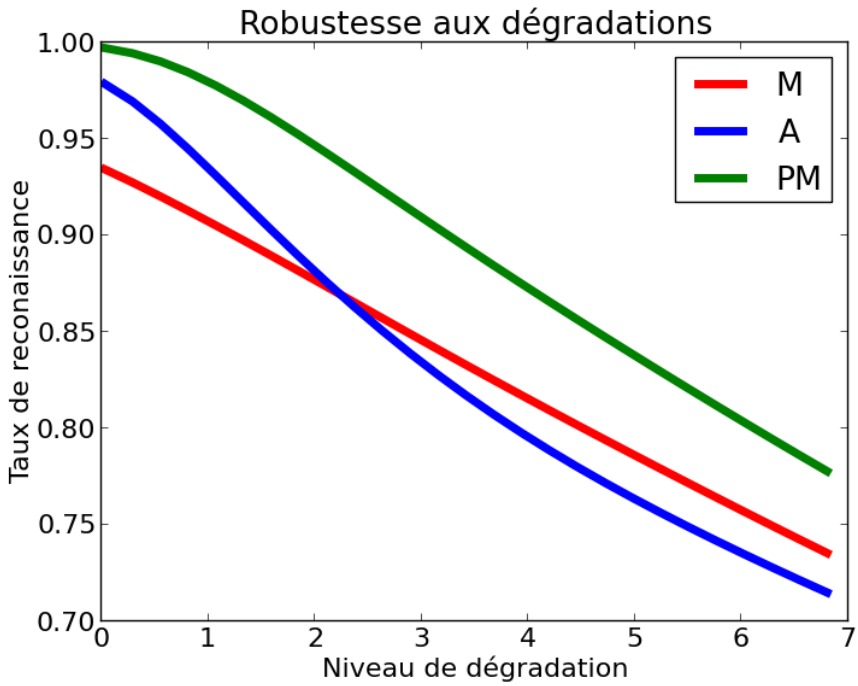


Figure 4.10: **Comparaison de la robustesse aux dégradations** des modèles moteur (M), auditif (A), et perceptuo-moteur (PM) dans le cas non-linéaire (pour une valeur de pente $a = 0.8$), après 1000 itérations de l'algorithme d'apprentissage.

La figure 4.10 montre que le modèle auditif a de meilleures performances que le modèle moteur en conditions normales, mais que celui-ci prend l'avantage en conditions dégradées. Pour interpréter ces résultats, il est bon de noter que, bien que l'on compare un modèle auditif avec un modèle moteur, ce qui est important est surtout le fait que le premier soit totalement appris, alors que le second n'a pas fini son apprentissage (le chapitre 3 suggère d'ailleurs qu'avec un apprentissage infini des deux côtés on retombe sur le cas d'indistinguabilité). Cette différence de performances entre le modèle auditif et le modèle moteur tient alors essentiellement au sous-apprentissage de la transformation articulatori-acoustique : le modèle interne appris par l'agent

est incomplet et imparfait, il ne connaît bien que les zones correspondant aux stimuli fournis par le maître au cours de l'apprentissage. Bien que cela puisse semble paradoxal, c'est justement cette connaissance seulement partielle du reste de l'espace sensori-moteur qui lui apporte sa robustesse en conditions dégradées.

Finalement, les résultats de la figure 4.10, obtenus dans le cas où l'apprentissage est limité, et pour une transformation articulatori-acoustique non-linéaire, peuvent se résumer ainsi :

- Le modèle auditif, facile à apprendre, est très spécialisé, et a d'excellentes performances en conditions normales.
- Le modèle moteur, moins bien appris, est également plus polyvalent et montre plus de robustesse aux conditions dégradées.
- Le modèle perceptuo-moteur, qui réalise la fusion des deux, combine ces avantages d'être à la fois spécifique et d'avoir des capacités de généralisation.
- L'imprécision du modèle moteur sur une tâche de perception en conditions normales s'explique par l'imprécision du modèle interne appris pour la transformation articulatori-acoustique (voir figure 4.6). En revanche, la figure 4.7 laisse penser que cette imprécision ne remettrait pas en cause la précision du modèle moteur sur des tâches de production dans le cadre d'une théorie motrice. [29]

4.3.3 Impact de la non-linéarité sur les performances des modèles

La figure 4.11 montre comment évolue le taux de reconnaissance des trois modèles en fonction du niveau de dégradation au moment du test, pour les différents degrés de non-linéarité, à pour différents volumes d'apprentissage (entre 1000 et 100000 itérations).

On y voit tout d'abord que, quel que soit le degré de non-linéarité, ou la quantité de données auxquelles les modèles moteur et auditif ont été exposés durant l'apprentissage, le modèle perceptuo-moteur, qui en réalise la fusion, est toujours meilleur que les deux autres. Pour ce qui est des modèles auditif et moteur, il semble qu'augmenter le degré de non-linéarité ait le même effet qu'augmenter le nombre d'itérations de l'algorithme d'apprentissage : dans un premier temps passer de 0.0008 à 0.8 augmente les performances des modèles, puis passer de 0.8 à 8 les dégrade. Cela est dû au fait que, comme le montre la figure 4.4, la non-linéarité a pour effet de concentrer les productions du maître dans de petites zones de l'espace acoustique, ce qui fait que les modèles appris par l'agent sont d'autant plus piqués que le degré de non-linéarité est élevé. Ainsi, dans un premier temps augmenter le degré de non-linéarité permet de gagner en précision, jusqu'à ce que le surapprentissage fasse diminuer les performances des modèles en conditions dégradées.

5 Conclusion

Dans ce chapitre nous avons proposé une première instance du modèle *COSMO* dans un cadre théorique dont la simplicité nous permet de bien mettre en évidence les propriétés des algo-

[29] En revanche, dans ce cas c'est le modèle auditif serait plus imprécis puisqu'il fait intervenir dans son calcul de la tâche de production le modèle interne de la transformation articulatori-acoustique (voir le tableau 3.5 du chapitre 3).

Évolution des taux de performance en fonction du niveau de dégradation

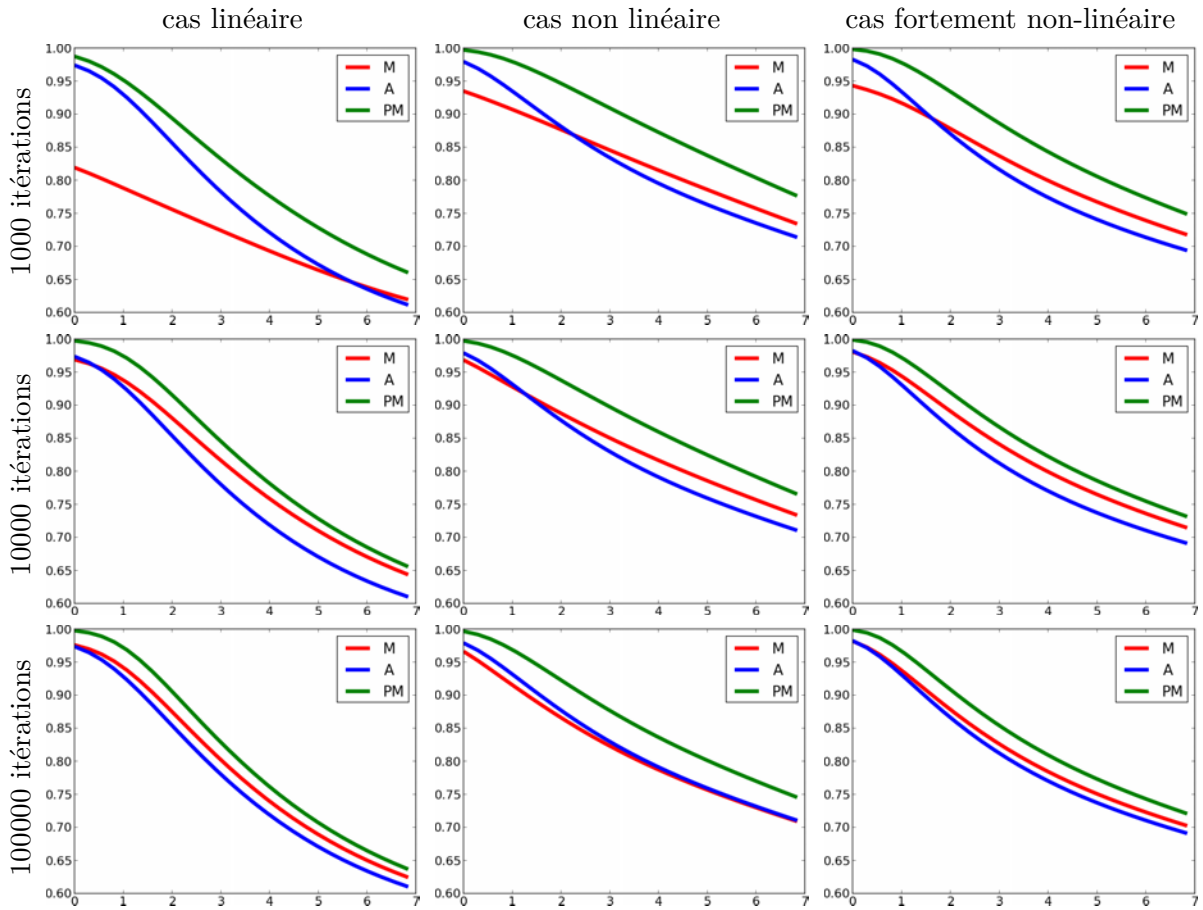


Figure 4.11: **Comparaison de la robustesse aux dégradations** des modèles moteur (M), auditif (A), et perceptuo-moteur (PM). Sur chacune de ces neufs figures les axes sont les mêmes que sur la figure 4.10, à savoir le niveau de dégradation en abscisse, et le taux de reconaissance en ordonnée. Chaque ligne correspond à un volume d'apprentissage donné, avec des valleurs allant de 1000 à 100000 itérations de l'algorithme d'apprentissage. Chaque colonne correspond à un degré de non-linéarité donné : la pente de 0.0008 correspond au cas linéaire, 0.8 au cas non-linéaire, et 8 au cas fortement non-linéaire.

rithmes d'apprentissages et des modèles appris, qui seront ensuite généralisées dans le chapitre 6 au cas des syllabes. Avec les simulations théoriques de ce chapitre, nous avons présenté les contributions suivantes.

Un algorithme original d'apprentissage par accomodation permet d'apprendre des compétences motrices à partir d'entrées perceptives uniquement. Cet algorithme combine un principe d'affinage progressif du modèle interne de la transformation articulatori-acoustique avec un mécanisme de renforcement permettant d'ancrer progressivement des choix moteurs et de développer ainsi des idiosyncrasies. Avec ce paradigme d'accomodation, l'agent concentre son apprentissage sur les zones de l'espace sensori-moteur qui lui permettent d'atteindre les cibles proposées par le maître, et n'a du reste que des représentations imprécises.

Une explication de la plus grande robustesse au bruit du modèle moteur : alors que le modèle auditif, qui est très facile à apprendre, se spécialise très vite sur les zones de l'espace correspondant à son apprentissage et n'est plus compétent ailleurs, le modèle moteur est intrinsèquement plus imprécis (à cause de la difficulté d'apprendre une transformation articulatori-acoustique complexe), ce qui lui confère de meilleures capacités de généralisation.

La voie motrice : précise en production, imprécise en perception. Le modèle moteur de la perception fait intervenir deux termes : des prototypes moteurs de gestes fréquemment associés aux objets, et un modèle interne qui encode des connaissances sur la transformation articulatori-acoustique. Alors que le premier terme peut encoder des idiosyncrasies et être très précis (il est facile à apprendre, et correspond à des distributions piquées, ce qui explique également la consistance observée sur des tâches de production) le second terme est bien plus difficile à apprendre, et c'est là que réside l'imprécision inhérente au modèle moteur qui lui apporte de la robustesse au bruit.

Le rôle de la non-linéarité. La non-linéarité de la transformation articulatori-acoustique rend possible le développement d'idiosyncrasies motrices, qui n'ont pas d'incidence perceptive. La non-linéarité a également pour effet d'accélérer la convergence des apprentissages. Par ailleurs, on observe de plus qu'en termes de scores sur des tâches de perception dans le bruit, augmenter le degré de non-linéarité revient essentiellement à augmenter le nombre d'itérations de nos algorithmes d'apprentissage.

Des dynamiques d'évolution différentes. Alors que dans le système auditif il suffit d'apprendre des associations directes entre les stimuli et les objets, ce qui est rapide, l'apprentissage du système moteur est plus lent car il nécessite d'apprendre à la fois un modèle interne de la transformation articulatori-acoustique et les associations entre les objets et les gestes moteurs. Notre analyse des dynamiques d'évolution lors de l'apprentissage peut se généraliser à de l'adaptation en ligne, et nous avons obtenu dans ce cadre des résultats quant à la robustesse du modèle moteur aux conditions dégradées similaires à ceux présentés à la figure 4.10 (voir Moulin-Frier *et al.* (2012) où nous avons comparé un modèle moteur constant à un modèle auditif qui capture parfaitement les propriétés de son environnement à chaque instant). Ainsi, le modèle moteur permet de garder des connaissances stables, peu évolutives, alors que le modèle auditif s'adapte rapidement.

Le rôle de la fusion perceptuo-motrice. La fusion audio-motrice en perception permet de tirer parti avantageusement de la combinaison d'un système auditif précis, facilement adaptable, mais peu robuste aux dégradations avec un modèle moteur évoluant plus lentement auquel son imprécision apporte de la robustesse.

Bien que le cadre de travail que nous nous sommes donné soit très fruste (les variables motrice M et sensorielle S sont unidimensionnelles, le choix perceptif est restreint à deux objets) il a pu permettre de mettre en évidence les principes que nous venons de lister. Nous obtenons ainsi une première illustration de l'intérêt potentiel de combiner des connaissances perceptives et motrices pour améliorer les performances d'un système de reconnaissance de parole. Ce résultat fait écho aux données expérimentales présentées dans la section 2.2 du chapitre 2 qui montraient une activité accrue et un rôle fonctionnel possible du système moteur en situation adverse. Il fournit également un support d'explication et de raisonnement aux résultats de modèles computationnels présentant un gain de performance lorsque la voie motrice est activée, particulièrement dans le bruit (Badino *et al.*, 2014).

Bien évidemment, cette simulation dans un cadre extrêmement simplifié demande à subir le « passage à l'échelle » vers des signaux plus complexes. C'est dans cette voie que nous allons nous engager dans le chapitre suivant, en remplaçant des espaces moteur et auditif unidimensionnels abstraits par des espaces plus proches de ceux de la parole naturelle.

Chapitre 5

Synthèse de syllabes réalistes dans le cadre de *COSMO* avec un modèle géométrique de conduit vocal : *VLAM*

1 Modélisation de la syllabe **88**

2 Synthèse de syllabes plosive-voyelle dans le cadre de *COSMO* **102**

3 Évaluation des syllabes plosive-voyelle produites **105**

4 Conclusion **110**

Nous avons présenté au chapitre 3 *COSMO*, notre modèle générique d'agent communicant, et nous avons décrit au chapitre 4 une instanciation de ce modèle dans un cadre purement théorique, ce qui a permis d'étudier des algorithmes d'apprentissage de capacités auditives et motrices ainsi que les propriétés des modèles appris (dynamiques d'évolution, développement d'idiosyncrasies, robustesse aux conditions dégradées...).

L'objectif du présent chapitre est de montrer comment nous avons construit un ensemble de données de syllabes, caractérisées à la fois dans un espace articulatoire et dans un espace acoustique, afin de pouvoir ensuite dans le chapitre 6 adapter le modèle *COSMO* et généraliser les résultats du chapitre 4 au cas des syllabes. Pour cela nous choisissons d'utiliser *VLAM* (the *Variable Lineary Articulatory Model*) qui est un modèle de conduit vocal permettant la synthèse de trajectoires articulatoires et acoustiques.[30]

Ce chapitre commence par un survol rapide de la bibliographie permettant d'expliquer les choix que nous avons faits sur la manière de modéliser dans des espaces moteurs et perceptifs le signal de parole en général et les syllabes de type plosive-voyelle en particulier. Nous en retenons des principes simples permettant de synthétiser avec *VLAM* des syllabes ayant des patrons de variabilité similaires aux données réelles de parole.

[30] Le modèle *VLAM* sera présenté plus en détails à la section 1.2.1 du présent chapitre.

1 Modélisation de la syllabe

Dans le modèle *COSMO* présenté aux chapitres 3 et 4 nous avons introduit des représentations M et S qui sont des abstractions de représentations motrices et auditives, associées au signal de parole. Dans cette section les variables génériques M et S sont instanciées de manière plus précise, et nous présentons les choix de modélisation que nous avons faits à deux niveaux :

1. Nous choisissons de ne sélectionner que certains paramètres pour décrire le signal de parole dans les espaces des représentations motrices et auditives.

2. Nous adoptons une vision très simplifiée de la syllabe selon laquelle une syllabe de type plosive-voyelle peut être décrite par deux événements : un état stable portant l'information de la voyelle, qui est précédé par un geste de plosion. Ces deux états sont reliés par des effets de coarticulation.

1.1 Différentes manières de modéliser le signal de parole

Nous avons présenté au chapitre 2 les deux grands courants cognitifs que sont les théories motrices et auditives de la parole. La principale différence qui les oppose concerne la nature des espaces dans lesquels ils choisissent de décrire l'objet parole : soit à partir de représentations motrices, soit à partir de représentations auditives. Dans le chapitre 2 nous avons décrit les principaux enjeux du débat théorique ; il s'agit maintenant de voir comment les théories motrices et auditives peuvent être implémentées de manière concrète dans un cadre computationnel. Plus précisément, la question est de savoir comment caractériser le signal de parole dans un espace moteur et dans un espace auditif, et quels sont les paramètres permettant de capturer l'information qui y est exprimée ?

1.1.1 Comment caractériser la parole dans un espace auditif

Nous proposons de regarder comment, en amont de tout traitement neuronal, notre système auditif analyse et recode le signal de parole. La description qui va suivre est largement inspirée de Romand (2000).

Alors que l'oreille externe et l'oreille moyenne ont principalement un rôle de conditionnement du signal de parole arrivant en entrée (conduction, protection, adaptation d'impédance acoustique), la cochlée, qui est l'organe principal de l'oreille interne, réalise la première étape majeure du processus de traitement auditif du signal acoustique. La figure 5.1 présente une vue schématique de la cochlée.

L'étrier transmet aux liquides de l'oreille interne une onde de pression acoustique qui déforme maximalement la membrane basilaire en un lieu qui dépend de la fréquence. Les propriétés mécaniques de la membrane basilaire varient systématiquement en fonction de la position longitudinale le long de la cochlée : la zone proche de la base, plus fine et légère, répond mieux aux hautes fréquences, et à l'inverse la zone proche de l'apex, plus lourde et épaisse, répond mieux aux basses fréquences. Finalement, chaque point de la membrane basilaire a une fréquence caractéristique à laquelle elle répond optimalement. Ensuite, les cellules ciliées convertissent les déformations mécaniques en messages nerveux qui sont transmis *via* le nerf auditif au cerveau.

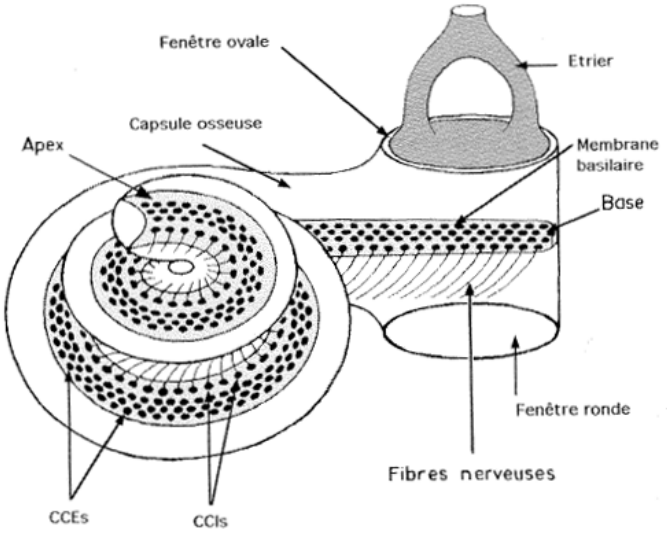


Figure 5.1: **Représentation schématique de la cochlée**, d'après Romand (2000). Les cellules ciliées (*CCis* pour internes, et *CCes* pour externes) sont réparties le long de la *membrane basilaire*, qui va de la *base* à son *apex*. Par le biais de la transduction mécano-électrique, elles transforment les vibrations sonores en influx nerveux ensuite transmis par les *fibres nerveuses* pour être interprétés par le cerveau.

En présence d'un son pur de fréquence f, seul un petit nombre de cellules ciliées seront sollicitées : celles dont la fréquence caractéristique est suffisamment proche de f. En résumé, la cochlée permet de convertir le signal acoustique vers le domaine fréquentiel, ce qui revient à réaliser une analyse de Fourier. Par conséquent, les traitements auditifs qui ont lieu ensuite dans le cerveau ont comme point de départ une représentation de type fréquentielle.

Une caractéristique importante de l'analyse fréquentielle ayant lieu dans la cochlée est la transformation non-linéaire de l'échelle des fréquences qui s'y produit : en effet, la distribution des fréquences caractéristiques en fonction des positions des cellules ciliées sur la membrane basilaire est semi-logarithmique, c'est-à-dire linéaire en basse fréquence puis logarithmique à partir de 1000 Hz environ. Pour rendre compte de cette transformation perceptive, plusieurs échelles ont été introduites, et nous utiliserons ici le Bark (Bk). Une formule a été proposée par Schroeder *et al.* (1979) pour convertir les Hertz en Bark :

$$z(Bark) = 7\ Argsh\left(\frac{F(Hz)}{650}\right)\ .$$

Chistovich (1980) propose que le traitement auditif des sons de parole soit réalisé dans le cerveau par deux systèmes schématisés figure 5.2 : le premier est spécialisé dans la détection d'événements temporels et le second dans l'analyse du spectre.

D'après Schwartz *et al.* (1992), la détection d'événements (« Quand ? ») pourrait se faire grâce à des neurones « on/off », dont l'activation serait déclenchée par des variations brusques du taux d'excitation des groupes de neurones associés à une zone du spectre. Wu *et al.* (1996) proposent ainsi un modèle physiologiquement plausible pour la détection d'événements articulatori-

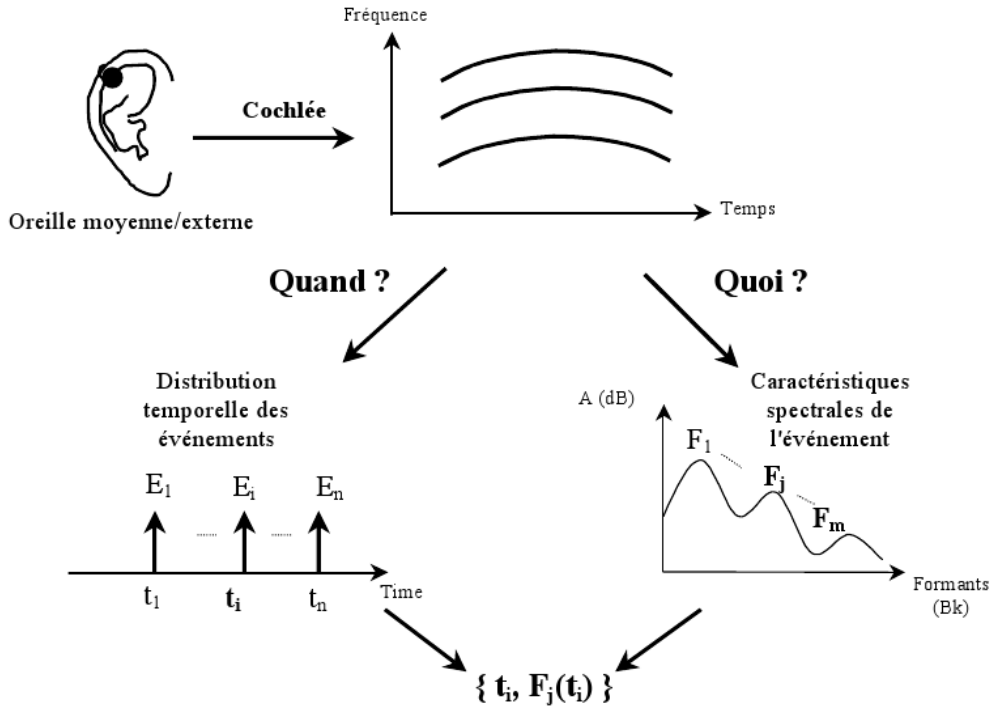


Figure 5.2: **Représentation schématique des traitements auditifs**, d'après Serkhane (2005).

acoustiques dans le noyau cochléaire, tels que le début et la fin du voisement, les bruits de plosion ainsi que le début et la fin des productions vocaliques.

Toujours d'après Schwartz *et al.* (1992), le système spécialisé dans la caractérisation des événements (« Quoi ? ») fonctionnerait grâce à une analyse continue du signal de parole permettant des traitements statistiques précis des variations de taux d'excitation en fonction des fréquences caractéristiques du signal. Ce type de traitements pourrait ainsi permettre, comme semblent le montrer Young et Sachs (1979), d'extraire du spectre ses pics d'intensité.

La question est alors de savoir quelle est la finesse des traitements temporels et spectraux, et le nombre (ou la granularité) des événements temporels et des caractéristiques spectrales ainsi extraits. En ce qui concerne les événements temporels, on peut imaginer une gamme de fréquences d'analyse allant de micro-événements locaux (comme des modulations locales d'intensité à une échelle infra-phonémique) jusqu'à des événements segmentaux majeurs (tels que le début du voisement, l'explosion acoustique (en anglais, *burst*) des plosives ou le début de friction des fricatives, les *maxima* et *minima* des trajectoires formantiques entre voyelles ou au sein de séquences consonne-voyelle ou voyelle-consonne, etc) ou même à des événements prosodiques (les pics accentuels, les marqueurs de tour de parole, ...).

Les développements de l'analyse de scènes auditives dans les années 90, avec l'introduction de la notion de primitives auditives, ont cristallisé ces discussions (voir par exemple Cooke et Ellis (2001)), qui ont ensuite été relancées par les mises en évidence du codage neuronal multiplexe (voir une revue récente par Giraud et Poeppel (2012)). Nous verrons dans une section ultérieure quels choix nous semblent adéquats pour traiter des séquences plosive-voyelle qui sont la cible

de nos travaux dans le présent chapitre et dans le suivant.

En ce qui concerne les caractéristiques spectrales extraites, les débats se sont cristallisés dans les années 70-80 sur le rôle des formants, résonances du conduit vocal (voir Fant (1960)) qui sont les *maxima* d'énergie du spectre sonore du son de parole. Si, du point de vue comportemental, les arguments sont contrastés (Pols, 1975 ; Carlson *et al.*, 1979), les mécanismes d'inhibition latérale disponibles dès le noyau cochléaire semblent fournir la trame nécessaire à des détections de *maxima* spectraux compatibles avec la détection des formants, qui sont les composantes spectrales cruciales pour caractériser le lien articulatori-acoustique.

Finalement, nous retiendrons de cette section l'idée que le signal de parole peut être décrit par un petit nombre d'événements significatifs, qui sont détectés lors des premières étapes du traitement auditif, et qui peuvent être caractérisés dans le domaine fréquentiel par les formants sur une échelle perceptive en Bark.

1.1.2 Comment caractériser la parole dans un espace moteur

La production de la parole est une tâche fortement complexe qui fait intervenir l'action coordonnée de plus d'une centaine de muscles respiratoires, laryngaux, pharyngaux et orofaciaux (Simonyan et Horwitz, 2011 ; Levelt, 1993). La déformation des structures respiratoires, laryngées et du conduit vocal permet de créer des sources sonores, de les modifier, de les amplifier et de les filtrer.

Devant cette complexité, plusieurs types de modèles du conduit vocal ont été proposés, qui présentent différents degrés de simplification.

Des modèles purements géométriques. Les caractéristiques acoustiques du signal de parole ne dépendant que de la géométrie du conduit vocal, certains travaux ont proposé de modéliser le conduit vocal sous forme de cavités résonantes équivalentes d'un point de vue acoustique.

Ainsi, Stevens et House (1955) et Fant (1971) ont proposé une première approximation qui consiste à résumer le comportement du conduit vocal par quatre tubes : un tube représentant l'espace entre les lèvres, deux tubes pour les cavités avant et arrière, qui sont séparées par un quatrième tube représentant la constriction du conduit vocal au niveau de la langue.

D'autres modèles plus réalistes (Mermelstein, 1973 ; Coker, 1976) utilisent des formes géométriques simples (des lignes, des arcs de cercle) contrôlées par des paramètres pour construire un contour approximatif d'un conduit vocal complet. Le modèle de Mermelstein (1973) permet par exemple, en choisissant des valeurs adéquates pour ses neufs paramètres, de reconstituer les formes de conduits vocaux correspondant aux voyelles et aux consonnes observées dans des données de cinéradiographie (Perkell, 1969).

Dans les modèles où le conduit vocal est représenté par une série de tubes cylindriques, la précision peut être améliorée en augmentant le nombre de tubes (Badin et Fant, 1984) ou en en changeant la forme (Schoentgen et Ciocea (1995) utilisent par exemple des segments de conduit ayant une forme conique, ce qui permet d'obtenir des fonctions d'aire continues).

Des modèles dont les paramètres articulatoires sont issus d'analyses statistiques. L'une des principales limites des modèles purement géométriques tient au fait que, leurs paramètres étant artificiels, ils sont difficiles à interpréter en termes de contrôle moteur, et certaines combinaisons de paramètres peuvent conduire à des formes de conduits vocaux qui ne sont pas réalisables en pratique par des humains. Par ailleurs, ces variables géométriques sont difficiles à défendre du point de vue cognitif : même si dans le cas des plosives il est envisageable de penser qu'un retour somatosensoriel puisse permettre de connaître la position de la constriction, dans le cas des voyelles, qui sont ouvertes, il est beaucoup plus discutable de supposer que l'on ait accès à tout moment à la géométrie précise du conduit vocal (et en particulier aux positions et tailles des constrictions).

À l'inverse, les modèles statistiques (comme par exemple ceux décrits par Harshman *et al.* (1977) ; Maeda (1979, 1990) ; Tiede *et al.* (1996) ; Beautemps *et al.* (2001)) sont basés sur un petit nombre de paramètre articulatoires, qui sont choisis (par exemple grâce à une Analyse en Composantes Principales, ou ACP, voir Jolliffe (2005)) à la fois pour permettre de rendre compte de la variabilité des formes de conduits vocaux dans les données d'apprentissage, et pour qu'ils soient interprétables en termes de commandes musculaires, phonétiques et acoustiques. Ces modèles sont typiquement appris à partir de contours de conduits vocaux dans le plan sagittal, qui sont extraits à partir d'images cinéradiographiques (Moll, 1960 ; Bothorel, 1986) ou obtenus par imagerie à résonance magnétique (Tiede *et al.*, 1996).

Des modèles biomécaniques. Les modèles biomécaniques sont des modèles de conduits vocaux qui intègrent des propriétés physiologiques des articulateurs et leurs interactions. Un des premiers modèles physiologiques de la langue a été développé par Perkell (1974). Il s'agit d'un modèle de type masses-ressorts, basé sur une étude anatomique de la langue, qui est composé de nœuds porteurs de masses qui sont reliés entre eux et/ou à des éléments générateurs de tensions, de manière active (pour les tissus musculaires) ou passive (pour les tissus conjonctifs et les structures rigides ou molles du conduit vocal).

Des travaux plus récents se distinguent par l'utilisation de mesures plus précises des activités musculaires, par l'utilisation de méthodes par éléments finis (Zienkiewicz et Taylor, 2005), par le développement de modèles tridimensionnels, et par la génération dynamique de trajectoires articulatoires. À titre d'illustration, on peut citer par exemple les travaux de Payan et Perrier (1997) ; Sanguineti *et al.* (1998) ; Perrier *et al.* (2003) ; Nazari *et al.* (2013) ; Stavness *et al.* (2014) qui sont emblématiques de ce qui se fait dans ce domaine, et on pourra trouver dans Perrier *et al.* (2011) une revue des principes sur lesquels s'appuient l'élaboration de tels modèles.

Ces modèles biomécaniques, qui prennent en compte des éléments de masse, de viscosité, de raideur ou d'élasticité contribuent à l'étude de phénomènes fondamentaux de la parole tels que l'anticipation, la coarticulation et l'adaptation aux perturbations. Plus ambitieux dans ce qu'ils essayent de décrire, ils ont également l'inconvénient d'être souvent plus gourmands en nombre de paramètres et en temps de calcul.

L'objectif de cette section n'était que de proposer un aperçu rapide de la variété des approches et des modèles de conduits vocaux proposés dans la littérature, le lecteur désireux de

plus de détails pourra se référer par exemple à Gabioud (1994) ; Busset (2013).

1.2 Le choix du modèle *VLAM* pour assurer le passage des représentations motrices aux représentations auditives

Les modèles que nous avons décrits se placent à différents niveaux d'abstraction pour décrire le conduit vocal. Pour le reste de ce chapitre nous en choisissons un qui se trouve en position intermédiaire entre un résumé géométrique de haut niveau (telle que la donnée des lieus et tailles des constrictions) et une description bas niveau des propriétés des muscles et des tissus. L'intérêt de ce niveau intermédiaire est qu'il est interprétable en terme de commandes musculaires, phonétiques et acoustiques.

1.2.1 *VLAM* : the *Variable Linear Articulatory Model*

VLAM, the *Variable Linear Articulatory Model* (Boë, 1999), est un modèle articulatoire basé sur le modèle de Maeda (1990) qui est issu de l'analyse statistique d'images radiographiques et labiographiques (Bothorel, 1986) correspondant à dix phrases en français. À partir de ces coupes dans le plan sagittal, 519 contours de conduit vocal ont été extraits à la main, puis analysés selon une grille semi-polaire de référence qui divise le conduit vocal en 28 sections (Maeda, 1988). Une analyse en composantes principales guidée aboutit à sept paramètres articulatoires, présentés figure 5.3. Ils décrivent la position de la mandibule et du larynx, la forme de la langue et des lèvres, sont interprétables en termes de commandes phonétiques, et sont très proches de commandes musculaires (Maeda et Honda, 1994).

Les quatre premiers paramètres permettent de rendre compte de 88% de la variance observée dans les contours de la langue (Maeda, 1990). La forme des lèvres a été modélisée à partir de mesures réalisées sur un locuteur différent (Abry et Boë, 1986). La sensibilité de chacun de ces sept paramètres articulatoires a été normalisée en utilisant l'écart-type autour de la position moyenne observée dans les données.

Une équation linéaire combinant ces sept paramètres permet de regénérer les contours d'un conduit vocal dans le plan sagittal (Gabioud, 1994). On calcule alors la surface S de chacune des 28 intersections de ces contours sagittaux avec la grille semi-polaire de référence en utilisant la formule de Heinz et Stevens (1965) : $S = \alpha d^{\beta}$, où d est la distance sagittale, et α et β sont des coefficients dérivés d'études tomographiques réalisées par Perrier *et al.* (1992). La fonction d'aire ainsi calculée permet ensuite de déduire la fonction de transfert du conduit vocal ainsi que les formants (Badin et Fant, 1984). Finalement, il est possible de générer du son à partir des valeurs des formants grâce à un module de synthèse développé par Berthommier *et al.* (2012).

Guenther (2006, page 352) défend l'utilisation de *VLAM* qu'il utilise dans *DIVA* : « Le modèle repose sur des simulations informatiques qui contrôlent un synthétiseur articulatoire (Maeda, 1990) capable de produire un signal acoustique. Les trajectoires articulatoires et acoustiques produits par le modèle sont comparables aux productions de locuteurs humains ; les résultats de nombreuses comparaisons de la sorte sont décrits par d'autres travaux (e.g., Callan *et al.* (2000) ; Guenther (1995) ; Guenther *et al.* (1998, 1999) ; Nieto-Castanon *et al.* (2005) ; Perkell *et al.* (2004a,b)). »[31]

[31] « The model is implemented in computer simulations that control an articulatory synthesizer (Maeda, 1990) in order to produce an acoustic signal. The articulator movements and acoustic signal produced by the model

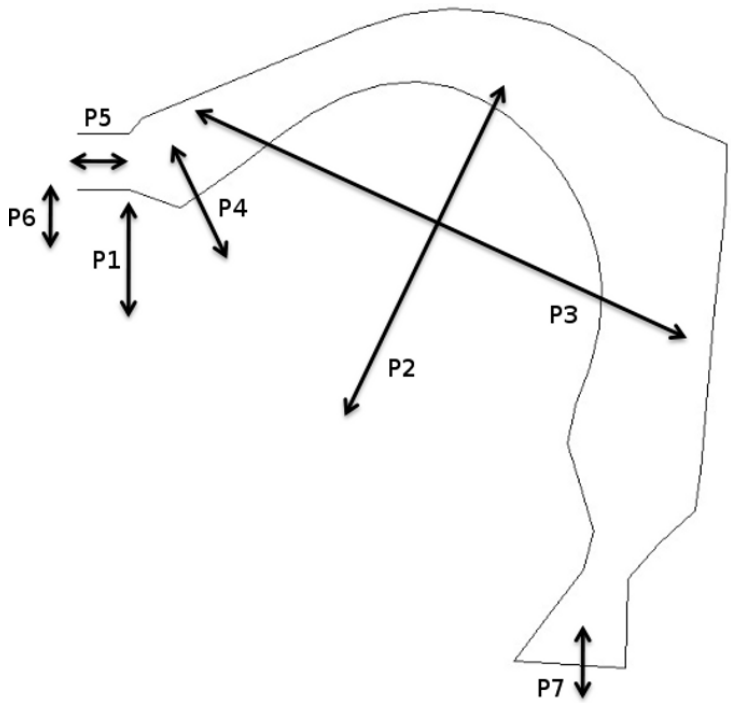


Figure 5.3: **Les paramètres articulatoires du modèle *VLAM***. P1 (*Jaw*) permet de contrôler les mouvements verticaux de la mandibule, P2 (*TongueDorsum*) la courbure/applatissement du dos de la langue, P3 (*TongueBody*) la protrusion/rétraction du corps de la langue, P4 (*Apex*) les mouvements verticaux de la pointe de la langue, P5 (*LipProtrusion*) la protrusion des lèvres, P6 (*LipHeight*) l'écartement des lèvres et P7 (*Larynx*) l'élévation du larynx.

Le modèle *VLAM* a été par la suite systématiquement testé et amélioré par les chercheurs du laboratoire GIPSA-Lab (anciennement ICP) : Boë (1999) ; Boë *et al.* (2002) ; Ménard (2002) ; Ménard *et al.* (2007) ; Serkhane *et al.* (2003, 2007) ; Schwartz *et al.* (2012b) ; Boë *et al.* (2013).

Finalement, pour schématiser, la vision que nous adoptons de *VLAM* dans cette thèse est de décrire le modèle par la relation entrées/sorties : il s'agit d'un modèle géométrique qui, à partir de commandes articulatoires (qui sont proches de commandes musculaires) produit des formants.

1.2.2 La transformation articulatori-acoustique et sa complexité

Une propriété fondamentale et bien connue des relations entre les variables articulatoires et les variables acoustiques est le caractère complexe et *many-to-one* (c'est-à-dire non injectif) de la transformation articulatori-acoustique. En effet, bien qu'une configuration articulatoire donnée, si elle est parfaitement spécifiée, permette de calculer précisément les caractéristiques acoustiques résultantes du son correspondant, un jeu de caractéristiques acoustiques donné ne suffit pas en général à spécifier de manière unique une configuration articulatoire. Pour être plus précis, si on se dote de suffisamment de caractéristiques spectrales, on peut vraisemblablement parvenir à une spécification assez précise de l'articulation ; mais si on se limite à quelques

can be compared to the productions of human speakers; the results of many such comparisons are described elsewhere (e.g., Callan *et al.* (2000) ; Guenther (1995) ; Guenther *et al.* (1998, 1999) ; Nieto-Castanon *et al.* (2005) ; Perkell *et al.* (2004a,b)). »

paramètres spectraux calculables, par exemple les deux, trois, voire même quatre premiers formants, on peut construire un large ensemble de configurations articulatoires fournissant exactement les mêmes valeurs de formants (voir Atal *et al.* (1978)). Néanmoins, ces configurations partagent souvent des caractéristiques géométriques similaires, et notamment des valeurs en général proches des paramètres de constrictions (aire et position de la constriction linguale, aire de la constriction labiale ; voir Boë *et al.* (1992)). Cette possible caractérisation géométrique a pourtant également ses limites, comme le montre bien le cas de la voyelle /u/, dont Boë *et al.* (2000) ont montré qu'elle pouvait être articulée en trois positions différentes du conduit vocal, dites *pharyngale*, *vélo-pharyngale* et *vélo-palatale*. Ainsi, la figure 5.4, montre deux configurations très différentes du conduit vocal conduisant à des valeurs formantiques quasiment identiques pour le /u/.

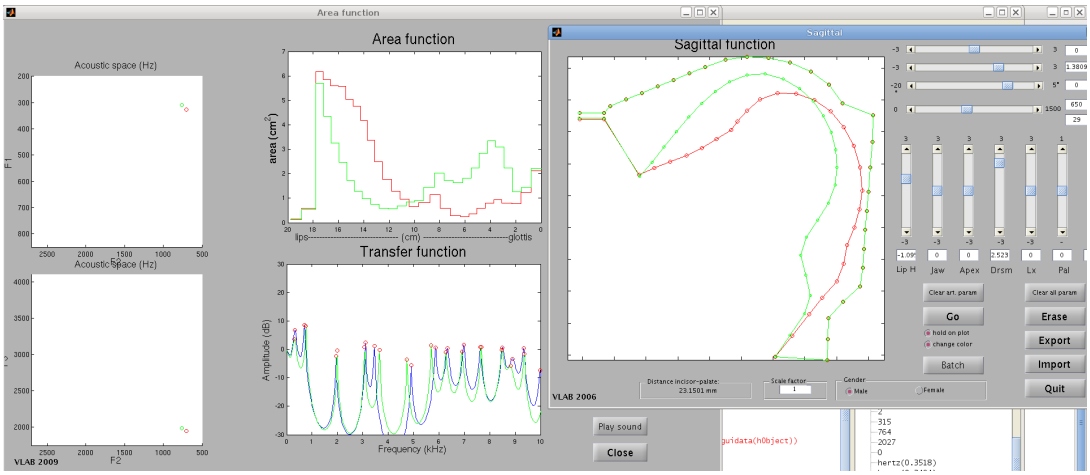


Figure 5.4: **Différentes manières de réaliser un /u/ :** deux coupes sagittales différentes, caractérisées par une constriction pharyngale pour l'une et vélo-palatale pour l'autre, conduisent à des fonctions d'aires très différentes, mais à des valeurs de formants quasiment identiques.

Avec le modèle *VLAM*, si l'on génère des configurations articulatoires avec sept paramètres et que l'on caractérise le son avec deux ou trois formants, le sous-dimensionnement de la sortie acoustique par rapport à l'entrée articulatoire est patent. Cela conduit à une indétermination articulatoire si l'on cherche à résoudre le problème de « l'inversion », c'est-à-dire si l'on cherche à récupérer le geste articulatoire à partir du son (voir une revue et des solutions proposées par exemple dans Ouni (2001) ; Demange et Ouni (2013)).

Dans la plupart des modèles existants, le problème de l'inversion est résolu par l'utilisation d'heuristiques reposant sur des principes de régularisation pour sélectionner un antécédent articulatoire particulier, par l'introduction de contraintes articulatoires telles que minimum de distance articulatoire à la configuration précédente, minimum d'écart au neutre, etc... En revanche, dans le cadre bayésien de notre modèle *COSMO*, le problème de l'inversion est résolu par le principe même de ce qu'est une distribution « inverse » dans un modèle bayésien. Ainsi, nous verrons que dans *COSMO* il nous suffira de construire par apprentissage une distribution de probabilité conjointe $P(MS)$, où M représente la configuration articulatoire et S la configuration acoustique associée, à partir de laquelle on obtient un « modèle direct » en calculant la distribution de probabilité conditionnelle $P(S \mid M)$, et un « modèle inverse » en calculant

$P(M \mid S)$.

1.3 Comment caractériser les syllabes de type plosive-voyelle

1.3.1 Variabilité des consonnes plosives et modèles de production de séquences plosive-voyelle

Une consonne plosive (ou occlusive) est produite par un geste de fermeture totale du conduit vocal, suivi d'un relâchement vers le son suivant (par exemple une voyelle). La fermeture, qui peut se produire en divers points du conduit, du plus arrière (fermeture dans le pharynx pour les plosives pharyngales) au plus avant (fermeture au niveau des lèvres pour les plosives bilabiales) produit une période de silence ou de voisement selon que les cordes vocales restent ou non en activité vibratoire pendant la fermeture (plosives voisées ou non voisées). Le relâchement, qui survient après une montée de pression en amont du lieu d'occlusion, fournit une brève période de bruit relativement intense que l'on nomme « explosion » (en anglais, *burst*).

Comme nous l'avons vu précédemment (dans la section 1 du chapitre 2), la production de séquences de phonèmes produit des phénomènes de coarticulation : chaque phonème est influencé, au niveau articulatoire, et au niveau acoustique, par la production des phonèmes voisins, antérieurs et postérieurs. Un exemple typique, abondamment commenté par les acteurs de la théorie motrice, est précisément celui de l'enchaînement entre une plosive et une voyelle : on observe, dans la figure 5.5, que le même phonème /d/ présente au niveau acoustique des transitions de formants radicalement différentes selon qu'il est suivi d'une voyelle /i/ (comme dans « dix ») ou /u/ (comme dans « doute »).

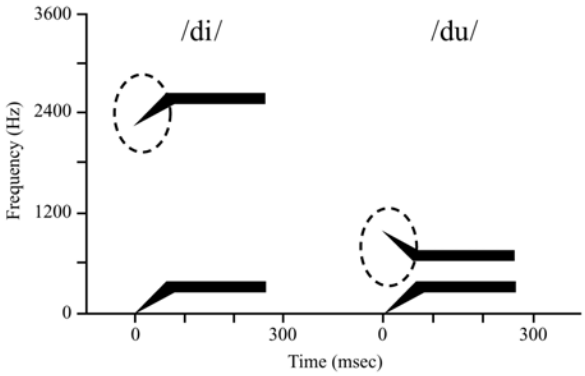


Figure 5.5: **Différentes caractéristiques acoustiques du /d/ en fonction du contexte vocalique :** le second formant monte de /d/ vers /i/, et au contraire descend de /d/ vers /u/, et ce avec des valeurs de départ elles-mêmes très différentes. (Cette figure est tirée de Galantucci *et al.* (2006) qui ont eux-mêmes adapté Liberman *et al.* (1967).)

L'interprétation de ce phénomène, du point de vue de la production de la parole, est simple. Comme le montre la figure 5.6, les consonnes plosives sont sous-spécifiées articulatoirement : ainsi, si la zone labiale est spécifiée pour les plosives /p b m/ avec une fermeture au niveau des lèvres, le reste du conduit vocal est libre : la langue n'est pas contrainte, et peut donc anticiper la configuration vocalique qui suit. Ainsi, la langue peut se positionner vers l'avant du conduit vocal pour le /bi/ ou vers l'arrière pour le /bu/. Bien évidemment, au moment du relâchement du *burst*, comme la langue n'est pas positionnée de la même manière pour /bi/ et pour /bu/, le

son n'est pas le même, ce qui produit la coarticulation. Il en est de même en français pour les dentales /t d n/, qui sont spécifiées au niveau de la pointe de la langue et libres ailleurs (dos de la langue et lèvres), et pour les vélaires /k g/ qui sont spécifiées au niveau du dos de la langue mais libres à l'avant et aux lèvres.

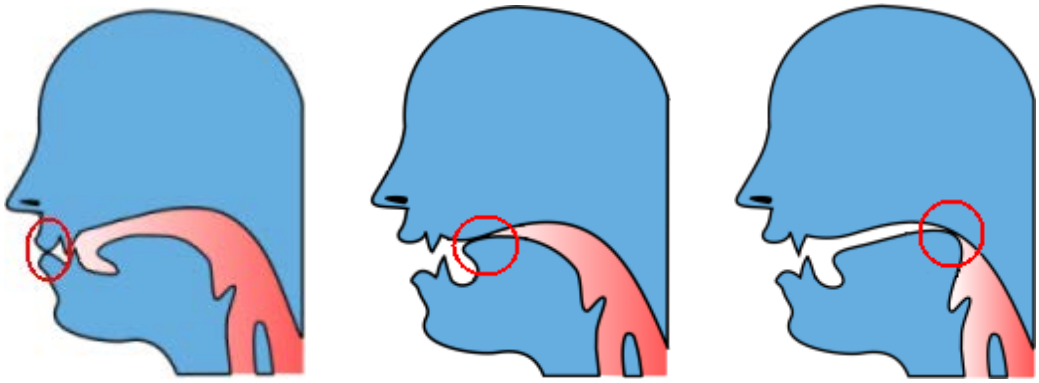


Figure 5.6: **Les plosives sont sous-spécifiées articulatoirement :** seul le lieu d'articulation (mis en évidence par une ellipse rouge) est contraint, les articulateurs restant sont libres. Les plosives /p b m/ sont uniquement caractérisées par une fermeture labiale, /t d n/ par une fermeture dentale, et /k g/ par une fermeture vélaire. (Ces figures ont été adaptées à partir de `http://www.phil-fak.uni-duesseldorf.de/ch-2-phonetics-phonology/`.)

Pour rendre compte de ce phénomène du point de vue de la production de la parole, un des premiers modèles introduits dans la littérature, et aussi l'un des plus simples, est celui d'Öhman (1966) qui propose de considérer la plosive comme une perturbation articulatoire locale appliquée sur la trajectoire vocalique : si une plosive P s'intercale entre deux voyelles $V1$ et $V2$ (pour former la séquence $V1PV2$), le contrôleur calculerait dans un premier temps la trajectoire $V1V2$, puis appliquerait en une zone donnée de cette trajectoire une perturbation sous la forme d'un geste labial (pour les bilabiales), de la pointe de la langue (pour les dentales) ou du dos de la langue (pour les vélaires). Dans le cas des séquences PV sur lesquelles nous allons nous focaliser dans ce chapitre, la programmation impliquerait de partir de la voyelle, de calculer la configuration plosive correspondante « rétrospectivement », par application de la perturbation, et de lancer alors la production de la trajectoire de P vers V.

Cette proposition, simple et séduisante, a été par la suite abondamment modifiée, améliorée et complexifiée (voir des revues bibliographiques dans Farnetani et Recasens (1997) ainsi que dans les thèses de Ma (2008) et Brunner (2008), et une discussion autour de plusieurs modèles dans le numéro spécial de *Behavioral and Brain Sciences* autour des travaux de Sussman *et al.* (1998), sur lesquels nous allons revenir). C'est cependant la proposition initiale de la théorie de la perturbation d'Öhman que nous retiendrons dans la suite de ce travail, à la fois par sa simplicité et parce que, nous le verrons, elle s'avère capable de produire des phénomènes de coarticulation plosive-voyelle globalement compatibles avec les données expérimentales.

1.3.2 Invariants potentiels associés au lieu d'articulation et modèles de perception de séquences plosive-voyelle

Puisque les formes acoustiques que peut prendre une plosive donnée sont variées et dépendent du contexte vocalique, existe-t-il un invariant calculable par le système de perception qui soit susceptible de fournir un corrélat unique à la plosive indépendamment de cette variation ? C'est une question qui a motivé bon nombre de travaux et débats dans les années 80. Là encore, nous ne ferons pas une revue exhaustive de cette littérature, mais en rappellerons quelques points d'ancrage.

Dans le cadre de sa « théorie quantique », que nous avons évoquée en divers points de ce manuscrit, Stevens propose que les lieux d'articulation sélectionnés majoritairement par les langues du monde, comme les articulations bilabiales, alvéo-dentales et palato-vélaires, soient associées à des discontinuités articulatori-acoustiques qui résultent en des invariants acoustiques (Stevens et Blumstein, 1978). Ces invariants sont calculés selon Stevens au niveau des 40 millisecondes qui suivent le *burst* acoustique de la consonne, et se traduisent par trois types de formes spectrales présentés figure 5.7 : « diffus descendant » pour les bilabiales, « diffus montant » pour les alvéo-dentales et « compact » pour les palato-vélaires.

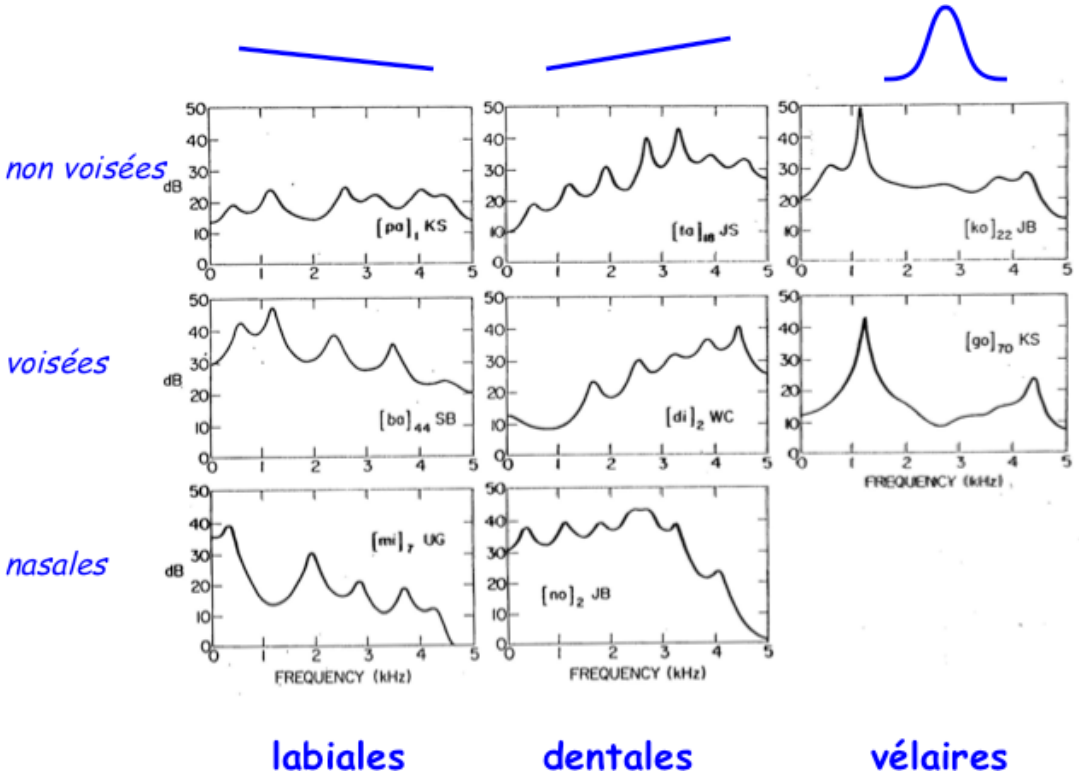


Figure 5.7: **Les plosives caractérisées par un invariant acoustique.** À chaque type de plosive semble correspondre un type de forme spectrale. (Cette figure est adaptée de Stevens (1980).)

À l'inverse, les tenants de la théorie motrice ont proposé, dans la logique de l'invariance articulatoire qui est à la base de leur réflexion, que l'invariant soit gestuel et non acoustique : puisque ce qui caractérise la plosive est articulatoire, alors que le son est variable, le système perceptif serait capable de récupérer le geste et son invariant (voir figure 5.8).

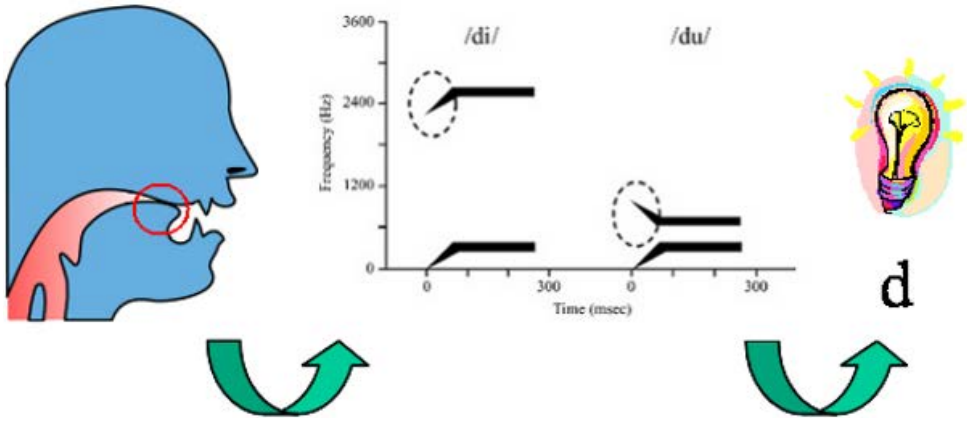


Figure 5.8: **Les plosives caractérisées par un invariant moteur.** Alors que le geste (à gauche) est fixe, le signal acoustique (au milieu) est variable, mais le percept (à droite) est constant. L'invariance est donc dans le geste vocal. (Cette figure est adaptée de Moulin-Frier *et al.* (2010).)

La théorie motrice se heurte néanmoins à des limites calculatoires puisqu'elle doit résoudre un problème « d'inversion articulatori-acoustique » dont nous avons dit dans la section précédente qu'il était compliqué. La théorie de l'invariance acoustique de Stevens a quant à elle montré ses limites dans les années 80 (voir par exemple Kewley-Port (1982, 1983)). Plus tard, dans les années 90, Sussman *et al.* (1991, 1998), prolongent des travaux célèbres des laboratoires Haskins dans les années 50 (Delattre *et al.*, 1955) en proposant non pas une invariance locale de la plosive dans la région de l'explosion acoustique, mais une invariance relationnelle faisant intervenir à la fois le spectre de la plosive (juste après le *burst*) et le spectre de la voyelle qui suit. Ainsi, en étudiant la relation entre le second formant de la voyelle et le second formant de la plosive au début de la trajectoire formantique vers la voyelle après le *burst*, ils observent une forte corrélation pour /b/ et pour /d/, et une corrélation par parties (en séparant voyelles antérieures et postérieures) pour /g/. La figure 5.9 montre ces relations linéaires qui sont d'après eux caractéristiques du lieu d'articulation de chaque plosive.

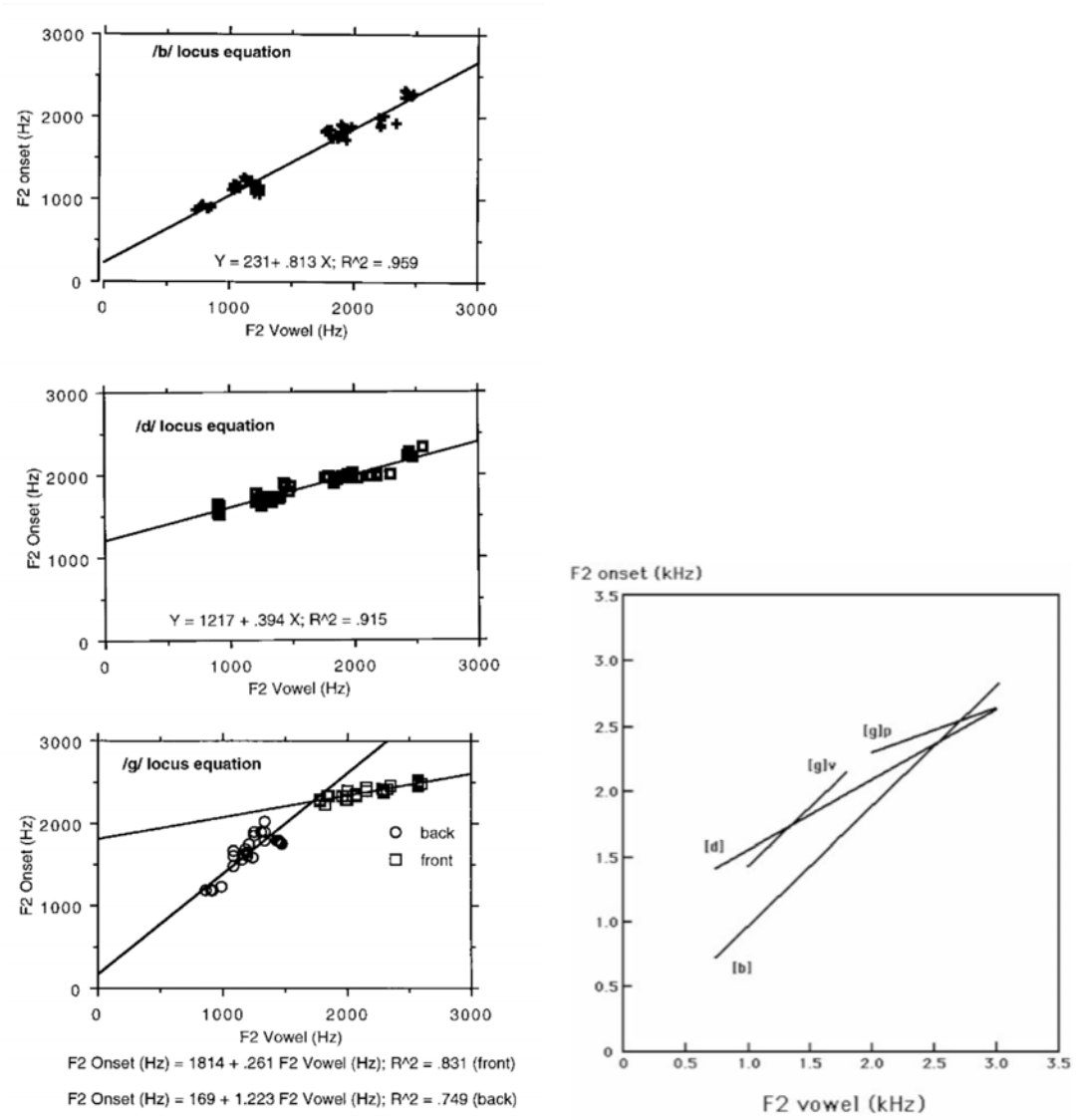


Figure 5.9: **Les plosives caractérisées par les équations du locus.** Il semble y avoir une relation affine entre le second formant de la plosive après le *burst* et le second formant de la voyelle. (D'après Sussman *et al.* (1998).)

1.3.3 Nos choix de modélisation

Nous avons donc décidé de nous appuyer sur le modèle de perturbation d'Öhman pour simuler des séquences plosives-voyelles. Ce modèle, simple, consistera pour nous à partir dans le modèle *VLAM* d'une configuration vocalique donnée, et à appliquer sur cette configuration articulatoire une perturbation permettant de fermer le conduit vocal aux lèvres pour un /b/, avec la pointe de la langue vers les dents pour /d/ et avec le corps de la langue contre le palais pour /g/. La configuration de *VLAM* correspondante sera alors considérée comme la plosive initiale, en « jouant » ainsi la séquence plosive-voyelle dans le sens inverse de celui qui a permis de la programmer.

Le modèle *VLAM* disposant de sept paramètres articulatoires (qui sont présentés à la figure 5.3) nous avons décidé de réduire ce nombre au minimum pour réduire la complexité calculatoire des simulations qui s'ensuivent. Il faut *a minima* trois paramètres sur *VLAM* pour générer de manière plausible l'ensemble des voyelles orales du français :

- *TongueBody* qui contrôle la position du corps de la langue ;
- *TongueDorsum* qui contrôle la position du dos de la langue ;
- *LipHeight* qui contrôle l'écartement entre les lèvres.

Pour produire des voyelles, tous les autres paramètres seront fixés à 0, ce qui correspond à une situation de repos.

Pour produire les plosives à partir de ces configurations, il faut ajouter la mandibule, qui permet d'élever globalement la langue et les lèvres, ainsi que le paramètre contrôlant la pointe de la langue pour réaliser le /d/, les paramètres de contrôle des lèvres pour le /b/ et de contrôle du dos de la langue pour le /g/ étant déjà disponibles. Ainsi, il nous faut *a minima* cinq paramètres sur *VLAM* pour générer de manière plausible les trois plosives /b d g/ :

- *Jaw* qui contrôle la hauteur de la mandibule ;
- *TongueBody* qui contrôle la position du corps de la langue ;
- *TongueDorsum* qui contrôle la position du dos de la langue ;
- *Apex* qui contrôle la position de la pointe de la langue ;
- *LipHeight* qui contrôle l'écartement entre les lèvres.

Pour produire des plosives, tous les autres paramètres seront fixés à 0, ce qui correspond à une situation de repos.

En termes auditifs, nous caractériserons les séquences plosive-voyelle par un couple de paramètres spectraux, correspondant respectivement à la plosive et à la voyelle, reprenant ainsi une version très simplifiée de l'architecture de la figure 5.2. Ces paramètres sont également réduits le plus possible : nous caractériserons les voyelles par leurs deux premiers formants $F1_V$ et $F2_V$, ce qui est un choix classique (voir par exemple Ladefoged et Johnstone (1982) ; Hayward (2000)). En ce qui concerne les plosives, nous les décrirons par leurs second et troisième formants $F2_P$ et $F3_P$, qui permettent selon Sussman *et al.* (1998) de caractériser complètement le lieu d'articulation indépendamment du contexte vocalique.

2 Synthèse de syllabes plosive-voyelle dans le cadre de *COSMO*

Nous supposons que chacune des catégories de voyelles produites par ce locuteur contient une certaine variabilité répartie autour d'une voyelle moyenne, prototypique. Nous simulons alors la production de voyelles par ce locuteur en tirant des gestes moteurs selon des distributions de probabilité gaussiennes centrées sur ces prototypes de voyelles. Des plosives sont alors réalisées à partir des voyelles ainsi obtenues avec un principe de coarticulation maximale : la voyelle est perturbée par un geste de constriction du conduit vocal qui lui est superposé.

2.1 Synthèse de voyelles : de la variabilité autour de gestes prototypiques

Nous présentons d'abord en détail la manière dont nous nous sommes dotés de prototypes moteurs de voyelles.

(1) Dans un premier temps, nous échantillonnons la transformation articulatori-acoustique réalisée par *VLAM* en parcourant l'espace des paramètres articulatoires avec un pas de discrétisation uniforme. Plus précisément, nous réalisons des voyelles en n'utilisant que les trois paramètres *TongueBody*, *TongueDorsum* et *LipHeight*, chacun étant discrétisé en 100 valeurs possibles. Pour chacune des configurations motrices ainsi obtenues, nous calculons la forme de conduit vocal correspondante. Si la surface d'une des 28 sections du conduit vocal ainsi obtenues est inférieure à un certain seuil ($0,15\ cm^2$) la configuration articulatoire est jugée trop fermée pour correspondre à une voyelle, et est donc rejetée. À partir de cette description géométrique du conduit vocal, *VLAM* calcule la fonction de transfert associée ainsi que les formants correspondants. Nous construisons ainsi un dictionnaire qui associe à 1 000 000 de gestes moteurs de voyelles uniformément répartis dans l'espace moteur les valeurs correspondantes des quatre premiers formants.[32]

(2) Dans un second temps, nous nous donnons des cibles acoustiques pour chacune des voyelles /a/, /i/ et /u/. Nous avons retenu comme valeurs prototypiques les valeurs moyennes de formants observées par Meunier (2007) chez des locuteurs français. Ainsi, la voyelle /a/ est caractérisée par $F1 = 684\ Hz$, $F2 = 1256\ Hz$ et $F3 = 2503\ Hz$; la voyelle /i/ par $F1 = 308\ Hz$, $F2 = 2064\ Hz$ et $F3 = 2976\ Hz$; et la voyelle /u/ par $F1 = 315\ Hz$, $F2 = 764\ Hz$ et $F3 = 2027\ Hz$.

(3) Dans un troisième temps, nous cherchons dans le dictionnaire construit en (1) des configurations articulatoires correspondant au mieux aux valeurs prototypiques des formants que nous avons retenues en (2) pour nos voyelles /a i u/. Nous utilisons pour cela deux critères : il faut d'une part que le geste moteur considéré soit suffisamment proche de la cible au sens d'une distance perceptive (une distance euclidienne sur les trois premiers formants exprimés en Bark), et d'autre part, lorsque plusieurs candidats sont quasiment équivalents du point de vue de cette distance perceptive, nous retenons celui qui est le plus proche (au sens d'une distance euclidienne sur les paramètres articulatoires) de la

[32] Même si dans ce qui est présenté dans cette thèse nous n'utilisons que les premier et second formants pour décrire les voyelles, les données que nous avons générées peuvent être utilisées dans d'autres contextes où les valeurs des quatre premiers formants seraient nécessaires.

position neutre. (*VLAM* est issu d'une analyse en composantes principales d'images de conduits vocaux, et ses paramètres articulatoires décrivent, composante par composante, la distance par rapport à la position moyenne observée dans les données. On considère que cette configuration moyenne du conduit vocal est une position de repos.) Il s'agit donc là d'une manière d'implémenter une certaine forme d'économie motrice.

C'est ainsi que nous nous sommes donné des prototypes moteurs de voyelles, que nous avons de plus validés auditivement en ajoutant en aval de *VLAM* un module permettant de synthétiser des voyelles à partir des valeurs de leurs formants. À partir de ces prototypes moteurs, nous tirons aléatoirement des réalisations de voyelles selon des distributions de probabilité gaussiennes centrées sur les gestes prototypiques. Là encore, si le critère d'ouverture du conduit vocal n'est pas satisfait (si l'aire au niveau d'une des sections est en dessous du seuil de $0,15$) la configuration articulatoire correspondante sera rejetée. Précisons également que, puisqu'il s'agit de simulations d'un phénomène physique, ce chapitre est le seul moment de la thèse où nos variables sont définies sur des intervalles continus ; partout ailleurs les variables correspondent à des représentations internes dans des modèles cognitifs et sont définies de manière discrétisée.

2.2 Synthèse de plosives : un geste de perturbation superposée à la voyelle

Nous présentons maintenant les principes que nous avons retenus pour générer des plosives à partir des voyelles précédemment décrites. Le processus de génération de plosives que nous avons mis en place repose sur les idées suivantes :

Un principe de coarticulation maximale. Nous adoptons la vision proposée par Öhman (1966) selon laquelle, dans le cas de syllabes de type plosive-voyelle, la plosive est une perturbation locale (un geste de fermeture du conduit vocal) qui vient se superposer à la configuration de la voyelle. Pour les syllabes /ba bi bu ga gi gu da di du/ que nous considérons, cela veut dire que la voyelle est anticipée au maximum au moment de l'ouverture du conduit vocal.

Un principe d'économie motrice dans le choix des articulateurs. Le geste de plosion est réalisé par l'action combinée de deux articulateurs : la mandibule (contrôlée par le paramètre *Jaw* de *VLAM*) et la fermeture des lèvres (contrôlée par le paramètre *LipHeight*) pour réaliser une labiale, la mandibule et le dos de la langue (paramètre *TongueDorsum*) pour réaliser une vélaire, et la mandibule et la pointe de la langue (paramètre *Apex*) pour réaliser une dentale.

De la variabilité liée à l'utilisation de la mandibule. Nous réalisons une plosive à partir d'une voyelle en venant fermer le conduit vocal. Cependant, pour que *VLAM* puisse calculer les formants associés à la plosive, on s'arrête juste avant que la fermeture ne soit complète : on impose que la surface de la section du conduit vocal au niveau de laquelle la constriction se fait ne descende pas en dessous de la valeur $0,05\ cm^2$. Nous avons ainsi implémenté un processus qui prend en entrée une configuration du conduit vocal ainsi qu'un articulateur (*LipHeight*, *TongueDorsum* ou *Apex*) et, par recherche dichotomique, calcule la nouvelle valeur de cet articulateur permettant d'obtenir une plosive. Il est à noter qu'en donnant en entrée la même configuration de voyelle et le même articulateur,

on obtient toujours la même plosive. Pour introduire davantage de variabilité, on choisit d'utiliser également la mandibule (paramètre *Jaw*). On génère alors les plosives en fermant le conduit vocal grâce à différentes combinaisons de l'action de la mandibule d'une part, et de l'autre articulateur considéré (*LipHeight*, *TongueDorsum* ou *Apex*) d'autre part. À partir d'une voyelle, on obtient donc ainsi tout un ensemble de plosives, caractérisées par différents degrés de fermeture de la mandibule (pour chacun de ces degrés de fermeture, on réutilise le mécanisme de recherche dichotomique précédemment décrit pour obtenir la valeur correspondante du second articulateur utilisé).

La modification du paramètre *Apex* de *VLAM*. Dans la version de *VLAM* dont nous disposions initialement, le paramètre *Apex* n'était pas assez corrélé avec la pointe de la langue, et trop avec le reste de la langue, ce qui fait que des variations de ce paramètre avaient pour effet de modifier un grand nombre de sections du conduit vocal.

Le manque de précision du modèle *VLAM* sur la position de la pointe de la langue est une limite du modèle qui a déjà été constatée par Gabioud (1994) et qui peut s'expliquer de la manière suivante. Les paramètres articulatoires de *VLAM* sont issus d'une Analyse en Composantes Principales qui a été guidée (Maeda, 1990) pour que les composantes qui en ressortent soient interprétables en terme de commandes musculaires, phonétiques et acoustiques. Un compromis a donc été réalisé entre deux enjeux : trouver un ensemble de composantes permettant d'exprimer la variabilité des données (les coupes cinéradiographiques de conduit vocal adulte) d'une part ; et assurer une forte corrélation entre ces composantes et des commandes musculaires d'autre part. Dans la version de *VLAM* dont nous disposions, le paramètre *Apex* servait plus le premier objectif que le second, et agissait à la fois sur la partie antérieure du conduit vocal (ce qui est normal) et sur des régions plus postérieures, ce qui n'a aucune vraisemblance anatomique.

Nous avons donc choisi de modifier la manière dont ce paramètre *Apex* modifie la forme du conduit vocal. Galván-Rodriguez (1997) a également modifié le paramètre d'apex du modèle de Maeda, en marquant à la main les coordonnées de la pointe de la langue dans les clichés de cinéradiographies, et l'analyse en composantes principales guidée de cette manière permet de faire émerger deux paramètres physiologiquement plausibles qui contrôlent la pointe de la langue. Nous avons été un peu plus radicaux dans nos choix, et nous avons modifié directement la manière dont le paramètre *Apex* intervient dans le calcul de la coupe sagittale pour qu'il n'influe plus que sur les premières sections du conduit vocal, de manière à produire une occlusion de type dentale. La figure 5.10 compare l'effet de l'activation du paramètre *Apex* sur la forme du conduit vocal dans l'ancienne version de *VLAM* et dans notre version modifiée.

La figure 5.10 illustre donc le choix que nous avons fait de retenir un /d/ plus proche d'une dentale (comme en français) que d'une alvéolaire (comme ce serait le cas en anglais).

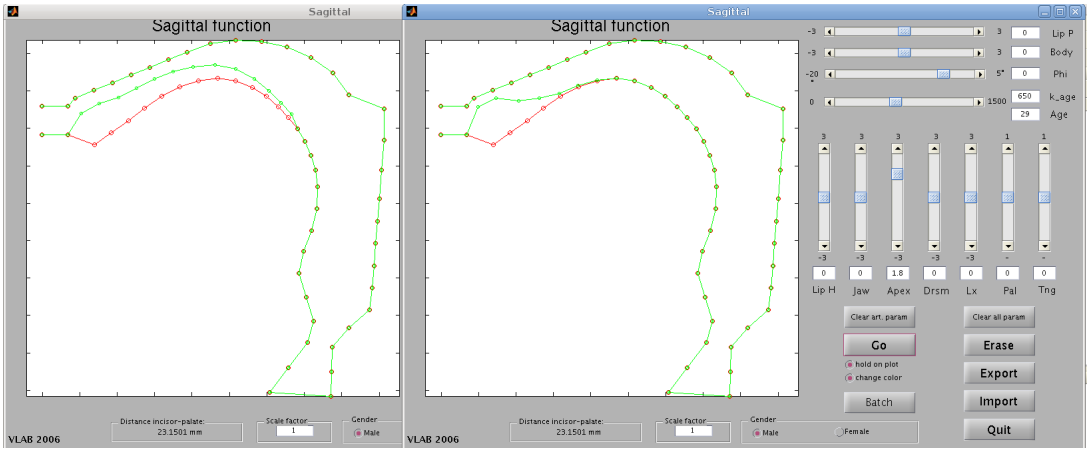


Figure 5.10: **Modification du paramètre *Apex* de *VLAM***. À gauche l'ancienne version de *VLAM*, à droite la version pour laquelle nous avons modifié le paramètre *Apex*. Dans les deux cas, on a en rouge la position moyenne (position neutre, dite de repos, où les valeurs des paramètres articulatoires sont toutes à 0) du conduit vocal, et en vert la forme de conduit vocal obtenue en actionnant l'*Apex*.

3 Évaluation des syllabes plosive-voyelle produites

3.1 Résultats de simulation

Dans un premier temps, les prototypes moteurs que nous nous sommes donnés conduisent aux productions vocaliques que la figure 5.11 illustre en les projetant dans les espaces acoustiques $F2 \times F1$ et $F2 \times F3$, et en montrant les formes de conduits vocaux correspondantes.

Sur cette base, notre modèle de génération de plosives par perturbation locale conduit aux données de la figure 5.12 qui présente les plosives ainsi obtenues dans l'espace acoustique[33].

L'ensemble de ces données apparaît compatible avec les résultats de simulation de Schwartz *et al.* (2012b) qui sont présentés figure 5.13.

On observe ainsi que les espaces acoustiques correspondant respectivement dans nos simulations à /a i u/ d'une part et à /ba da ga bi di gi bu du gu/ d'autre part s'inscrivent correctement dans les espaces maximaux générés. Plus précisément, on observe sur la figure 5.13 que les trajectoires engendrées par les variations contextuelles pour chaque syllabe sur la figure 5.12 correspondent parfaitement aux trajectoires globales (tous contextes vocaliques confondus) pour chaque lieu d'articulation obtenues par Schwartz *et al.* (2012b).

Globalement, toutes ces données ayant été obtenues sur des versions très proches de *VLAM* (Schwartz *et al.* (2012b) ont réalisé une légère modification de la forme du palais pour permettre une plosion vélaire ; ils utilisent de plus les sept articulateurs de *VLAM* qu'ils font varier de -3 à $+3$ écarts-types, alors que nous n'en utilisons que trois pour les voyelles et cinq pour les consonnes, en autorisant une plage de variations légèrement plus large), la correspondance n'est pas surprenante. Mais le fait que nous explorions à peu près l'ensemble des trajectoires globales dans nos simulations montre que, malgré la réduction de complexité fournie par nos hypothèses de production simplistes (notamment en réduisant le nombre de paramètres variés), nous avons

[33] La valeur du premier formant étant sensiblement la même (300 Hz environ) pour les plosives /b/, /d/ et /g/, nous ne montrons les plosives générées que dans l'espace $F2 \times F3$.

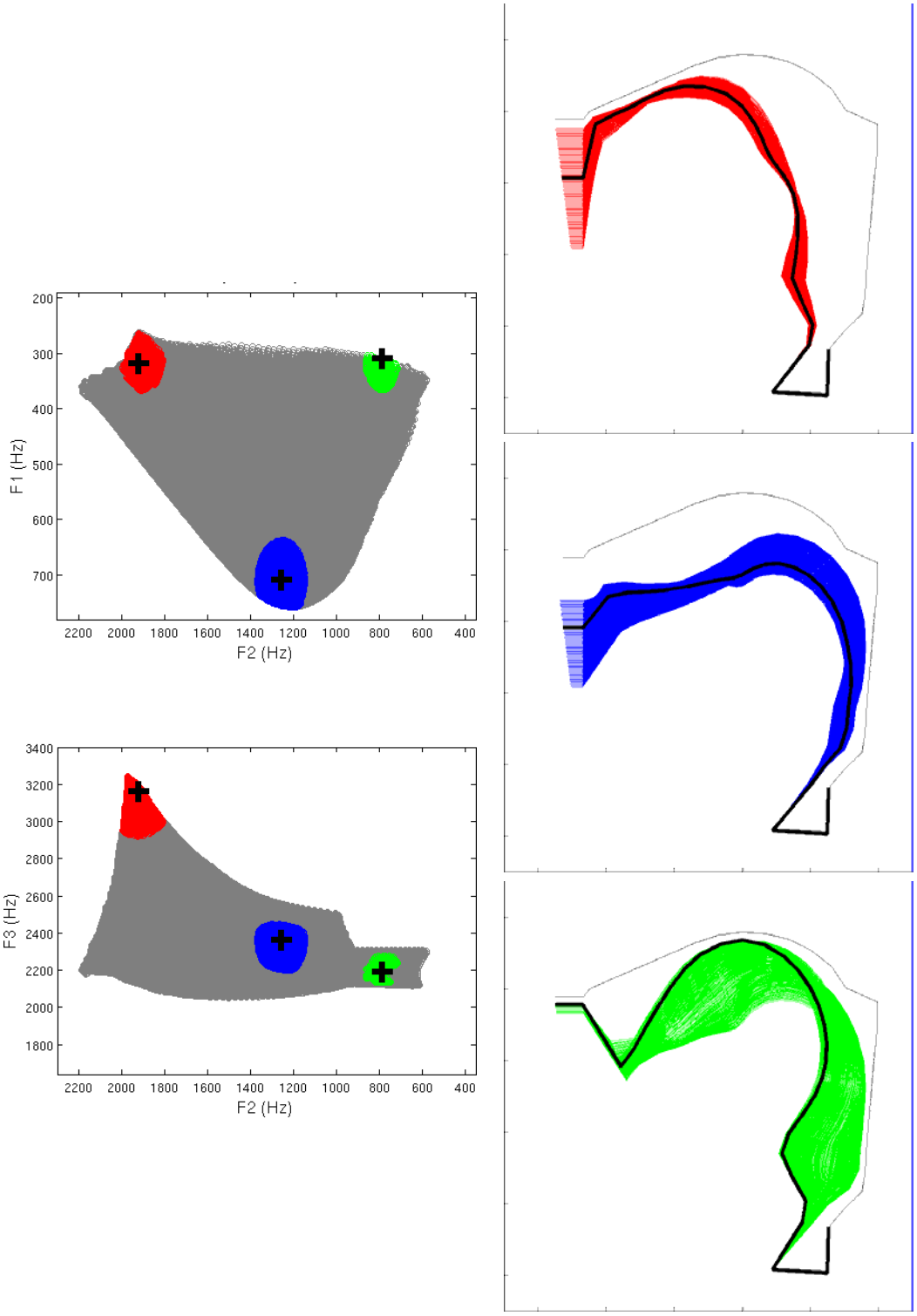


Figure 5.11: **Nos voyelles synthétiques**, décrites dans l'espace acoustique $F2 \times F1$ (en haut à gauche), dans l'espace $F2 \times F3$ (en bas à gauche), et par les coupes sagittales des conduits vocaux correspondants (dans la colonne de droite). La couleur rouge correspond aux différentes réalisations de la voyelle /i/, le bleu à la voyelle /a/, le vert au /u/, et la couleur noire correspond aux valeurs prototypiques choisies de la manière décrite à la section 2.1. La couleur gris représente l'espace acoustique maximal (pour le « locuteur » considéré, c'est-à-dire notre version de *VLAM*), qui est obtenu en projetant dans l'espace acoustique le dictionnaire de voyelles que nous avons construit en 1 par exploration exhaustive.

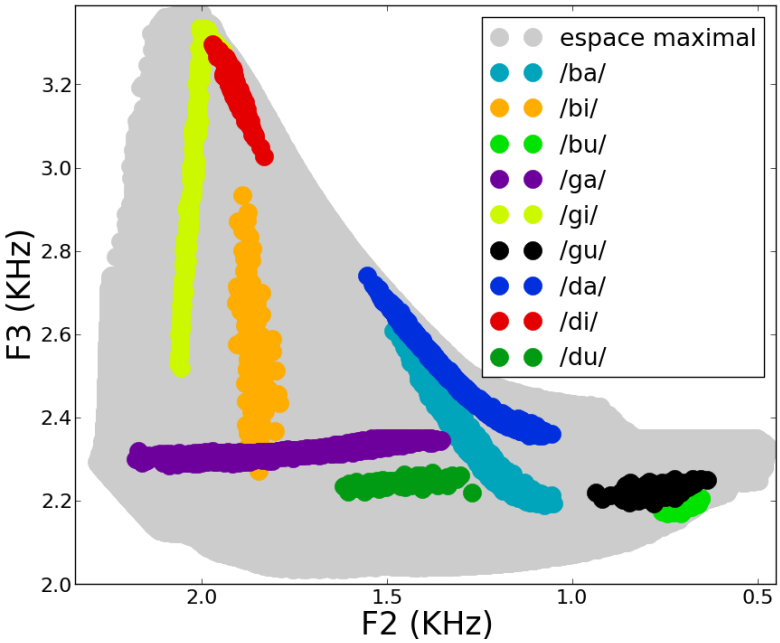


Figure 5.12: **Nos plosives synthétiques dans l'espace acoustique**. En gris, l'espace acoustique maximal correspondant aux plosives pouvant être produites avec notre version de *VLAM*.

bien généré des syllabes très variées, ce qui permettra de tester dans le chapitre suivant des matériaux raisonnablement complexes même s'ils ne sont pas naturels. Reste à savoir si ces portraits de variation acoustique sont conformes aux données réelles : c'est ce que nous allons voir maintenant.

De plus, les syllabes plosive-voyelle que nous avons générées ont été validées auditivement grâce aux outils de synthèse développés par Berthommier *et al.* (2012).

3.2 Comparaisons avec des données réelles : étude des lois du locus

Il existe de très nombreuses données acoustiques sur les syllabes consonne-voyelle dans la littérature. Nous n'avons pas entrepris une comparaison systématique, hors du contexte de ce travail. Mais, pour obtenir un point d'évaluation nous permettant de valider nos simulations, nous avons décidé de les comparer aux données de locus de Sussman *et al.* (1998). Pour ce faire, nous prenons les plosives présentées figure 5.12, synthétisées à partir des voyelles présentés figure 5.11, dont nous résumons la variabilité liée à l'action de la mandibule en calculant la valeur moyenne du second formant pour chaque lieu d'articulation. Nous reportons ces données sur la figure 5.14, sur laquelle nous avons fait figurer également les droites de régression correspondant pour ces données à chaque lieu d'articulation (/b/, /d/ ou /g/).

Alors que la partie droite de la figure 5.14 montre une régression sur des données (Sussman *et al.*, 1991) de plosives produites par vingt locuteurs (dix hommes et dix femmes) dans dix contextes vocaliques différents, les données de plosives que nous avons générées (voir figure 5.12) correspondent à un seul « locuteur » (un seul conduit vocal artificiel : notre version de *VLAM*) et à trois contextes vocaliques seulement. Les régressions que nous faisons sont donc basées sur peu

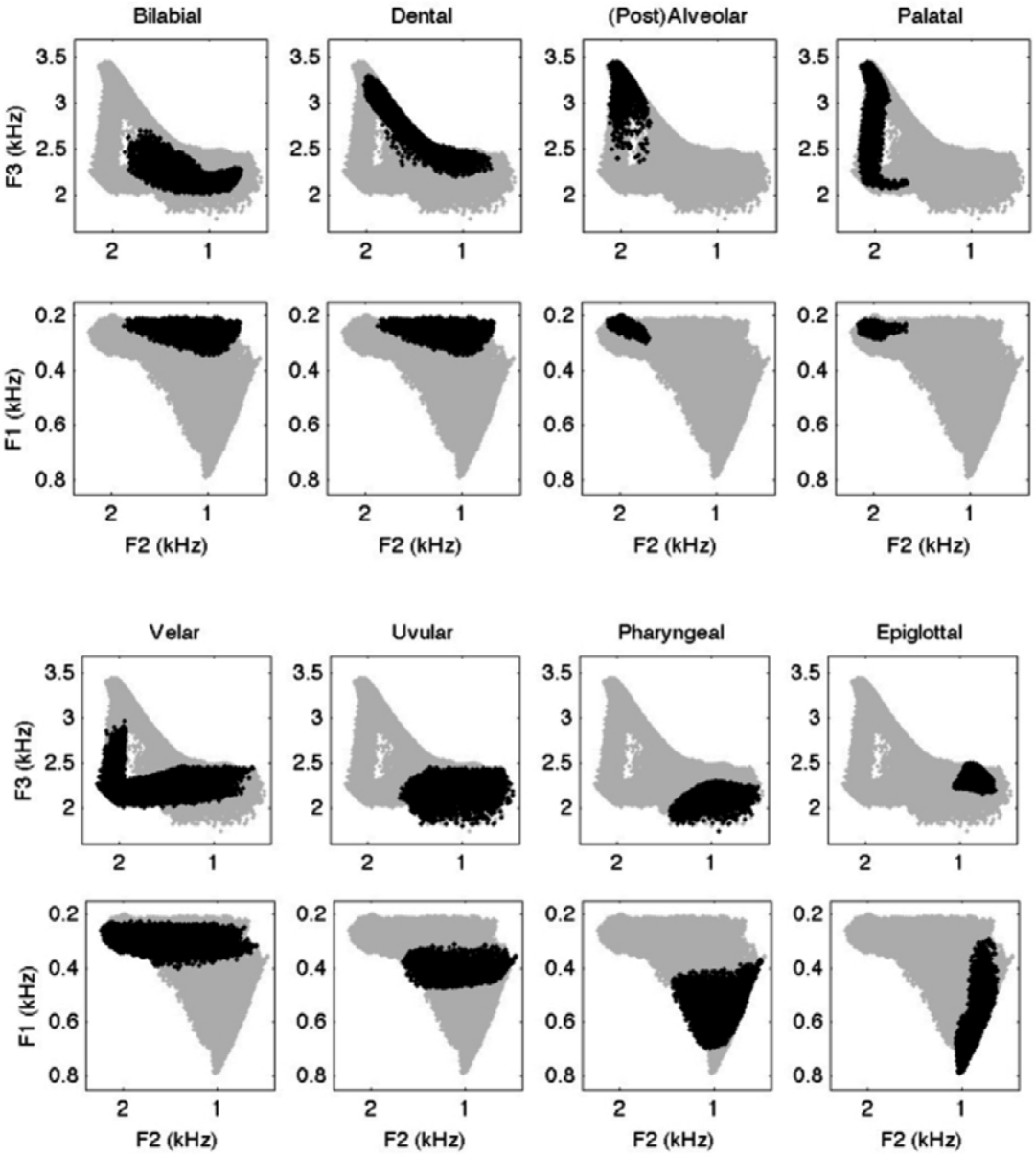


Figure 5.13: **Comparaison avec les plosives synthétisées par Schwartz *et al.* (2012b)**. Les valeurs (en noir) des trois premiers formants juste après le relâchement du *burst* des plosives, sont superposées à l'espace maximal des plosives (en gris). Cette figure se lit en huit blocs verticaux de deux figures. Chaque bloc, qui correspond à un lieu d'articulation différent (/b/ bilabial, /d/ dental, /d/ alvéolaire ou post-alvéolaire, /ɟ/ palatal, /g/ vélaire, /ɢ/ uvulaire, /ʢ/ pharyngal ou /ʔ/ épiglottal), comporte deux figures montrant respectivement les valeurs des formants dans l'espace $F2 \times F3$ pour celle du haut et dans l'espace $F2 \times F1$ pour celle du bas.

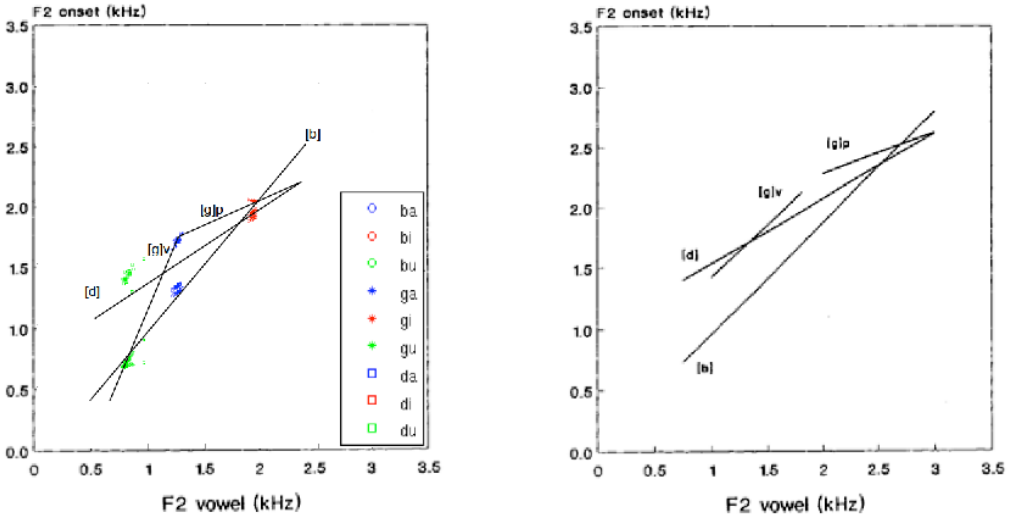


Figure 5.14: **Comparaison aux équations du locus :** à gauche un résumé de nos données synthétiques est comparé avec les régressions obtenues par Sussman *et al.* (1998) sur des données réelles.

de points (et la régression pour les palatales /g/ est basée sur un petit nombre de réalisations de la seule voyelle /i/). Néanmoins il apparaît sur la figure 5.14 que les configurations d'équations du locus que nous obtenons ne sont pas aberrantes par rapport aux données de terrain reportées par Sussman *et al.* (1998).

On peut noter quelques éléments de discordance. La valeur de $F2$ pour la plosive dans /da/ est basse, probablement parce que nous avons choisi pour /d/ une articulation dentale plutôt qu'alvéolaire. Cela se traduit, comme on le voit sur la figure 5.13, par des valeurs moins élevées de $F2$. Les valeurs de $F2$ pour la plosive dans /du/ et dans /gu/ sont également plus faibles que dans les données de locus de Sussman *et al.* (1998). Ceci situe probablement les limites du modèle de perturbation que nous avons mis en place, dans lequel on peut faire un /du/ ou un /gu/ sans avancer le corps de la langue, par un simple geste de fermeture – alors qu'il y a aussi un geste d'avancée de la langue à prendre en compte pour produire la fermeture en situation réelle. C'est particulièrement le cas pour /du/, qui peut être fait, avec notre *Apex* simplifié, sans aucune avancée, ce qui est totalement impossible étant données les contraintes morphologiques de la langue (on ne peut pas placer la pointe de la langue contre les dents si le corps de la langue est trop en arrière).

Malgré ces limites, qui s'expliquent par nos choix simplificateurs, et qui sont d'importance assez marginale par rapport à nos objectifs, l'ensemble des données présentées dans ce chapitre montre que nos simulations permettent de générer de la variabilité contextuelle sur les plosives, selon la voyelle qui suit, et ceci d'une manière qui n'est pas qualitativement très éloignée des données réelles telles qu'elles sont résumées par les équations du locus de Sussman *et al.* (1998).

4 Conclusion

Dans ce chapitre, nous avons réalisé une étude bibliographique dans le but d'en extraire les principes de modélisation nécessaires pour pouvoir instancier le modèle *COSMO* générique présenté au chapitre 3 dans un cadre moins abstrait que celui du chapitre 4. Dans un premier temps, nous nous sommes intéressés à la caractérisation dans des espaces perceptifs et moteurs du signal de parole en général, puis à celle des plosives en particulier. Nous avons notamment décrit le modèle de perturbation d'Öhman (1966), qui propose de considérer la plosive comme une perturbation articulatoire locale appliquée sur la trajectoire vocalique, et que nous avons utilisé pour générer un ensemble de données de syllabes de type plosive-voyelle à partir de *VLAM*, un modèle géométrique de conduit vocal, pour les utiliser au chapitre suivant où le modèle *COSMO* sera étendu au traitement des syllabes.

Les syllabes que nous avons synthétisées sont réalistes au sens où elles montrent des patrons de variabilité qui correspondent aux données réelles. L'enjeu principal de cette thèse est le traitement de ces données et de cette variabilité qui est présenté au chapitre suivant. L'algorithme d'apprentissage que nous avons déjà décrit au chapitre 4 et qui va être appliqué à ces données acoustiques de syllabes synthétiques ne fait aucune hypothèse sur la manière dont les données ont été générées. Nous aurions donc pu, sans rien changer à l'intérêt de notre approche, utiliser dans cette thèse d'autres principes plus complexes de génération de syllabes, ou un autre modèle de synthèse articulatoire que *VLAM* (par exemple, le modèle développé par Busset (2013) semble être basé sur des paramètres articulatoires un peu plus réalistes que ceux de *VLAM*, notamment en ce qui concerne la pointe de la langue), voire même utiliser un modèle biomécanique comme celui de Winkler *et al.* (2010) pour la génération de séquences articulatoires et acoustiques.

Chapitre 6

Apprentissages sensori-moteurs réalistes au sein de *COSMO* et application à des tâches de perception de syllabes

1 *COSMO-S* : le modèle *COSMO* étendu au traitement des syllabes **111**

2 Implémentation d'algorithmes d'apprentissage dans le cadre de *COSMO-S* . . **119**

3 Principaux résultats **138**

4 Conclusion **147**

COSMO est un modèle générique permettant d'étudier les interactions perceptuo-motrices en parole. Nous avons présenté son élaboration détaillée au chapitre 3, et nous en avons proposé une instanciation au chapitre 4 dans un cadre théorique monodimensionnel, ce qui a permis de présenter un algorithme original d'apprentissage par accomodation et d'en illustrer les dynamiques ainsi que les propriétés des modèles appris. Il s'agit dans le présent chapitre d'étendre le modèle *COSMO* présenté au chapitre 4 au traitement des syllabes, grâce aux données qui ont été générées de la manière décrite au chapitre 5, afin de pouvoir étudier avec nos outils computationnels la manière dont COSMO peut gérer la complexité de véritables objets phonétiques, et nous aider à comprendre comment ils sont appris, représentés et traités.

1 *COSMO-S* : le modèle *COSMO* étendu au traitement des syllabes

Dans cette section nous montrons comment la structure de notre modèle bayésien d'agent cognitif *COSMO* s'adapte pour que l'agent cognitif qu'il décrit ait des représentations internes auditives et motrices associées à des objets de type syllabe. Dans le reste de cette thèse, nous baptisons le modèle ainsi construit *COSMO-S*, pour *COSMO-Syllabes*.

1.1 Principes, hypothèses de modélisation et enjeux de simulations

Le modèle *COSMO-S* est construit autour des idées suivantes.

- On conserve le principe au cœur du modèle *COSMO*, à savoir la combinaison dans les représentations internes d'un agent d'un système moteur, d'un système sensori-moteur et d'un système auditif. Le modèle *COSMO-S* procure également un cadre computationnel intégrateur dans lequel on peut comparer les déclinaisons motrice, auditive et perceptuo-motrice de la production et de la perception des syllabes.

- Au chapitre 5 nous avons montré comment la syllabe plosive-voyelle pouvait être modélisée par deux états : un état ouvert correspondant à la voyelle, et un état fermé correspondant à la consonne. Les variables motrice M et perceptive S du modèle générique sont donc dédoublées en M_V et M_C qui dénotent les gestes moteurs de la voyelle et de la consonne d'une part, et en S_V et S_C qui dénotent les représentations auditives associées à la voyelle et à la consonne d'autre part.

- Le modèle de coarticulation d'Öhman (1966) décrit les plosives comme étant en partie déterminées par les voyelles auxquelles elles sont associées au sein d'une syllabe. Cette dépendance sera exprimée dans le modèle par des distributions de probabilité conditionnelles.

- L'agent communicant π_{Ag} est doté d'un vocabulaire d'actions qui comporte trois types de gestes de constriction, c'est-à-dire trois manières de perturber la voyelle en fermant le conduit vocal pour produire une plosive de type bilabial (fermeture des lèvres), dental (fermeture de la pointe de la langue derrière les dents) ou vélaire (fermeture du dos de la langue contre le palais mou, aussi appelé *velum*).

- Du point de vue technique, nous faisons le choix de représenter les connaissances décrites par le modèle dans trois sous-systèmes distincts et indépendants (les systèmes moteur, perceptif, et sensori-moteur), que nous relions ensuite par des variables de cohérence[34]. Il s'agit d'un outil mathématique qui permet d'assurer durant l'inférence la cohérence entre certaines variables du modèle en imposant un lien mathématique entre ces variables. Par exemple, dans les chapitres 3 et 4, la variable C est une variable de cohérence, et conditionner une distribution de probabilité par le terme C=1 a pour effet d'imposer l'égalité des variables O_S et O_L. Dans le modèle *COSMO* comme dans *COSMO-S*, cette variable de cohérence C permet de définir mathématiquement les instanciations des tâches de production et de perception de la parole dans le cadre des théories perceptuo-motrices. Dans le modèle *COSMO-S*, nous ajoutons des variables de cohérence supplémentaires λ qui permettent d'assembler rigoureusement les connaissances exprimées dans les différents sous-systèmes au sein d'un unique modèle intégrateur. Ces variables λ jouent un rôle d'interrupteur probabiliste, et permettent d'activer ou non les connaissances stockées dans les différents systèmes (voir Gilet *et al.* (2011) pour une utilisation similaire des variables de cohérence).

[34] Voir par exemple Bessière *et al.* (2013) pour une définition, et différentes utilisations dans Colas *et al.* (2010) et Gilet *et al.* (2011).

Sur cette base, les questions que nous poserons à nos simulations sont les suivantes :

- Comment gérer, dans un modèle qui commence à présenter un certain niveau de complexité, l'apprentissage de compétences motrices selon notre processus d'interaction dans lequel l'agent apprend à partir des données sensorielles fournies par le maître ?
- Quel place ce processus d'apprentissage laisse-t-il au développement d'idiosyncrasies motrices, et quelles en sont les conséquences pour le traitement perceptif ?
- Comment nos différents modèles de perception peuvent-ils rendre compte de l'émergence de la phonologie, et en particulier de l'émergence de la notion de « lieu d'articulation des plosives » en dépit de la variabilité due au contexte vocalique ?
- Quelles sont les différences entre les différents modèles en termes de robustesse au bruit dans les tâches de perception ?

1.2 Variables du modèle et domaines de définition

Le modèle *COSMO* est un modèle générique qui peut être instancié de différentes manières. Alors que la démonstration du théorème d'indistinguabilité faite au chapitre 3 est indépendante de la définition précise des variables, dans le chapitre 4 nous avons instancié le modèle *COSMO* pour étudier un cas théorique simple, dans lequel les variables motrices et perceptives sont monodimensionnelles. Dans le présent chapitre nous en proposons une instanciation plus réaliste, pour pouvoir manipuler des syllabes. Le modèle *COSMO-S* comporte maintenant les variables suivantes.

Les objets O_S et O_L sont des variables discrètes pouvant prendre neuf valeurs possibles : les syllabes plosive-voyelle /ba/, /bi/, /bu/, /da/, /di/, /du/, /ga/, /gi/, /gu/.

La consigne motrice, M_V ou M'_V selon que l'on se place dans le système sensori-moteur ou dans le système moteur, est une variable dont les trois dimensions correspondent aux trois paramètres $TongueBody$, $TongueDorsum$ et $LipHeight$ de *VLAM* qui permettent de décrire les configurations motrices des voyelles /a/, /i/ et /u/. Chacune de ces variables est discrétisée de manière uniforme en 25 valeurs possibles[35].

La consigne motrice M_C, qui apparaît uniquement dans le système sensori-moteur, est une variable dont les cinq dimensions correspondent aux cinq paramètres Jaw, $TongueBody$, $TongueDorsum$, $TongueApex$ et $LipHeight$ de *VLAM* qui permettent de décrire les configurations des trois consonnes plosives /b/, /d/, et /g/ produites dans les trois contextes vocaliques /a/, /i/ et /u/. Chacune de ces variables est également discrétisée de manière uniforme en 25 valeurs possibles.

Le geste de constriction G'_C est une variable discrète dont les trois valeurs possibles (*bilabiale*, *dentale* ou *vélaire*) constituent le vocabulaire d'actions dont dispose l'agent. Cette variable dénote le choix du lieu d'articulation de la consonne plosive.

[35] La définition précise des intervalles de définition de chaque variable est donnée au chapitre 5, et correspond aux bornes de l'espace couvert dans les dictionnaires produits avec *VLAM*.

La perturbation Δ'_{MC} est une variable qui décrit la manière dont le geste de constriction choisi vient perturber la configuration du conduit vocal correspondant à la voyelle. Δ'_{MC} est donc une variable dont les cinq dimensions Jaw, $TongueBody$, $TongueDorsum$, $TongueApex$ et $LipHeight$ sont les mêmes que celles de la variable M_C.

La représentation sensorielle, S_V ou S'_V selon que l'on se place dans le système sensori-moteur ou dans le système auditif, est une variable dont les deux dimensions sont les deux premiers formants de la voyelle : $F1_V$ et $F2_V$. Ces deux variables $F1_V$ et $F2_V$ sont discrétisées selon une échelle perceptive dont le pas de discrétisation est de $0,3$ $Bark$. Il y a 16 valeurs possibles pour $F1_V$ et 33 valeurs possibles pour $F2_V$.

La représentation sensorielle, S_C ou S'_C selon que l'on se place dans le système sensori-moteur ou dans le système auditif, est une variable dont les deux dimensions sont les second et troisième formants de la consonne : $F2_C$ et $F3_C$. Ces deux variables $F2_C$ et $F3_C$ sont discrétisées selon une échelle perceptive dont le pas de discrétisation est de $0,3$ $Bark$. Il y a 33 valeurs possibles pour $F2_C$ et 14 valeurs possibles pour $F3_C$.

Le succès de la communication est décrit par la variable booléenne C qui a donc deux valeurs possibles : 1 et 0.

Les variables de cohérence λ_{SV}, λ_{SC}, λ_{MV} et λ_{MC} sont des variables booléennes, ayant donc deux valeurs possibles (1 et 0), qui permettent d'assurer à toute étape de l'inférence probabiliste la correspondance entre les variables qu'elles relient.

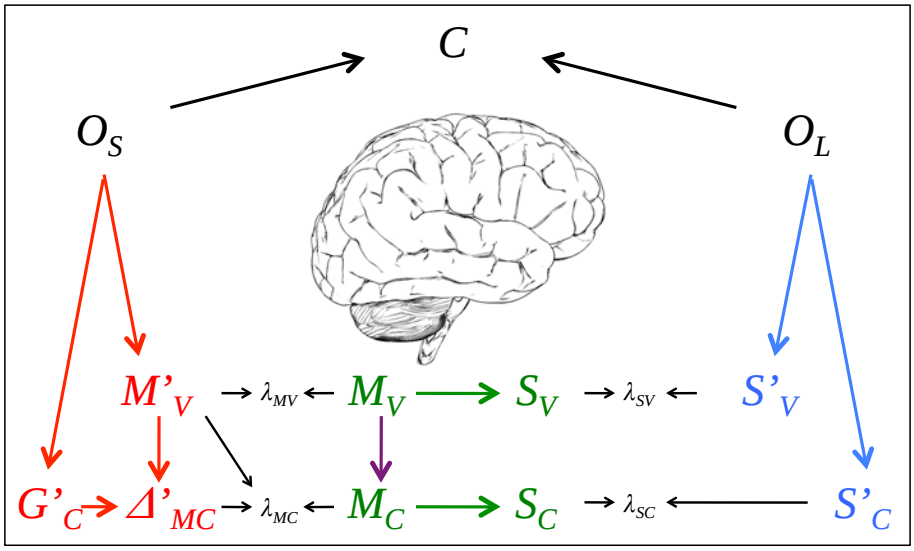
1.3 Distribution de probabilité conjointe du modèle *COSMO-S*

L'équation 3.13 du chapitre 3 montre comment on peut calculer toute question probabiliste de la forme $P(Cherchées \mid Connues)$ à partir de la distribution de probabilité conjointe d'un modèle. Dans le cas du modèle *COSMO-S*, cette distribution de probabilité conjointe est définie sur un espace dont la taille est le produit des tailles des ensembles de définition de chacune des variables, c'est-à-dire un nombre de l'ordre de 10^{32} ! De manière similaire à ce qui est fait dans la section 2.2 du chapitre 3, nous faisons maintenant une série d'hypothèses d'indépendances conditionnelles afin de réduire la complexité du modèle et des calculs d'inférence, ce qui a aussi pour effet de rendre le modèle plus facile à interpréter.

La distribution de probabilité conjointe du modèle se décompose alors de la manière présentée sur la figure 6.1 : d'une part nous illustrons par un modèle graphique[36] les relations de dépendances conditionnelles entre les variables, et d'autre part nous définissons par une équation la décomposition de la distribution de probabilité conjointe du modèle.

Cette décomposition est similaire à celle du modèle *COSMO* présentée au chapitre 3 dans la mesure où elle regroupe au sein d'un modèle intégrateur un système moteur, un système perceptif et un système sensori-moteur. Ayant fait ces hypothèses d'indépendances conditionnelles et ce choix de décomposition, l'espace mémoire nécessaire pour stocker l'ensemble des termes intervenant dans la décomposition de la distribution de probabilité conjointe est alors de l'ordre de 10^{12}, ce qui est sans commune mesure avec 10^{32}.

[36] En fait il ne s'agit pas tout à fait d'un réseau bayésien puisque le terme $P(S'_V\ S'_C \mid O_L\ \pi_{Ag})$ traduit les deux flèches partant de O_L par une seule distribution de probabilité.



$$
\begin{aligned}
&P(O_S\ G'_C\ M'_V\ \Delta'_{MC}\ \lambda_{MV}\ \lambda_{MC}\ M_V\ M_C\ S_V\ S_C\ \lambda_{SV}\ \lambda_{SC}\ S'_V\ S'_C\ O_L\ C \mid \pi_{Ag})\\
&\quad = P(O_S \mid \pi_{Ag}) \times P(M'_V \mid O_S\ \pi_{Ag}) \times P(G'_C \mid O_S\ \pi_{Ag}) \times P(\Delta'_{MC} \mid M'_V\ G'_C\ \pi_{Ag}) \times\\
&\qquad P(\lambda_{MV} \mid M'_V\ M_V\ \pi_{Ag}) \times P(\lambda_{MC} \mid M'_V\ \Delta'_{MC}\ M_C\ \pi_{Ag}) \times\\
&\qquad P(M_V \mid \pi_{Ag}) \times P(S_V \mid M_V\ \pi_{Ag}) \times P(M_C \mid M_V\ \pi_{Ag}) \times P(S_C \mid M_C\ \pi_{Ag}) \times\\
&\qquad P(\lambda_{SV} \mid S_V\ S'_V\ \pi_{Ag}) \times P(\lambda_{SC} \mid S_C\ S'_C\ \pi_{Ag}) \times\\
&\qquad P(O_L \mid \pi_{Ag}) \times P(S'_V\ S'_C \mid O_L\ \pi_{Ag}) \times\\
&\qquad P(C \mid O_S\ O_L\ \pi_{Ag})
\end{aligned}
$$

Figure 6.1: **Le modèle *COSMO-S* de traitement des syllabes**, décrit par un modèle graphique (en haut), et par sa distribution de probabilité conjointe (en bas).

1.4 Formes paramétriques

Nous spécifions maintenant les choix des formes paramétriques utilisées pour représenter les distributions de probabilités qui encodent les connaissances stockées dans les différents sous-systèmes de notre modèle *COSMO-S* (voir figure 6.1).

Le système moteur décrit les connaissances qu'a l'agent sur les gestes moteurs qu'il est habitué à associer aux objets. Le système moteur de l'agent fait intervenir dans sa décomposition les éléments suivants : un prior sur les objets, un répertoire moteur de gestes associés aux voyelles, le choix d'un geste de constriction dans un vocabulaire d'actions, et un modèle de perturbation décrivant la manière dont le geste de fermeture choisi vient modifier la configuration voyelle.

Le prior sur les objets $P(O_S \mid \pi_{Ag})$ encode la connaissance a priori qu'a l'agent sur les objets. $P(O_S \mid \pi_{Ag})$ prend la forme d'une distribution de probabilité uniforme pour traduire le fait que les différents objets ont la même fréquence d'apparition

dans l'environnement (ou l'ignorance de cette fréquence d'apparition). On a donc $P\left([O_S{=}o] \mid \pi_{Ag}\right) = 1/9$ pour chaque objet o (parmi /ba-bi-bu-da-di-du-ga-gi-gu/).

Le répertoire moteur de voyelles $P(M'_V \mid O_S\ \pi_{Ag})$ décrit la connaissance qu'a l'agent sur les gestes moteurs associés à la partie voyelle des objets syllabes. On choisit d'encoder ce terme par des distributions de probabilité gaussiennes de la forme $P\left(M'_V \mid [O_S{=}o]\ \pi_{Ag}\right) = Gauss\left(\mu_o^{M_V}, \sigma_o^{M_V}\right)$. Puisque O_S a neuf valeurs possibles, il y a donc neuf gaussiennes différentes, qui sont paramétrées par neuf vecteurs moyennes $\mu_o^{M_V}$ à trois dimensions et neuf matrices de covariance $\sigma_o^{M_V}$ de taille 3×3.

Le répertoire de gestes de constriction $P(G'_C \mid O_S\ \pi_{Ag})$ décrit la connaissance qu'a l'agent sur le choix des articulateurs à actionner pour fermer le conduit vocal en fonction de la syllabe à réaliser. Puisque le vocabulaire d'actions est limité à un choix parmi trois gestes de fermeture (bilabiale, dentale, ou vélaire), on choisit d'encoder le terme $P(G'_C \mid O_S\ \pi_{Ag})$ sous forme d'histogramme. Chacun des neuf histogrammes est décrit par un vecteur de trois paramètres que l'on note g'_o.

Il est donc important de noter ici que l'hypothèse que nous faisons de l'existence de trois gestes de base pour la production des plosives n'implique pas en elle-même la notion de lieu d'articulation comme donnée phonologique. En effet, rien ne signale au départ que les syllabes /di/, /du/ et /da/ partagent le même geste, et c'est au contraire un histogramme complet sur les trois gestes possibles qui est associé au départ à chaque syllabe. Tout l'enjeu des simulations sera de déterminer si cette notion de lieu d'articulation et d'invariance associée émerge de l'apprentissage dans les différents modèles testés.

Le modèle de perturbation $P(\Delta'_{MC} \mid M'_V\ G'_C\ \pi_{Ag})$ décrit la connaissance qu'a l'agent sur la manière dont il faut solliciter les articulateurs associés au geste de fermeture G'_C pour fermer le conduit vocal à partir de la voyelle M'_V. La fermeture du conduit vocal se fait en combinant le paramètre Jaw de $VLAM$ avec un autre paramètre : $LipHeight$ pour produire une bilabiale, $TongueApex$ pour produire une dentale, et $TongueDorsum$ pour produire une vélaire. Nous choisissons d'encoder le terme $P(\Delta'_{MC} \mid M'_V\ G'_C\ \pi_{Ag})$ sous la forme d'une distribution de probabilité gaussienne sur l'espace produit des deux paramètres associés à G'_C pour obtenir une constriction à partir de la voyelle M'_V. Chacune de ces distributions de probabilité gaussiennes (il y en a trois fois le nombre de valeurs possibles de M'_V) est paramétrée par un vecteur moyenne à deux dimensions μ^{Δ}_{g,m'_v} et une matrice de covariance σ^{Δ}_{g,m'_v} de taille 2×2.

Le système sensori-moteur décrit la connaissance qu'a l'agent sur la relation entre les consignes motrices (gestes articulatoires) et leurs conséquences perceptives. Le système sensori-moteur de l'agent fait intervenir les éléments suivants dans sa décomposition : un prior sur les gestes moteurs de voyelles, un modèle interne de la transformation articulatori-acoustique pour les voyelles, un modèle interne pour les consonnes, et un terme encodant des contraintes biomécaniques de coarticulation.

Le prior sur les gestes de voyelles $P(M_V \mid \pi_{Ag})$ encode la connaissance a priori qu'a l'agent sur les gestes moteurs de voyelles qui peuvent être produits. $P(M_V \mid \pi_{Ag})$

prend la forme d'une distribution de probabilité uniforme pour traduire l'ignorance de la fréquence de réalisation de ces gestes.

Le modèle interne pour les voyelles $P(S_V \mid M_V\ \pi_{Ag})$ encode un modèle interne de la transformation articulatori-acoustique pour les voyelles. On choisit d'utiliser des distributions de probabilité gaussiennes : pour chacune des valeurs possibles de la variable M_V, les connaissances de l'agent sur les conséquences perceptives de cette consigne articulatoire m_v sont encodées sous forme d'une distribution de probabilité gaussienne, c'est-à-dire $P\left(S_V \mid [M_V{=}m_v]\ \pi_{Ag}\right) = Gauss(\mu^S_{m_v}, \sigma^S_{m_v})$. Il y a donc autant de gaussiennes que de valeurs possibles pour M_V, qui sont paramétrées par des vecteurs moyennes $\mu^S_{m_v}$ à deux dimensions, et des matrices de covariance $\sigma^S_{m_v}$ de taille 2×2.

Les contraintes de coarticulation $P(M_C \mid M_V\ \pi_{Ag})$ sont codées « en dur » dans le modèle. Ce terme encode un principe d'économie motrice dans le choix d'un articulateur : il décrit la manière dont l'agent choisit de produire ses consonnes M_C à partir des voyelles M_V, c'est-à-dire en appliquant un geste de constriction labiale, dentale ou vélaire (par l'action combinée de la mandibule avec les lèvres, la pointe de la langue ou le dos de la langue). Le terme $P(M_C \mid M_V\ \pi_{Ag})$ est représenté par un ensemble de distributions de probabilité (une pour chaque valeur m_v de M_V) uniformes par zones et nulles ailleurs : $P\left([M_C{=}m_c] \mid [M_V{=}m_v]\ \pi_{Ag}\right)$ vaut $1/k$ lorsque la configuration consonne m_c correspond à l'une des k consonnes pouvant s'obtenir à partir de la voyelle m_v en réalisant l'un de ces trois types de geste de constriction, et 0 sinon. Ce terme $P(M_C \mid M_V\ \pi_{Ag})$ impose donc un lien fort entre voyelle et consonne qui structurera l'exploration de l'espace moteur lors de l'apprentissage.

Le modèle interne pour les consonnes $P(S_C \mid M_C\ \pi_{Ag})$ encode un modèle interne de la transformation articulatori-acoustique pour les consonnes. On choisit d'utiliser des distributions de probabilité gaussiennes : pour chacune des valeurs possibles de la variable M_C, les connaissances de l'agent sur les conséquences perceptives de cette consigne articulatoire m_c sont encodées sous forme d'une distribution de probabilité gaussienne, c'est-à-dire $P\left(S_C \mid [M_C{=}m_c]\ \pi_{Ag}\right) = Gauss(\mu^S_{m_c}, \sigma^S_{m_c})$. Il y a donc autant de gaussiennes que de valeurs possibles pour M_C, qui sont paramétrées par des vecteurs moyennes $\mu^S_{m_c}$ à deux dimensions, et des matrices de covariance $\sigma^S_{m_c}$ de taille 2×2.

Le système auditif décrit la connaissance qu'a l'agent sur la relation entre les entrées perceptives et les objets. Le système auditif fait intervenir les éléments suivants dans sa décomposition : un prior sur les objets, et des prototypes auditifs associant les percepts aux objets.

Le prior sur les objets $P(O_L \mid \pi_{Ag})$ encode la connaissance a priori qu'a l'agent sur les objets. $P(O_L \mid \pi_{Ag})$ prend la forme d'une distribution de probabilité uniforme pour traduire le fait que les différents objets ont la même fréquence d'apparition dans l'environnement (ou l'ignorance de cette fréquence d'apparition). On a donc $P\left([O_L{=}o] \mid \pi_{Ag}\right) = 1/9$ pour chaque objet o (parmi /ba-bi-bu-ga-gi-gu-da-di-du/).

Les prototypes auditifs $P(S'_V\ S'_C \mid O_L\ \pi_{Ag})$ décrivent la connaissance qu'a l'agent sur les percepts associés aux objets syllabes. On choisit d'encoder ce terme par des distributions de probabilité gaussiennes de la forme $P(S'_V\ S'_C \mid O_L\ \pi_{Ag}) = Gauss(\mu_o^S, \sigma_o^S)$. Puisque O_L a neuf valeurs possibles, il y a donc neuf gaussiennes différentes, qui sont paramétrées par neuf vecteurs moyennes μ_o^S à quatre dimensions (les deux premiers formants $F1_V$ $F2_V$ de la voyelle et les second et troisième formants $F2_C$ et $F3_C$ de la consonne) et neuf matrices de covariance σ_o^S de taille 4×4.

Le système de validation de la communication $P(C \mid O_S\ O_L\ \pi_{Ag})$ décrit la connaissance qu'a l'agent sur le succès de la communication. C est une variable booléenne qui vaut 1 avec une probabilité 1 lorsque les objets considérés du point de vue du locuteur et de l'auditeur sont les mêmes ($O_S = O_L$), et qui vaut 0 sinon. Mathématiquement, cela se traduit par l'utilisation d'un Dirac fontionnel, ce qui peut s'écrire $P\left([C{=}1] \mid O_S\ O_L\ \pi_{Ag}\right) = \delta_{O_S=O_L}$.

L'intégration des sous-systèmes moteur, auditif et sensori-moteur au sein d'un même modèle π_{Ag} se fait grâce à l'utilisation de variables de cohérence, qui permettent d'assurer à toute étape de l'inférence probabiliste la correspondance entre les variables qu'elles relient. On a ainsi quatre termes à définir.

Le terme $P(\lambda_{SV} \mid S_V\ S'_V\ \pi_{Ag})$ assure la cohérence entre la représentation du percept voyelle dans le système sensori-moteur et dans le système auditif. On utilise un Dirac fonctionnel, ce qui peut s'écrire $P\left([\lambda_{SV}{=}1] \mid S_V\ S'_V\ \pi_{Ag}\right) = \delta_{S_V=S'_V}$.

Le terme $P(\lambda_{SC} \mid S_C\ S'_C\ \pi_{Ag})$ assure la cohérence entre la représentation du percept consonne dans le système sensori-moteur et dans le système auditif. On utilise un Dirac fonctionnel, ce qui peut s'écrire $P\left([\lambda_{SC}{=}1] \mid S_C\ S'_C\ \pi_{Ag}\right) = \delta_{S_C=S'_C}$.

Le terme $P(\lambda_{MV} \mid M_V\ M'_V\ \pi_{Ag})$ assure la cohérence entre la représentation du geste moteur de voyelle dans le système sensori-moteur et dans le système moteur. On utilise un Dirac fonctionnel, ce qui peut s'écrire $P\left([\lambda_{MV}{=}1] \mid M_V\ M'_V\ \pi_{Ag}\right) = \delta_{M_V=M'_V}$.

Le terme $P(\lambda_{MC} \mid M'_V\ \Delta'_{MC}\ M_C\ \pi_{Ag})$ assure la cohérence entre la représentation du geste moteur de consonne dans le système sensori-moteur et dans le système moteur. Plus précisément, ce terme assure que l'on obtient la consonne M_C en superposant la perturbation Δ'_{MC} à la voyelle M'_V. On utilise un Dirac fonctionnel, ce qui peut s'écrire $P\left([\lambda_{MC}{=}1] \mid M'_V\ \Delta'_{MC}\ M_C\ \pi_{Ag}\right) = \delta_{M_C=M'_V!+\Delta'_{MC}}$.

Les connaissances encodées dans notre modèle dépendent de la valeur de plusieurs paramètres : $\mu_o^{M_V}$ et $\sigma_o^{M_V}$ qui contrôlent les prototypes moteurs des voyelles correspondant à chaque objet o, les vecteurs g'_o qui contrôlent le choix des gestes de constriction, $\mu_{m_v}^S$ et $\sigma_{m_v}^S$, et $\mu_{m_c}^S$ et $\sigma_{m_c}^S$ qui contrôlent la représentation dans le modèle interne de l'agent des conséquences perceptives des gestes moteurs de voyelle m_v et de consonne m_c, et μ_o^S et σ_o^S qui contrôlent les prototypes acoustiques correspondant à chaque objet. Ces paramètres peuvent prendre des valeurs différentes dans chaque instance de notre modèle, et leur valeur peut évoluer au cours de phases d'apprentissage lors desquelles l'agent va mettre à jour ses connaissances à partir des observations qu'il fait sur son environnement.

2 Implémentation d'algorithmes d'apprentissage dans le cadre de *COSMO-S*

Cette section décrit comment les algorithmes d'apprentissage présentés au chapitre 4 sont adaptés pour traiter les syllabes au sein du modèle *COSMO-S*.

2.1 Choix d'un cycle développemental

Bien qu'il soit probable que chez l'humain les différents apprentissages (perception et production de phonèmes, lexique, syntaxe, sémantique, ...), même s'ils ne convergent pas à la même vitesse, se fassent de manière plus ou moins concomittante, nous choisissons dans le cadre de cette thèse de procéder en séquence, pour des raisons computationnelles, mais aussi dans un certain accord avec les schémas développementaux proposés dans la littérature, comme nous allons le voir.

Le modèle *COSMO-S*, qui est résumé par la figure 6.1, a été construit en combinant trois sous-systèmes : le système auditif, le système sensori-moteur et le système moteur. Il s'agit donc par l'apprentissage d'acquérir des compétences perceptives, des compétences motrices, et de construire un modèle interne de la transformation articulatori-acoustique. Dans quel ordre ?

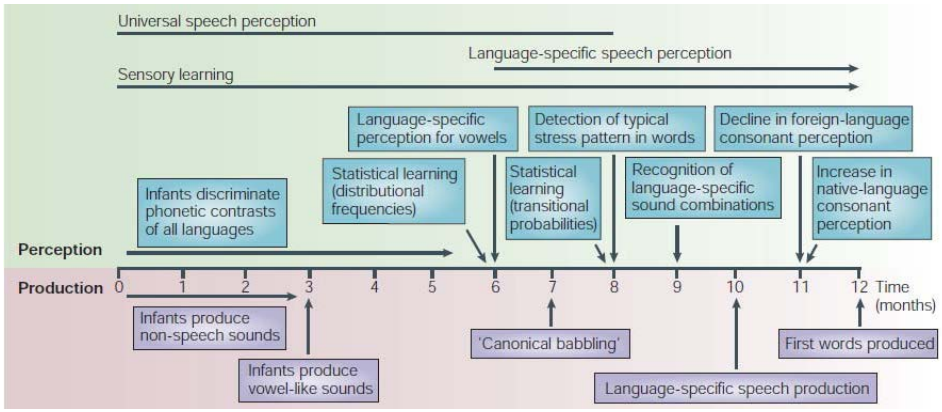


Figure 6.2: **Chronologie du développement** de la perception et de la production, d'après Kuhl (2004).

Comme le montre la figure 6.2, Kuhl (2004) propose une chronologie du développement de la perception et de la production. Nous en retiendrons principalement trois étapes :

1. Si les nouveaux nés disposent déjà de capacités contrastives, les mécanismes de catégorisation des voyelles caractéristiques de la langue semblent être en place au bout de six mois.

2. La phase dite de « babillage canonique » intervient aux alentours de sept mois. Il s'agit de séquences d'ouverture-fermeture de la mâchoire étudiées en phonétique développementale, comme par exemple dans le cadre de la théorie Frame/Content (MacNeilage, 1998). Cette phase du développement est souvent interprétée par les modélisateurs comme une étape d'exploration (éventuellement aléatoire) permettant la construction d'un modèle interne du comportement du conduit vocal.

③ La production de sons de parole spécifiques de la langue commence à se faire à partir de dix mois, de manière concomitante à une baisse des capacités de perception des consonnes des langues étrangères, et une amélioration de la perception des consonnes de la langue native.

À partir de cette série d'observations, nous choisissons de réaliser l'apprentissage du modèle *COSMO-S* en trois étapes, qui consistent à apprendre chacun des sous-systèmes auditif, sensori-moteur et moteur selon la séquence développementale suivante.

① La première étape correspond à la phase d'apprentissage supervisé du système auditif, qui est décrite à la section 2.3. L'existence de capacités contrastives chez le nouveau né traduit l'existence d'une certaine notion de distance perceptive, que nous implémentons sous forme de distributions de probabilité gaussiennes à partir desquelles l'agent apprend son classifieur auditif.

② La seconde étape correspond à la phase d'apprentissage du système sensori-moteur par accomodation, qui est décrite à la section 2.4. Parmi les stratégies d'exploration possibles pour construire un modèle interne du conduit vocal, nous en avons choisi une selon laquelle les productions du bébé sont guidées par la seule volonté de reproduire les sons de parole présents dans son environnement.

③ La troisième étape correspond à la phase d'apprentissage du système moteur par accomodation, qui est décrite à la section 2.5. Une fois que l'agent dispose d'un modèle interne lui permettant de retrouver les gestes moteurs associés aux signaux acoustiques, il peut se construire des répertoires moteurs de gestes pour les objets correspondants.

Par ailleurs, les algorithmes d'apprentissage présentés dans ce chapitre étant basés sur les mêmes principes que ceux présentés au chapitre 4, ils permettent pour les mêmes raisons de s'éloigner des conditions du théorème d'indistinguabilité présentées au chapitre 3.

2.2 Données d'apprentissage

Les algorithmes d'apprentissage présentés ici sont basés sur le même principe d'interactions entre un agent maître et un agent apprenant que ce qui a été décrit à la section 3 du chapitre 4. L'apprentissage fait intervenir d'une part le modèle du maître, qui a des répertoires moteurs de gestes de voyelles et produit des consonnes selon le modèle de coarticulation de Öhman (1966) présenté au chapitre 5 ; et d'autre part le modèle de l'environnement décrivant le passage des gestes articulatoires réalisés par le maître aux valeurs de formants perçues par l'agent, qui sont données par *VLAM*, le modèle géométrique de conduit vocal.

Les apprentissages que nous allons décrire ensuite sont réalisés en se basant sur des distributions de probabilité qui résument les propriétés du modèle du maître et du modèle de l'environnement. En effet, à partir des données des syllabes plosive-voyelle générées de la manière décrite au chapitre 5, nous construisons neuf (une pour chaque valeur /ba-bi-bu-ga-gi-gu-da-di-du/ de $O^{Maître}$) distributions de probabilité de type histogramme de la forme $P\left(S_V\ S_C \mid O^{Maître}\right)$. Les interactions entre l'agent apprenant et l'agent maître au sein de leur environnement, lors desquelles le maître communique à l'agent apprenant un objet o à la fois par un mécanisme d'attention partagée et par la transmission des caractéristiques acoustiques

$(f1_v, f2_v, f2_c, f3_c)$ de la syllabe correspondante, sont alors simulées en choisissant des objets o selon une distribution de probabilité uniforme, et en tirant aléatoirement les valeurs des formants selon la distribution de probabilité $P\left(S_V\ S_C \mid \left[O^{Maître}{=}o\right]\right)$. Ainsi, la variabilité des données de syllabes qui ont été synthétisées au chapitre 5 est capturée par des distributions de probabilité de type histogramme, selon lesquelles des tirages aléatoires permettent d'obtenir autant d'items que nécessaire aux besoins de l'apprentissage, ce qui revient à rééchantillonner.

Finalement, les trois phases de l'apprentissage de notre modèle *COSMO-S* (apprentissage du classifieur auditif, apprentissage du modèle interne de la transformation articulatori-acoustique, et apprentissage des répertoires de gestes moteurs) sont réalisées en utilisant les mêmes données : les syllabes produites par le maître et transmises par l'environnement, qui sont obtenues en réalisant des tirages aléatoires selon les distributions de probabilité $P\left(S_V\ S_C \mid O^{Maître}\right)$ que nous venons de mentionner.

2.3 Apprentissage du système auditif par association

Cette section décrit la manière dont l'agent apprend par association le lien entre les objets syllabes O_L et les valeurs de formants $F1_V$, $F2_V$, $F2_C$ et $F3_C$ correspondantes. À partir des productions du maître, l'agent se construit des prototypes auditifs gaussiens de la forme $P(S'_V\ S'_C \mid O_L\ \pi_{Ag}) = Gauss(\mu_o^S, \sigma_o^S)$, qui encodent la connaissance que l'agent se construit progressivement sur les zones de l'espace acoustique associées aux syllabes. Ces prototypes de syllabes sont encodés dans le modèle sous forme de distributions de probabilité gaussiennes à quatre dimensions ($F1_V$, $F2_V$, $F2_C$ et $F3_C$) qui contiennent donc de l'information sur la voyelle et la consonne de la syllabe, sans que cette distinction soit rendue explicite.

Le principe de cette phase d'apprentissage supervisé est le même que ce qui a été présenté à la section 3.2 du chapitre 4 : le maître choisit des objets o puis produit les syllabes (s_v, s_c) correspondantes en tirant aléatoirement selon la distribution de probabilité $P\left(S_V\ S_C \mid [O{=}o]\ \pi_{App}\right)$, et transmet également à l'agent la nature des objets o par un mécanisme d'attention partagée. Pour chacune de ces syllabes $< (s_v, s_c),\ o >$ transmises par le maître, l'agent met à jour ses connaissances avec l'observation selon laquelle l'objet o peut correspondre au signal perceptif (s_v, s_c). Plus précisément, pour chacune de ces syllabes, l'agent utilise la valeur (s_v, s_c) pour mettre à jour les valeurs μ_o^S et σ_o^S de la moyenne et de la matrice 4×4 de covariance de la gaussienne sur l'espace perceptif à quatre dimensions ($F1_V$, $F2_V$, $F2_C$ et $F3_C$) qui est associée à l'objet o.

Les valeurs initiales des paramètres μ_o^S et σ_o^S sont choisies pour que ces gaussiennes se comportent comme des distributions de probabilité uniformes : on met la moyenne μ_o^S au centre de l'espace perceptif, et la matrice initiale σ_o^S est diagonale avec de très grandes valeurs comparativement à la taille de l'espace perceptif des $S_V \times S_C$. Ensuite, au cours de l'apprentissage, l'agent affine progressivement sa représentation du lien entre objets et percepts et, en mettant à jour leurs moyennes et leurs matrices de covariance, l'agent ajuste ses prototypes gaussiens $P(S'_V\ S'_C \mid O_L\ \pi_{Ag})$ pour qu'ils correspondent aux productions du maître.

Ce sont ces prototypes acoustiques de syllabes appris de manière supervisée que l'agent utilise ensuite (avec une inversion probabiliste) pour réaliser des tâches de perception de syllabes en n'utilisant que des compétences perceptives (comme le prévoient les théories purement auditives de la perception).

2.4 Apprentissage du système sensori-moteur par accomodation

Le cœur de cette section décrit un algorithme d'apprentissage non supervisé permettant à l'agent de se construire progressivement un modèle interne de la transformation articulatori-acoustique en imitant les productions syllabiques du maître. Nous commençons par définir formellement la tâche d'imitation de syllabes, puis nous décrivons l'algorithme d'apprentissage, dont nous illustrons enfin la convergence.

2.4.1 Inférence probabiliste pour la tâche d'imitation de syllabes

Dans le cas de notre modèle *COSMO-S*, la tâche d'imitation se formalise de la manière suivante : étant donné un stimulus de syllabe (s_v, s_c) fourni par le maître, l'agent π_{Ag} doit choisir une consigne motrice (m_v, m_c) dans le but d'atteindre la cible (s_v, s_c). Dans cette version de l'algorithme d'imitation, l'agent est placé dans un bain audio de stimuli fournis par le maître, qu'il va essayer d'imiter, mais sans qu'intervienne aucune notion de catégorie symbolique : l'agent perçoit un son qu'il va essayer de reproduire, sans chercher à savoir s'il s'agit d'un /ba/, d'un /di/ ou d'un /gu/, etc... Dans ce contexte, réaliser une tâche d'imitation, cela revient à tirer aléatoirement un geste moteur (m_v, m_c) selon la distribution de probabilité $P\,(M_V\ M_C \mid [S_V{=}s_v]\ [S_C{=}s_c]\ \pi_{Ag})$. Nous montrons maintenant par inférence probabiliste comment se calcule cette distribution de probabilité qui encode la tâche d'imitation. D'après le théorème de Bayes, on a :

$$P\,(M_V\ M_C \mid [S_V{=}s_v]\ [S_C{=}s_c]\ \pi_{Ag}) = \frac{P\,(M_V\ M_C\ [S_V{=}s_v]\ [S_C{=}s_c] \mid \pi_{Ag})}{P\,([S_V{=}s_v]\ [S_C{=}s_c] \mid \pi_{Ag})}\ . \tag{6.1}$$

Il s'agit du calcul d'une distribution de probabilité sur les variables M_V et M_C, dans lequel le dénominateur $P\,([S_V{=}s_v]\ [S_C{=}s_c] \mid \pi_{Ag})$ est une constante. Ce terme permet d'assurer la normalisation de la distribution de probabilité conditionnelle $P\,(M_V\ M_C \mid [S_V{=}s_v]\ [S_C{=}s_c]\ \pi_{Ag})$, et on peut se contenter de le calculer a posteriori. On peut donc écrire, en utilisant l'opérateur de proportionalité $\propto$:

$$P\,(M_V\ M_C \mid [S_V{=}s_v]\ [S_C{=}s_c]\ \pi_{Ag}) \propto P\,(M_V\ M_C\ [S_V{=}s_v]\ [S_C{=}s_c] \mid \pi_{Ag})\ . \tag{6.2}$$

Par utilisations successives de la règle du produit, ceci peut s'écrire :

$$\begin{aligned} &P\,(M_V\ M_C \mid [S_V{=}s_v]\ [S_C{=}s_c]\ \pi_{Ag}) \propto \\ &\quad P\,(M_V \mid \pi_{Ag}) \times \\ &\quad P\,(M_C \mid M_V\ \pi_{Ag}) \times \\ &\quad P\,([S_V{=}s_v] \mid M_C\ M_V\ \pi_{Ag}) \times \\ &\quad P\,([S_C{=}s_c] \mid [S_V{=}s_v]\ M_C\ M_V\ \pi_{Ag})\ . \end{aligned} \tag{6.3}$$

La décomposition de la distribution de probabilité conjointe du modèle, qui est présentée figure 6.1 repose sur des hypothèses d'indépendances conditionnelles, parmi lesquelles certaines permettent de simplifier l'expression précédente. En réorganisant les termes, on obtient :

$$\begin{aligned} &P\,(M_V\ M_C \mid [S_V{=}s_v]\ [S_C{=}s_c]\ \pi_{Ag}) \propto \\ &\quad P\,(M_V \mid \pi_{Ag}) \times P\,([S_V{=}s_v] \mid M_V\ \pi_{Ag}) \times \\ &\quad P\,(M_C \mid M_V\ \pi_{Ag}) \times P\,([S_C{=}s_c] \mid M_C\ \pi_{Ag})\ . \end{aligned} \tag{6.4}$$

Le terme $P(M_V \mid \pi_{Ag})$ étant une distribution de probabilité uniforme, il reste donc trois termes qui ont un impact sur le calcul de la tâche d'imitation :

- la probabilité que le geste moteur de voyelle M_V permette d'atteindre la cible s_v d'après le modèle interne $P(S_V \mid M_V\ \pi_{Ag})$,
- la probabilité que le geste moteur de consonne M_C permette d'atteindre la cible s_c d'après le modèle interne $P(S_C \mid M_C\ \pi_{Ag})$,
- et la probabilité que le geste moteur de consonne M_C soit combiné avec le geste moteur de voyelle M_V d'après les contraintes de coarticulation $P(M_C \mid M_V\ \pi_{Ag})$.

Ainsi, ce calcul d'inférence probabiliste montre que cette tâche d'imitation pure est réalisée en ne sollicitant que le système sensori-moteur de l'agent, et en particulier ses modèles internes de la transformation articulatori-acoustique pour les voyelles et pour les consonnes. Le terme $P(M_C \mid M_V\ \pi_{Ag})$ quant à lui impose un lien fort entre la consonne et la voyelle, ce qui a pour effet de faciliter le calcul de la tâche d'imitation. En effet, la distribution de probabilité correspondante, $P\,(M_V\ M_C \mid [S_V{=}s_v\ ]\ [S_C{=}s_c]\ \pi_{Ag})$, est une distribution de probabilité sur $M_V \times M_C$, c'est-à-dire un espace de taille $25^3 \times 25^5 \approx 10^{11}$. Notre principe d'économie motrice encode le fait que l'agent choisit de réaliser ses consonnes en superposant à la voyelle un geste de constriction, en combinant le paramètre *Jaw* avec le paramètre *LipHeight*, *TongueApex* ou *TongueDorsum* pour produire respectivement une labiale, une dentale ou une vélaire. Le choix d'une distribution de probabilité uniforme par plateaux pour le terme $P(M_C \mid M_V\ \pi_{Ag})$ réduit le nombre de consonnes que l'on peut obtenir à partir d'une voyelle m_v donnée à 25^2 (la constriction s'obtient en sollicitant deux articulateurs) ; à tous les autres gestes elle associe une valeur de probabilité nulle. Grâce à ce principe d'économie motrice, il suffit donc de calculer la distribution $P\,(M_V\ M_C \mid [S_V{=}s_v\ ]\ [S_C{=}s_c]\ \pi_{Ag})$ sur un espace de taille $25^3 \times 25^2 \approx 10^7$, ce qui est bien inférieur à 10^{11}.

2.4.2 L'algorithme d'apprentissage des modèles internes par accomodation

L'algorithme que nous proposons ici permet à l'agent d'apprendre les liens entre les variables motrices et les variables perceptives. Cela dit, certains des termes qui correspondent, dans la décomposition de la distribution de probabilité conjointe du modèle, au système sensori-moteur de l'agent ne sont pas appris. En effet, le prior sur les gestes de voyelles $P(M_V \mid \pi_{Ag})$ est fixé comme étant une distribution de probabilité uniforme pour traduire le fait que l'agent ne choisit aucun geste plus fréquemment que les autres, et le terme de contraintes de coarticulation $P(M_C \mid M_V\ \pi_{Ag})$, qui encode le principe d'économie motrice, est également fixé une fois pour toutes.

En revanche l'agent met à jour les connaissances stockées dans ses modèles internes de la transformation articulatori-acoustique pour les consonnes et pour les voyelles au cours de l'apprentissage. Initialement, les moyennes $\mu^S_{m_v}$ et $\mu^S_{m_c}$ des gaussiennes $P\,(S_V \mid [M_V{=}m_v]\ \pi_{Ag})$ et $P\,(S_C \mid [M_C{=}m_c]\ \pi_{Ag})$ sont fixées au centre des espaces S_V et S_C, et leurs matrices de covariance $\sigma^S_{m_v}$ et $\sigma^S_{m_c}$ sont initialement diagonales, avec des valeurs très grandes par rapport à la taille des espaces S_V et S_C. De cette manière, les gaussiennes sont dégénérées et se comportent comme des distributions de probabilité uniformes, ce qui permet d'encoder le fait

qu'initialement l'agent ne dispose d'aucune connaissance sur le lien qui existe entre les consignes motrices et leurs conséquences perceptives.

L'algorithme d'apprentissage itère alors les étapes suivantes.

La production d'une syllabe par le maître se fait en deux temps. Tout d'abord l'agent maître choisit un objet syllabe o uniformément, puis la production de la syllabe correspondante est simulée en effectuant un tirage aléatoire selon la distribution de probabilité $P\,(S_V\ S_C \mid [O{=}o]\ \pi_{App})$. L'agent reçoit ainsi un percept de syllabe (s_v, s_c), qu'il va essayer de reproduire.

Le choix de gestes moteurs pour imiter le maître se fait en tirant aléatoirement des gestes moteurs (m_v, m_c) susceptibles de correspondre à la cible (s_v, s_c) selon la distribution de probabilité $P\,(M_V\ M_C \mid [S_V{=}s_v\ ]\ [S_C{=}s_c]\ \pi_{Ag})$ qui, calculée de la manière décrite à la section 2.4.1, permet de réaliser la tâche d'imitation de syllabe. Ce calcul fait intervenir l'état de connaissance stocké dans les modèles internes de l'agent à un instant donné.

La réalisation de la consigne motrice : l'agent envoie cette consigne motrice (m_v, m_c) à son système moteur, ce qui a pour effet de produire une conséquence perceptive (s'_v, s'_c). Cette valeur de (s'_v, s'_c) est obtenue en utilisant directement un modèle physique externe, qui est bien distinct du modèle interne de l'agent, qui, lui, est cognitif. Alors que le premier est un modèle qui encode la physique de la transformation articulatori-acoustique (telle qu'elle est modélisée par *VLAM*), le second est un modèle cognitif qui encode un état de connaissance à un moment donné, que l'agent se construit progressivement à partir de l'observation des conséquences de ses actions sur sa perception. Le couple de valeurs (s'_v, s'_c) peut évidemment différer de la cible (s_v, s_c) que le maître a donné à imiter à l'agent.

La mise à jour des représentations internes : l'agent met à jour ses modèles internes de la transformation articulatori-acoustique à deux niveaux : l'observation selon laquelle la consigne motrice m_v a la conséquence perceptive s'_v lui permet de mettre à jour la moyenne $\mu^S_{m_v}$ et la matrice de covariance $\sigma^S_{m_v}$ de la gaussienne $P\,(S_V \mid [M_V{=}m_v]\ \pi_{Ag})$; et l'observation selon laquelle la consigne motrice m_c a la conséquence perceptive s'_c lui permet de mettre à jour la moyenne $\mu^S_{m_c}$ et la matrice de covariance $\sigma^S_{m_c}$ de la gaussienne $P\,(S_C \mid [M_C{=}m_c]\ \pi_{Ag})$.

Ainsi, au fur et à mesure que l'agent accumule de nouvelles observations, notre algorithme passe progressivement d'un stade initial dépourvu de connaissances, qui est caractérisé par des productions aléatoires pouvant ressembler à du babillage canonique (ou à de l'exploration uniforme), à un second stade d'apprentissage de plus en plus fin des zones de l'espace sensori-moteur correspondant aux cibles à reproduire, jusqu'à convergence de l'apprentissage.

2.4.3 Illustrations de la convergence de l'algorithme

Pour pouvoir analyser la convergence de l'algorithme, il convient de se doter de mesures quantitatives dont l'évolution au cours de l'apprentissage nous renseigne sur ce qui est appris par l'agent au fur et à mesure de ses interactions avec le maître. C'est dans ce but que nous nous

intéressons à l'évolution de l'entropie des distributions de probabilité qui représentent les modèles internes de l'agent. Étant donnée une variable aléatoire discrète X ayant n réalisations x_i possibles, l'entropie H de $P(X)$ mesurée en bits est alors définie de la manière suivante : $H(P(X)) = -\sum_{i=1}^{n} P(X{=}x_i)\ log_2\ (P(X{=}x_i))$.

Intuitivement, la correspondance entre entropie et quantité d'information nous permet de mesurer quantitativement ce qui est appris par l'agent au cours de l'apprentissage. La figure 6.3 montre l'évolution de l'entropie $H(P(S_V\ S_C \mid \pi_{Ag}))$ des modèles internes de l'agent au cours de l'apprentissage.

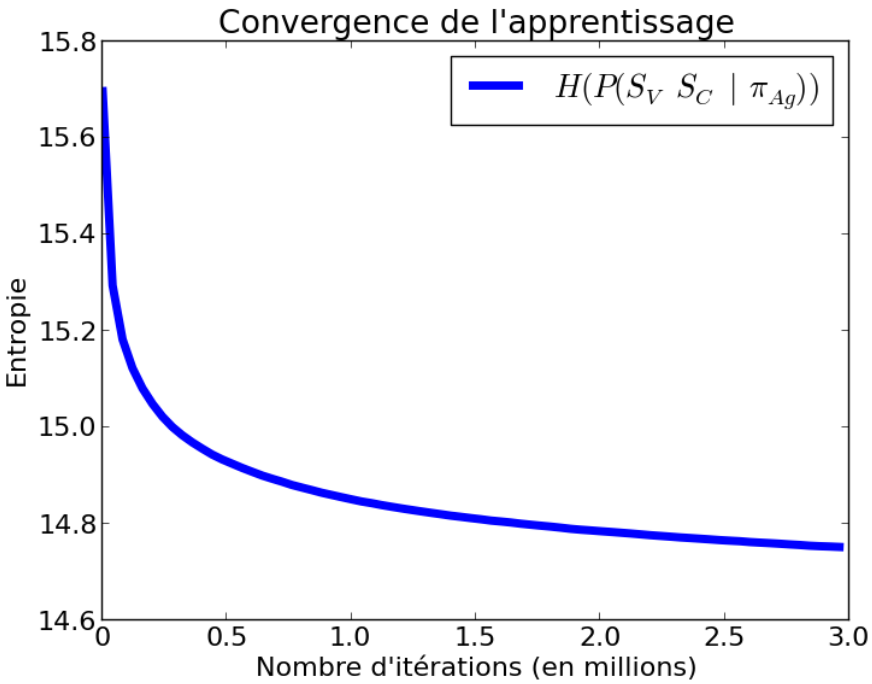


Figure 6.3: **Évolution de l'entropie** $H(P(S_V\ S_C \mid \pi_{Ag}))$ des modèles internes de l'agent au cours de l'apprentissage.

Au début de l'apprentissage, l'entropie a une valeur élevée, ce qui caractérise le fait que l'agent n'a encore rien appris. L'information stockée dans son modèle interne est alors très proche de distributions de probabilités uniformes : les gaussiennes dégénérées que nous utilisons pour l'initialisation. Lors des premières itérations de l'algorithme d'apprentissage, l'entropie diminue fortement : chaque nouvelle observation des conséquences de la réalisation d'une commande motrice apporte de l'information nouvelle à l'agent qui s'en sert pour mettre à jour son modèle interne. Au fur et à mesure que l'agent afine son modèle interne, il choisit de plus en plus fréquemment des gestes moteurs lui permettant de bien reproduire les cibles de syllabes proposées par le maître. Mais, comme il connaît déjà bien les conséquences de ces gestes-là, ces productions lui apportent de moins en moins d'information supplémentaire. En fin d'apprentissage, l'entropie est donc presque complètement stable, mais peut éventuellement diminuer un peu lorsque l'agent choisit pour imiter le maître un geste moteur qui a une faible probabilité de permettre d'atteindre la cible, et que la réalisation de ce geste procure à l'agent

une nouvelle information.

La figure 6.3, qui montre l'évolution de l'entropie $H(P(S_V\ S_C \mid \pi_{Ag}))$ des modèles internes de l'agent au cours de l'apprentissage, donne donc une mesure globale de ce qui est appris par l'agent. La valeur initiale de $15,7$ est à comparer avec l'entropie de la distribution de probabilité uniforme qui sert à l'initialisation. Étant données les tailles de nos espaces, l'entropie d'une distribution de probabilité uniforme définie partout serait de $\log_2(16 \times 33 \times 33 \times 14) = 17,9$. Mais, dans notre cas, la distribution de probabilité $P(S_V\ S_C \mid \pi_{Ag})$ n'est définie que sur les espaces acoustiques maximaux (qui sont introduits au chapitre 5, et apparraissent en gris sur les figures 6.11 et 6.12), ce qui explique la valeur initiale de $15,7$. En fin d'apprentissage, la valeur de $14,75$ est plus difficile à interpréter : elle dépend des connaissances accumulées dans toutes les distributions de probabilité du modèle interne de l'agent. Pour donner une idée, une configuration volontairement caricaturale, qui se traduirait par le même écart observé d'environ 1 par rapport à l'entropie de la distribution de probabilité uniforme sur notre espace $(15,7)$, serait le cas où l'agent a parfaitement appris la transformation articulatori-acoustique sur la moitié de l'espace (avec des distributions de probabilité de type Dirac, d'entropie nulle) et ne saurait rien sur la seconde moitié de l'espace (distributions de probabilité uniformes, d'entropie $15,7 - \log_2(2) = 14,7$).

L'entropie $H(P(S_V\ S_C \mid \pi_{Ag}))$ fournit donc une mesure globale, qui nous permet de nous assurer de la convergence de l'apprentissage, mais qui est difficile à interpréter finement. C'est pourquoi nous proposons maintenant d'illustrer plus en détails ce qui est appris au niveau du modèle interne de l'agent pour la transformation articulatori-acoustique des voyelles. Il s'agit du terme $P(S_V \mid M_V\ \pi_{Ag})$ de la distribution de probabilité conjointe du modèle, qui est constitué d'autant de distributions de probabilité $P\left(S_V \mid [M_V{=}m_v]\ \pi_{Ag}\right)$ qu'il y a de valeurs m_v possibles pour la variable M_V. On peut donc calculer l'entropie de chacune de ces distributions, et comparer les valeurs obtenues pour voir quelles sont les zones de l'espace moteur pour lesquelles la transformation articulatori-acoustique des voyelles a été mieux apprise.

Cependant, mesurer l'entropie des distributions de probabilité apprises ne suffit pas à décrire complètement ce qui est appris. En effet, bien qu'intuitivement des valeurs d'entropie faibles ou élevées traduisent différents degrés de certitude ou d'incertitude pour l'agent, ces valeurs n'ont de sens que relativement à la complexité du phénomène internalisé par l'agent. Ainsi, un modèle interne ayant une valeur d'entropie élevée peut, en grossissant le trait, correspondre à deux cas extrêmes : soit l'agent a mal appris un phénomène « simple » (au sens où la réalisation de ce phénomène montre peu de variabilité) ; soit l'agent a bien appris un phénomène plus « complexe » (au sens où son issue est incertaine, plus variable, difficile à prévoir). C'est pourquoi nous proposons de décrire l'apprentissage des distributions de probabilité $P\left(S_V \mid [M_V{=}m_v]\ \pi_{Ag}\right)$[37] en utilisant comme critère de mesure la différence entre l'entropie de la distribution apprise et l'entropie de référence de la distribution de probabilité correspondant au phénomène physique étant appris. Il s'agit donc de comparer l'entropie H du modèle interne $P(S_V \mid M_V\ \pi_{Ag})$ de l'agent avec l'entropie H_{ref} du modèle de l'environnement $P(S_V \mid M_V\ \pi_{Env})$ simulé par $VLAM$. [38]. Ainsi, nous définissons une grandeur $\Delta H = H - H_{ref}$ qui permet de quantifier pour chaque

[37]Pour ne pas surcharger l'analyse, nous ne discutons ici que du modèle interne de la transformation articulatori-acoustique pour les voyelles, les résultats sont similaires pour les distributions de probabilité $P\left(S_C \mid M_C\ \pi_{Ag}\right)$ apprises.

[38]Dans le cas d'un phénomène physique représenté de manière entièrement déterministe, cette entropie de référence est nulle. En revanche, le phénomène dont il est question ici est la transformation articulatori-acoustique

distribution ce que l'agent a effectivement appris par rapport à ce qu'il pouvait apprendre. En particulier, si l'hypothèse **H1** d'apprentissage parfait du théorème d'indistinguabilité est vérifiée, on a $\Delta H = 0$ pour toutes les valeurs m_v de M_V.

La figure 6.4 présente sous forme d'histogramme la répartition de l'effectif des différentes distributions $P\,(S_V \mid [M_V{=}m_v]\ \pi_{Ag})$ apprises par l'agent en fonction de leur valeur de ΔH.

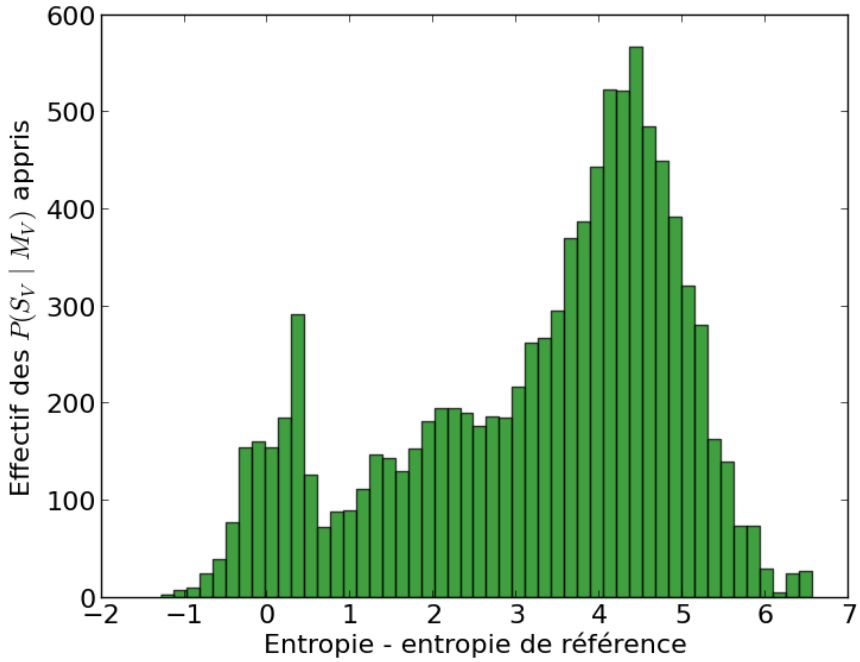


Figure 6.4: **Entropie des distributions $P\,(S_V \mid [M_V{=}m_v]\ \pi_{Ag})$ apprises** : sur cette figure l'ensemble des valeurs de ΔH est divisé en cinquante classes, et l'on montre le nombre de distributions $P\,(S_V \mid [M_V{=}m_v]\ \pi_{Ag})$ correspondant à chaque classe au terme de l'apprentissage.

Cette figure 6.4 présente deux régions particulièrement intéressantes. On observe tout d'abord un premier pic autour de 0, qui correspond aux distributions qui ont été très bien apprises. Une valeur de ΔH à 0 pour une distribution de probabilité $P\,(S_V \mid [M_V{=}m_v]\ \pi_{Ag})$ signifie que l'agent a convergé dans son modèle interne vers une distribution dont l'entropie est la même que celle de la distribution de probabilité de référence qui décrit les conséquences acoustiques que *VLAM* prévoit pour les gestes situés dans le cube de l'espace moteur correspondant à m_v. Une valeur de ΔH négative signifie que l'agent s'est construit un modèle interne simplifié, qui ne rend pas compte de toute la variabilité prévue par la distribution *VLAM* de référence. Cela peut être lié au choix qui a été fait pour la forme paramétrique : le terme $P(S_V \mid M_V\ \pi_{Ag})$ est encodé dans le modèle sous forme de distribution de probabilité gaussienne, et ce résumé gaussien peut causer une perte d'information. Une valeur négative de ΔH peut également

réalisée par *VLAM*, dont nous avons construit une représentation probabiliste. En effet, pour des raisons computationnelles, nous avons limité dans le modèle cognitif *COSMO-S* le pas de discrétisation des espaces M_V et M_C qui est plus grossier que celui ayant été utilisé pour échantillonner le phénomène physique réalisé par *VLAM*. À chaque point m_v ou m_c correspond donc une distribution de probabilité sur S_V ou S_C qui traduit la variabilité de la transformation articulatori-acoustique dans l'hypercube centré sur le point considéré.

s'expliquer comme une conséquence du surapprentissage qui peut éventuellement avoir lieu si les tirages aléatoires des conséquences acoustiques s'_v des gestes d'imitation m_v conduisent à des valeurs moins différentes que ne le prévoit en moyenne la distribution de référence.

On observe également un second pic autour de la valeur 4.5, qui correspond à des distributions de probabilité qui ont été peu apprises, pour lesquelles l'agent ne s'est construit qu'un modèle interne très approximatif, imprécis. La zone entre ces deux pics correspond à des distributions de probabilité qui commencent à se piquer, c'est-à-dire que dans cette zone, le modèle interne de l'agent contient déjà un certain niveau de description du modèle de l'environnement, même s'il est encore imprécis.

Dans l'ensemble, cette figure 6.4 montre que l'algorithme d'apprentissage que nous avons proposé permet à l'agent de concentrer son apprentissage sur certaines des distributions de probabilité $P\,(S_V \mid [M_V{=}m_v]\ \pi_{Ag})$, sans avoir besoin de devoir explorer exhaustivement tout l'espace moteur M_V ni d'apprendre totalement la transformation articulatori-acoustique (ce qui ne serait pas réaliste). Regardons maintenant quels sont ces gestes moteurs m_v pour lesquels l'agent a internalisé la transformation articulatori-acoustique de manière précise.

La figure 6.5 montre grâce à une carte de couleurs la valeur de ΔH associée à chacune des distributions $P(S_V \mid [M_V{=}m_v]\ \pi_{Ag})$.

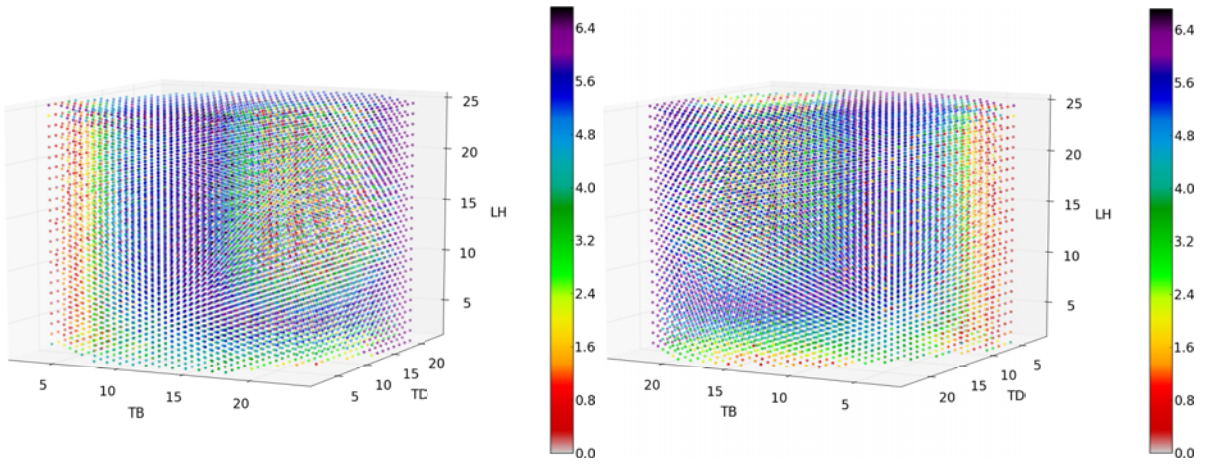


Figure 6.5: **Visualisation de la précision du modèle interne appris** en fonction des zones de l'espace moteur des voyelles. Pour chaque point m_v, caractérisé par les valeurs des paramètres *TongueBody* (TB), *TongueDorsum* (TD) et *LipHeight* (LH) de *VLAM*, la couleur correspond à la valeur de ΔH qui quantifie l'information apprise par l'agent dans sa distribution $P\,(S_V \mid [M_V{=}m_v]\ \pi_{Ag})$ correspondante. Les figures de gauche et de droite montrent les mêmes données sous deux angles de vues opposés.

Comme le montre la figure 6.5, l'algorithme d'apprentissage proposé fait que l'agent apprend la transformation articulatori-acoustique pour les voyelles de manière très inégale sur l'espace des gestes moteurs M_V : quelques zones sont bien apprises, ce qui se traduit par une faible valeur de ΔH, mais la majeure partie de l'espace reste peu explorée.

Pour une meilleure lisibilité, nous montrons maintenant figure 6.6 quels sont les gestes moteurs pour lesquels l'agent s'est construit un modèle interne plus précis.

Prises ensembles, les figures 6.4, 6.5 et 6.6 montrent comment, au lieu d'explorer exhaustivement l'espace moteur, l'agent apprend son modèle interne de la transformation articulatori-

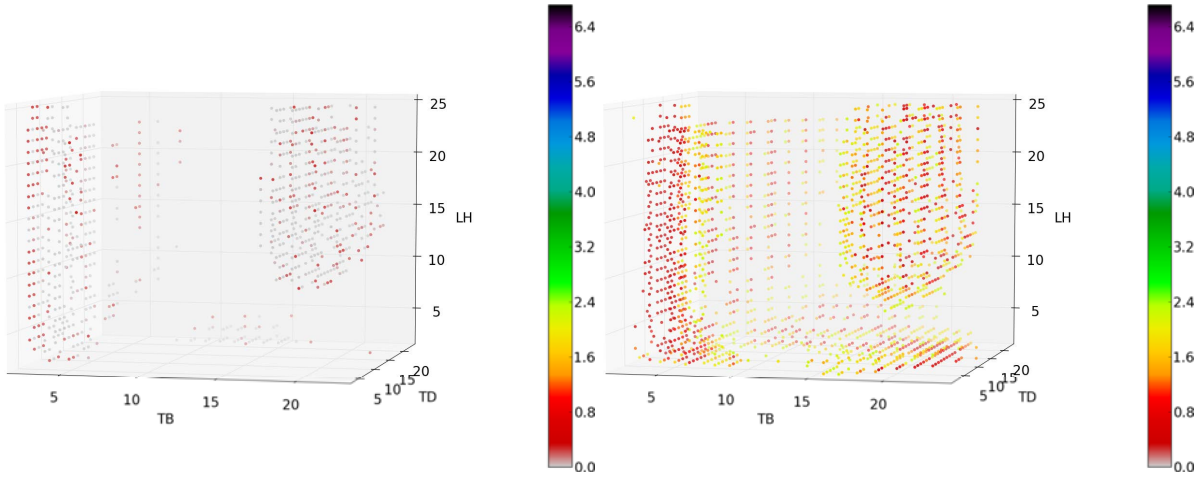


Figure 6.6: **Les gestes moteurs de voyelles les mieux appris.** Ne tracer que les points correspondant à certaines des valeurs de ΔH permet de rendre les données précédentes un peu plus facile à lire. À gauche, on ne présente que les valeurs de ΔH inférieures à 0.3, qui correspondent à des distributions $P\left(S_V \mid [M_V{=}m_v]\ \pi_{Ag}\right)$ très bien apprises ; et à droite on ne présente que les valeurs de H comprises entre 0.3 et 2.3, qui correspondent à des distributions un peu plus imprécises, mais encore relativement bien connues par l'agent à la fin de l'apprentissage. Pour chaque point m_v, caractérisé par les valeurs des paramètres *TongueBody* (TB), *TongueDorsum* (TD) et *LipHeight* (LH) de *VLAM*, la couleur correspond à la valeur de ΔH qui quantifie l'information apprise par l'agent dans sa distribution $P\left(S_V \mid [M_V{=}m_v]\ \pi_{Ag}\right)$ correspondante.

acoustique en se focalisant sur certaines zones de l'espace. Il s'agit des gestes moteurs qui lui permettent de bien imiter les syllabes produites par le maître et d'atteindre les cibles acoustiques qu'il lui propose. En effet, l'équation 6.4 montre que le choix du geste moteur m_v pour la tâche d'imitation fait intervenir le terme $P\left(S_V \mid [M_V{=}m_v]\ \pi_{Ag}\right)$, ce qui montre que le geste m_v est choisi avec une probabilité d'autant plus grande qu'il correspond bien à la cible acoustique à reproduire. Ainsi, au fur et à mesure que l'agent se construit son modèle interne de la transformation articulatori-acoustique, il choisit également des gestes permettant d'atteindre les cibles de mieux en mieux. C'est ce qu'illustre la figure 6.7 qui montre les zones de l'espace acoustique atteintes par l'agent à différents moments de l'apprentissage : au début (les 500 premières voyelles réalisées), en cours d'apprentissage (les 500 voyelles suivant la 28000-ième itération) et à la fin (les 500 dernières voyelles produites par l'agent au bout de trois millions d'itérations).

La figure 6.7 montre d'une part que, au début de l'apprentissage, les productions de l'agent peuvent tomber un peu n'importe où dans l'espace acoustique ; et d'autre part que, par la suite, l'agent atteint de mieux en mieux les cibles à repoduire : ses productions s'alignent progressivement sur celles du maître.

La figure 6.8 présente les mêmes données que la figure 6.7, mais dans l'espace moteur. Au début de l'apprentissage par imitation, les gestes moteurs choisis par l'agent sont répartis de manière quasi uniforme dans l'espace moteur, puis ils se concentrent progressivement autour des zones permettant de bien atteindre les cibles proposés par le maître.

En considérant simultanément les figures 6.7 et 6.8, il appert d'un côté que, en fin d'apprentissage, les productions de l'agent sont alignées sur celles du maître, très localisées dans l'espace

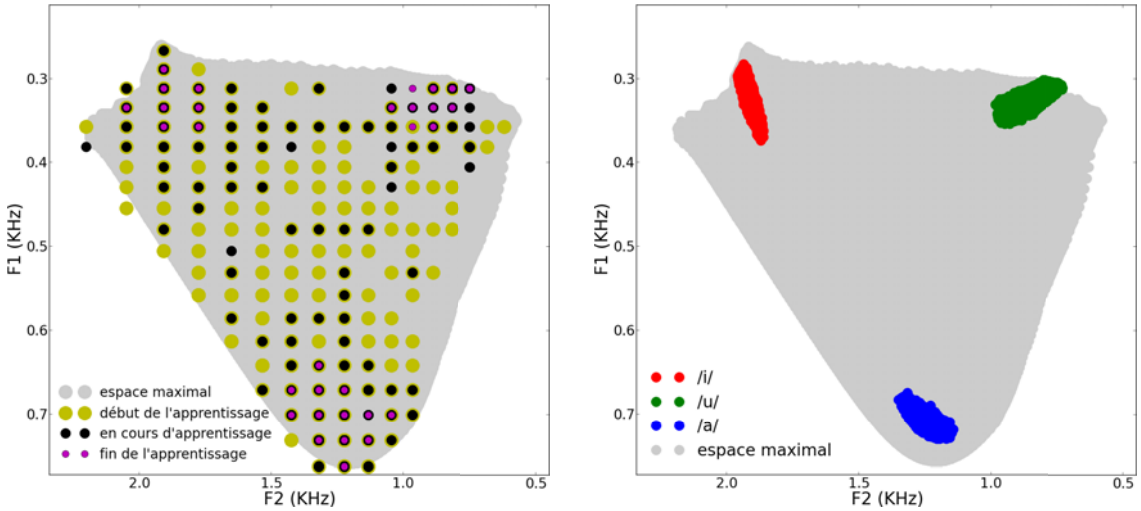


Figure 6.7: **Les voyelles produites au cours de l'apprentissage par imitation** (à gauche) sont de plus en plus proches des cibles fournies par le maître (à droite). Les valeurs des formants de l'espace maximal et des productions du maître sont des valeurs de variables continues, alors que sur la figure de gauche les formants sont présentés avec la discrétisation du modèle interne de l'agent.

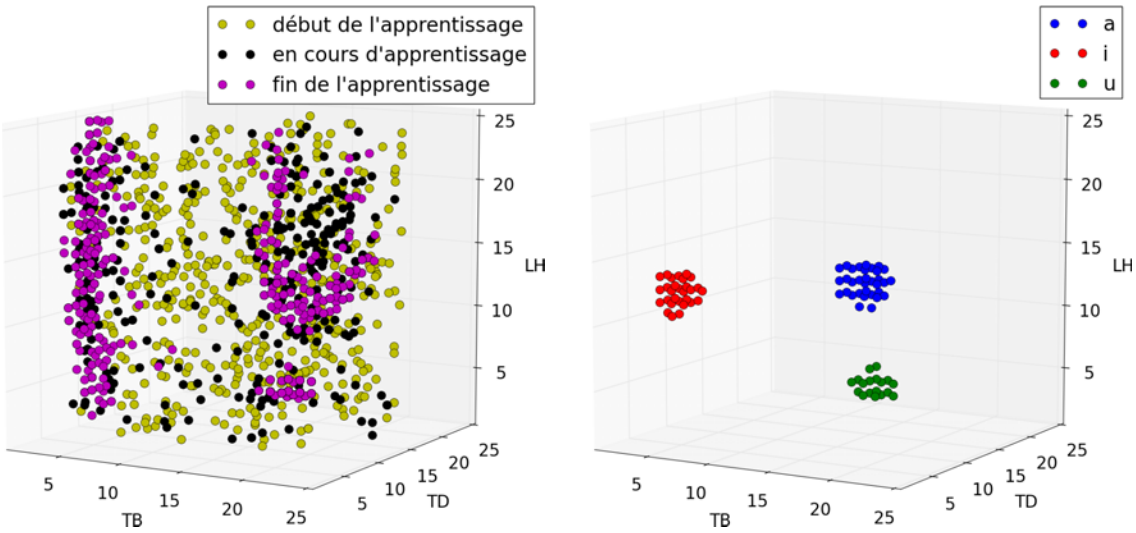


Figure 6.8: **Les gestes moteurs de voyelles produits au cours de l'apprentissage par imitation** (à gauche), sont comparés aux gestes moteurs correspondant aux cibles produites par le maître (à droite). Les voyelles sont présentées dans l'espace des paramètres *TongueBody* (TB), *TongueDorsum* (TD) et *LipHeight* (LH) de *VLAM*.

acoustique ; et de l'autre on voit qu'il y a davantage de variabilité dans les différents gestes moteurs choisis par l'agent. Il s'agit là d'une conséquence du caractère *many-to-one* (c'est-à-dire non injectif) de la transformation articulatori-acoustique : une même cible acoustique peut avoir de nombreux antécédents moteurs différents. Avec l'algorithme d'apprentissage que nous avons présenté, l'agent inclut dans son modèle interne l'ensemble des antécédents moteurs correspondant aux syllabes fournies par le maître. Cela peut sembler contradictoire avec ce qui est observé en pratique : bien que l'on ait peu de données sur des tâches d'imitation, on observe en revanche qu'il y a assez peu de variabilité dans les gestes moteurs que réalise un locuteur pour produire une même voyelle. La section suivante propose un mécanisme permettant d'ancrer des préférences au cours de l'apprentissage, des choix idiosyncrasiques de gestes moteurs qui font que l'agent, au lieu de considérer la diversité des antécédents correspondant à une même cible, se focalise sur une portion de l'espace moteur sur laquelle il se spécialise.

2.5 Apprentissage du système moteur par accomodation

Cette section décrit comment modifier l'algorithme d'apprentissage par accomodation présenté à la section précédente pour l'adapter à l'apprentissage de compétences motrices. Nous commençons par définir formellement la tâche d'imitation de syllabes, puis nous décrivons l'algorithme d'apprentissage, dont nous illustrons enfin la convergence.

2.5.1 Inférence probabiliste pour la tâche d'imitation supervisée de syllabes

Alors que l'apprentissage présenté à la section 2.4 se fait de manière non-supervisée, l'agent étant simplement immergé dans un bain audio composé des productions syllabiques (s_v, s_c) du maître, dans la présente section on procède de manière supervisée : le maître fournit également à l'agent, par un mécanisme d'attention partagée, la catégorie symbolique o de l'objet syllabe communiqué. La tâche d'imitation se formalise alors de la manière suivante : étant donné un stimulus (s_v, s_c) fourni par le maître et sa catégorie d'objet syllabe o, l'agent doit choisir une consigne motrice (m_v, m_c) correspondant à la fois à la catégorie d'objet o et à la cible (s_v, s_c) à atteindre.

Dans ce contexte, réaliser une tâche d'imitation revient à tirer aléatoirement un geste moteur (m_v, m_c) selon une distribution de probabilité de la forme $P\left(M_V\ M_C \mid S_V\ S_C\ O_S\ \pi_{Ag}\right)$. Contrairement au calcul de la section 2.4.1 qui pouvait se faire uniquement au sein du système sensori-moteur de l'agent, le calcul que nous faisons maintenant sollicite également le système moteur de l'agent, qui associe gestes moteurs et objets. Dans notre modèle *COSMO-S* (qui est représenté figure 6.1), la fusion des informations apportées par le système sensori-moteur et par le système moteur se fait grâce aux variables λ_{MV} et λ_{MC} qui assurent la cohérence entre les représentations motrices dans les deux sous-systèmes du modèle de l'agent. Par exemple, conditionner le calcul d'une distribution de probabilité par la connaissance $\lambda_{MV} = 1$ impose que les variables M_V et M'_V aient des valeurs égales à tout moment du calcul d'inférence. Alors que dans le système sensori-moteur de l'agent les syllabes sont représentées par les variables M_V et M_C, dans son système moteur elles sont représentées par les variables M'_V, G'_C et Δ'_{MC}. Dès lors, la tâche d'imitation supervisée de syllabe consiste donc à calculer la distribution de probabilité suivante :

$$P\left(M'_V\ G'_C\ \Delta'_{MC} \mid S_V\ S_C\ O_S\ [\lambda_{MV}{=}1]\ [\lambda_{MC}{=}1]\ \pi_{Ag}\right)\ .$$

La structure de notre modèle dans lequel les deux sous-systèmes sont reliés par des variables de cohérence fait que, pour toute syllabe $(m_v, g, \delta_{mc}) \in M_V \times G'_C \times \Delta'_{MC}$, le calcul de cette distribution de probabilité se décompose de la manière suivante :

$$\begin{aligned}
&P\left([M'_V{=}m_v]\ [G'_C{=}g]\ [\Delta'_{MC}{=}\delta_{mc}] \mid [S_V{=}s_v]\ [S_C{=}s_c]\ [O_S{=}o]\ [\lambda_{MV}{=}1]\ [\lambda_{MC}{=}1]\ \pi_{Ag}\right) \propto \\
&\quad P\left([M_V{=}m_v]\ [M_C{=}m_v{+}\delta_{mc}] \mid [S_V{=}s_v\ ]\ [S_C{=}s_c]\ \pi_{Ag}\right) \times \\
&\quad P\left([M'_V{=}m_v]\ [G'_C{=}g]\ [\Delta'_{MC}{=}\delta_{mc}] \mid [O_S{=}o]\ \pi_{Ag}\right)\ .
\end{aligned} \tag{6.5}$$

Le premier terme, $P\left([M_V{=}m_v]\ [M_C{=}m_v{+}\delta_{mc}] \mid [S_V{=}s_v\ ]\ [S_C{=}s_c]\ \pi_{Ag}\right)$, traduit le fait que les gestes moteurs choisis par l'agent doivent permettre d'atteindre la cible (s_v, s_c) le mieux possible. Comme dans la tâche d'imitation précédente, il se calcule dans le système sensori-moteur de la manière décrite par l'équation 6.4. On obtient alors :

$$\begin{aligned}
&P\left([M'_V{=}m_v]\ [G'_C{=}g]\ [\Delta'_{MC}{=}\delta_{mc}] \mid [S_V{=}s_v]\ [S_C{=}s_c]\ [O_S{=}o]\ [\lambda_{MV}{=}1]\ [\lambda_{MC}{=}1]\ \pi_{Ag}\right) \propto \\
&\quad P\left([M_V{=}m_v] \mid \pi_{Ag}\right) \times P\left([S_V{=}s_v] \mid [M_V{=}m_v]\ \pi_{Ag}\right) \times \\
&\quad P\left([M_C{=}m_v{+}\delta_{mc}] \mid [M_V{=}m_v]\ \pi_{Ag}\right) \times P\left([S_C{=}s_c] \mid [M_C{=}m_v{+}\delta_{mc}]\ \pi_{Ag}\right) \times \\
&\quad P\left([M'_V{=}m_v]\ [G'_C{=}g]\ [\Delta'_{MC}{=}\delta_{mc}] \mid [O_S{=}o]\ \pi_{Ag}\right)\ .
\end{aligned} \tag{6.6}$$

Le second terme, $P\left([M'_V{=}m_v]\ [G'_C{=}g]\ [\Delta'_{MC}{=}\delta_{mc}] \mid [O_S{=}o]\ \pi_{Ag}\right)$, traduit le fait que l'agent choisit de manière privilégiée les gestes qu'il a l'habitude d'associer à l'objet o dans ses répertoires de gestes moteurs. Le calcul de ce terme se fait au sein du système moteur uniquement. D'après le théorème de Bayes, on a :

$$P([M'_V{=}m_v][G'_C{=}g][\Delta'_{MC}{=}\delta_{mc}] \mid [O_S{=}o]\ \pi_{Ag}) = \frac{P\left([O_S{=}o][M'_V{=}m_v][G'_C{=}g][\Delta'_{MC}{=}\delta_{mc}] \mid \pi_{Ag}\right)}{P\left([O_S{=}o] \mid \pi_{Ag}\right)}. \tag{6.7}$$

Le dénominateur étant constant puisque la distribution $P(O_S \mid \pi_{Ag})$ est uniforme, cette expression se simplifie ainsi :

$$P\left([M'_V{=}m_v][G'_C{=}g][\Delta'_{MC}{=}\delta_{mc}] \mid [O_S{=}o]\ \pi_{Ag}\right) \propto P\left([O_S{=}o][M'_V{=}m_v][G'_C{=}g][\Delta'_{MC}{=}\delta_{mc}] \mid \pi_{Ag}\right). \tag{6.8}$$

On utilise alors la décomposition de la distribution de probabilité conjointe du modèle, qui se trouve figure 6.1, ce qui permet d'obtenir :

$$\begin{aligned}
&P\left([M'_V{=}m_v]\ [G'_C{=}g]\ [\Delta'_{MC}{=}\delta_{mc}] \mid [O_S{=}o]\ \pi_{Ag}\right) \propto \\
&\quad P\left([O_S{=}o] \mid \pi_{Ag}\right) \times P\left([M'_V{=}m_v] \mid [O_S{=}o]\ \pi_{Ag}\right) \times \\
&\quad P\left([G'_C{=}g] \mid [O_S{=}o]\ \pi_{Ag}\right) \times P\left([\Delta'_{MC}{=}\delta_{mc}] \mid [M'_V{=}m_v]\ [G'_C{=}g]\ \pi_{Ag}\right)\ .
\end{aligned} \tag{6.9}$$

En combinant les équations 6.6 et 6.9 on obtient finalement le calcul d'inférence pour la tâche d'imitation supervisée :

$$
\begin{aligned}
&P\left([M'_V{=}m_v]\ [G'_C{=}g]\ [\Delta'_{MC}{=}\delta_{mc}]\ \mid\ [S_V{=}s_v]\ [S_C{=}s_c]\ [O_S{=}o]\ [\lambda_{MV}{=}1]\ [\lambda_{MC}{=}1]\ \pi_{Ag}\right) \propto \\
&\quad P\left([M_V{=}m_v]\ \mid \pi_{Ag}\right) \times P\left([S_V{=}s_v]\ \mid\ [M_V{=}m_v]\ \pi_{Ag}\right) \times \\
&\quad P\left([M_C{=}m_v{+}\delta_{mc}]\ \mid\ [M_V{=}m_v]\ \pi_{Ag}\right) \times P\left([S_C{=}s_c]\ \mid\ [M_C{=}m_v{+}\delta_{mc}]\ \pi_{Ag}\right) \times \\
&\quad P\left([O_S{=}o]\ \mid \pi_{Ag}\right) \times P\left([M'_V{=}m_v]\ \mid\ [O_S{=}o]\ \pi_{Ag}\right) \times \\
&\quad P\left([G'_C{=}g]\ \mid\ [O_S{=}o]\ \pi_{Ag}\right) \times P\left([\Delta'_{MC}{=}\delta_{mc}]\ \mid\ [M'_V{=}m_v]\ [G'_C{=}g]\ \pi_{Ag}\right)\ . \qquad (6.10)
\end{aligned}
$$

Bien que la structure du modèle *COSMO-S* soit ici plus complexe, le résultat de ce calcul d'inférence pour la tâche d'imitation est de la même forme que celui auquel on aboutit à l'équation 4.8 du chapitre 4. En effet, le calcul de la tâche d'imitation fait intervenir essentiellement deux types de termes : ceux qui appartiennnent au système moteur de l'agent, et ceux qui appartiennent à son système sensori-moteur. Ainsi, dans le choix des gestes moteurs susceptibles d'atteindre une cible (s_v, s_c) associée à l'objet o, interviennent à la fois l'état de connaissance de la transformation articulatori-acoustique stocké dans le modèle interne de l'agent, et les connaissances encodées dans les répertoires moteurs de l'agent sur les gestes qu'il est habitué à associer à l'objet o.

On retrouve donc les deux principes, déjà décrits à la section 3.3.4 du chapitre 4, sur lesquels repose notre algorithme d'apprentissage : l'affinage progressif du modèle interne de l'agent, et l'ancrage dans les répertoires moteurs de choix idiosyncrasiques. Toutefois, dans le présent chapitre, ces deux principes opèrent de manière successive : l'agent apprend ses modèles internes de la manière décrite à la section 2.4, et il n'évoluent plus ensuite lorsque dans un second temps l'agent se contruit ses répertoires de gestes moteurs à associer aux syllabes de la manière décrite dans la section 2.5. Ainsi, dans le calcul de la distribution de probabilité correspondant à la tâche d'imitation supervisée qui est donné à l'équation 6.10, les termes correspondant au système sensori-moteur de l'agent n'évoluent pas (et du point de vue computationnel ils peuvent être précalculés une fois pour toutes), mais l'apprentissage se fait au niveau des termes correspondant au système moteur de l'agent qui sont mis à jour au fur et à mesure que l'agent ancre progressivement des choix préférentiels de gestes à associer aux différentes syllabes.

2.5.2 L'algorithme d'apprentissage des répertoires moteurs par imitation supervisée

L'algorithme que nous proposons ici permet à l'agent de construire ses répertoires de gestes moteurs à associer aux différentes syllabes. Initialement, les moyennes $\mu_o^{M_V}$ et μ_{g,m_v}^{Δ} des gaussiennes $P(M'_V \mid O_S\ \pi_{Ag})$ et $P(\Delta'_{MC} \mid M'_V\ G'_C\ \pi_{Ag})$ sont fixées au centre des espaces, et leurs matrices de covariance $\sigma_o^{M'_V}$ et σ_{g,m'_v}^{Δ} sont des matrices diagonales dont les coefficients sont fixés à une valeur très grande par rapport aux tailles des espaces concernés. De cette manière, les gaussiennes sont dégénérées et se comportent comme des distributions de probabilité uniformes. De plus, les vecteurs de paramètres g'_o caractérisant les histogrammes $P(G'_C \mid O_S\ \pi_{Ag})$ encodent initialement le répertoire de gestes de constrictions par une distribution de probabilité uniforme. L'utilisation de ces distributions uniformes ou quasi-uniformes permet d'encoder le

fait qu'initialement l'agent ne dispose d'aucune connaissance sur les gestes moteurs qui peuvent être associés aux syllabes. L'algorithme d'apprentissage itère alors les étapes suivantes.

La production d'une syllabe par le maître se fait en deux temps. Tout d'abord l'agent maître choisit un objet syllabe o uniformément, puis la production de la syllabe correspondante est simulée en effectuant un tirage aléatoire selon la distribution de probabilité $P\left(S_V\ S_C \mid [O{=}o]\ \pi_{App}\right)$. En plus du percept de syllabe (s_v, s_c) reçu par l'agent, le maître lui fournit également, grâce à un mécanisme d'attention partagée, la valeur o de l'objet à reproduire.

Le choix de gestes moteurs pour imiter le maître se fait en tirant aléatoirement des gestes moteurs (m_v, g, δ_{mc}) susceptibles de correspondre à la cible (s_v, s_c) et à l'objet o selon la distribution de probabilité permettant de réaliser la tâche d'imitation de syllabe, $P\left([M'_V{=}m_v]\ [G'_C{=}g]\ [\Delta'_{MC}{=}\delta_{mc}] \mid [S_V{=}s_v]\ [S_C{=}s_c]\ [O_S{=}o]\ [\lambda_{MV}{=}1]\ [\lambda_{MC}{=}1]\ \pi_{Ag}\right)$, qui est calculée de la manière décrite à la section 2.5.1. Ce calcul fait intervenir l'état de connaissance stocké dans les modèles internes de l'agent (qui a été appris de la manière présentée à la section 2.4 et n'évolue plus) ainsi que l'état de connaissance stocké dans son système moteur (qui lui est mis à jour au cours de cet apprentissage).

La mise à jour des représentations internes : ayant choisi de produire le geste (m_v, g, δ_{mc}) l'agent met à jour ses représentations internes à plusieurs niveaux. Il associe le geste moteur voyelle m_v à la syllabe o en mettant à jour la moyenne $\mu_o^{M'_V}$ et la matrice de covariance $\sigma_o^{M'_V}$ de la distribution de probabilité gaussienne $P\left([M'_V{=}m_v] \mid [O_S{=}o]\ \pi_{Ag}\right)$ avec la valeur de m_v. Il associe également le geste de constriction g à la syllabe o en mettant à jour le vecteur g'_o de paramètres de l'histogramme $P\left([G'_C{=}g] \mid [O_S{=}o]\ \pi_{Ag}\right)$ grâce à la valeur de g. L'agent modifie également ses connaissances sur la manière dont le choix du geste de constriction g vient modifier la voyelle m_v pour produire la consonne $m_c{=}m_v{+}\delta_{mc}$ en mettant à jour la moyenne μ^{Δ}_{g,m_v} et l'écart-type σ^{Δ}_{g,m_v} de la distribution de probabilité gaussienne $P\left([\Delta'_{MC}{=}\delta_{mc}] \mid [M'_V{=}m_v]\ [G'_C{=}g]\ \pi_{Ag}\right)$.

Ainsi, au fur et à mesure qu'il réalise des gestes moteurs de syllabes pour imiter les cibles fournies par le maître, l'agent les ajoute à ses répertoires moteurs. Initialement, le choix se fait de manière quasi uniforme parmi les gestes permettant (selon le modèle interne de l'agent) d'atteindre les cibles. Par la suite, le mécanisme de renforcement que nous proposons fait que l'agent aura tendance à privilégier les gestes qu'il a souvent utilisés par le passé, et l'agent ancre ainsi au cours de l'apprentissage des choix idiosyncrasiques de gestes moteurs qui ne dépendent essentiellement que de la réalisation des premiers tirages aléatoires.

2.5.3 Illustrations de la convergence de l'apprentissage : développement des idiosyncrasies

De manière similaire à ce qui est fait à la section 2.4.3, nous illustrons ici la convergence de l'apprentissage du système moteur en nous focalisant sur les représentations motrices des voyelles, qui sont plus facilement visualisables.

Dans un premier temps nous nous donnons une mesure globale de ce qui est appris par l'agent : l'évolution de l'entropie $H(P(M'_V \mid O_S\ \pi_{Ag}))$ des répertoires moteurs de voyelle nous

renseigne sur la manière dont l'agent organise ses connaissances au cours de l'apprentissage. Pour chaque syllabe o parmi /ba-bi-bu-da-di-du-ga-gi-gu/, la figure montre l'évolution de l'entropie $H\left(P\left(M'_V \mid [O_S{=}o]\ \pi_{Ag}\right)\right)$ au cours de l'apprentissage.

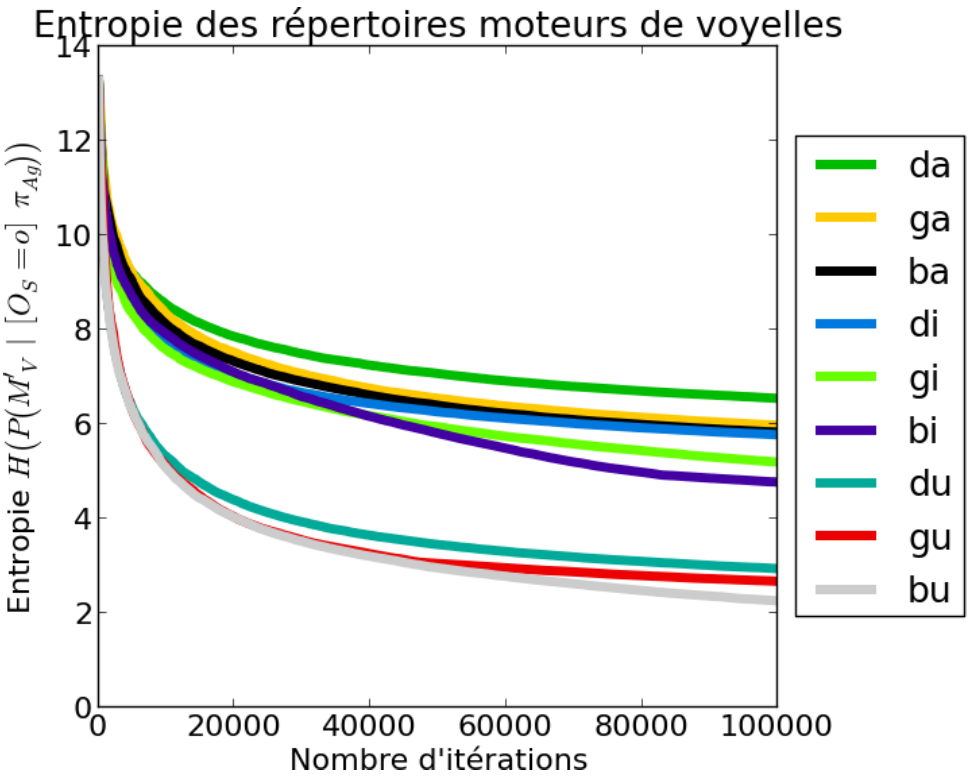


Figure 6.9: **Évolution de l'entropie des répertoires moteurs de voyelles au cours de l'apprentissage**. Chacune de ces neuf courbes décrit l'évolution au cours de l'apprentissage de l'entropie de la distribution de probabilité $P\left(M'_V \mid [O_S{=}o]\ \pi_{Ag}\right)$ pour chacune des neuf valeurs de o parmi /ba bi bu da di du ga gi gu/.

Au début de l'apprentissage, les neuf distributions de probabilité $P\left(M'_V \mid [O_S{=}o]\ \pi_{Ag}\right)$ ont une valeur d'entropie élevée, ce qui caractérise le fait que l'agent n'a encore rien appris. L'information stockée dans ses répertoires moteurs de voyelles est alors très proche de distributions de probabilité uniformes, du fait de l'utilisation de nos gaussiennes dégénérées. A titre de comparaison, l'entropie d'une distribution de probabilité uniforme sur un espace de taille 25^3 est d'environ $13,9$ bits.

Lors des premières itérations de l'algorithme, l'entropie diminue fortement : chaque nouvelle syllabe produite par l'agent pour imiter le maître contribue fortement à diminuer la variance de la distribution de probabilité $P\left(M'_V \mid [O_S{=}o]\ \pi_{Ag}\right)$ associant les gestes moteurs de voyelles M'_V à l'objet o. Par la suite, les productions de l'agent apportent de moins en moins d'information nouvelle par rapport à l'historique des connaissances accumulées par l'agent, et l'entropie diminue de moins en moins vite.

Il est intéressant de remarquer que, bien que les trois groupes de syllabes /ba-da-ga/, /bi di gi/ et /bu du gu/ ont la voyelle en commun, les répertoires de gestes moteurs de voyelles que l'agent leur associe ont des profils différents. Par exemple, on peut voir sur la figure 6.9 que dans les répertoires de gestes de voyelles que l'agent s'est construit, le /i/ de /di/ est plus variable que

le /i/ de /bi/ puisque $H\left(P\left(M'_V \mid [O_S{=}di]\ \pi_{Ag}\right)\right) \approx 6$ alors que $H\left(P\left(M'_V \mid [O_S{=}bi]\ \pi_{Ag}\right)\right) \approx 5$. Ainsi, bien que la décomposition de la distribution de probabilité conjointe du modèles *COSMO-S* (voir figure 6.1) ne fasse pas apparaître de dépendence explicite de la voyelle vis-à-vis de la consonne, l'algorithme d'apprentissage que nous proposons a la propriété remarquable de permettre à l'agent de capturer des effets de coarticulation au niveau des répertoires de gestes moteurs que l'agent se construit pour les voyelles.

Regardons maintenant de manière plus précise quels sont les gestes moteurs que l'agent choisit d'associer préférentiellement aux différentes voyelles. La figure 6.10 compare les gestes moteurs réalisés par l'agent à différents stades de l'apprentissage : au début (les 500 premières voyelles réalisées), en cours d'apprentissage (les 500 voyelles suivant la 2000-ième itération) et à la fin (les 500 dernières voyelles produites par l'agent après 100000 itérations).

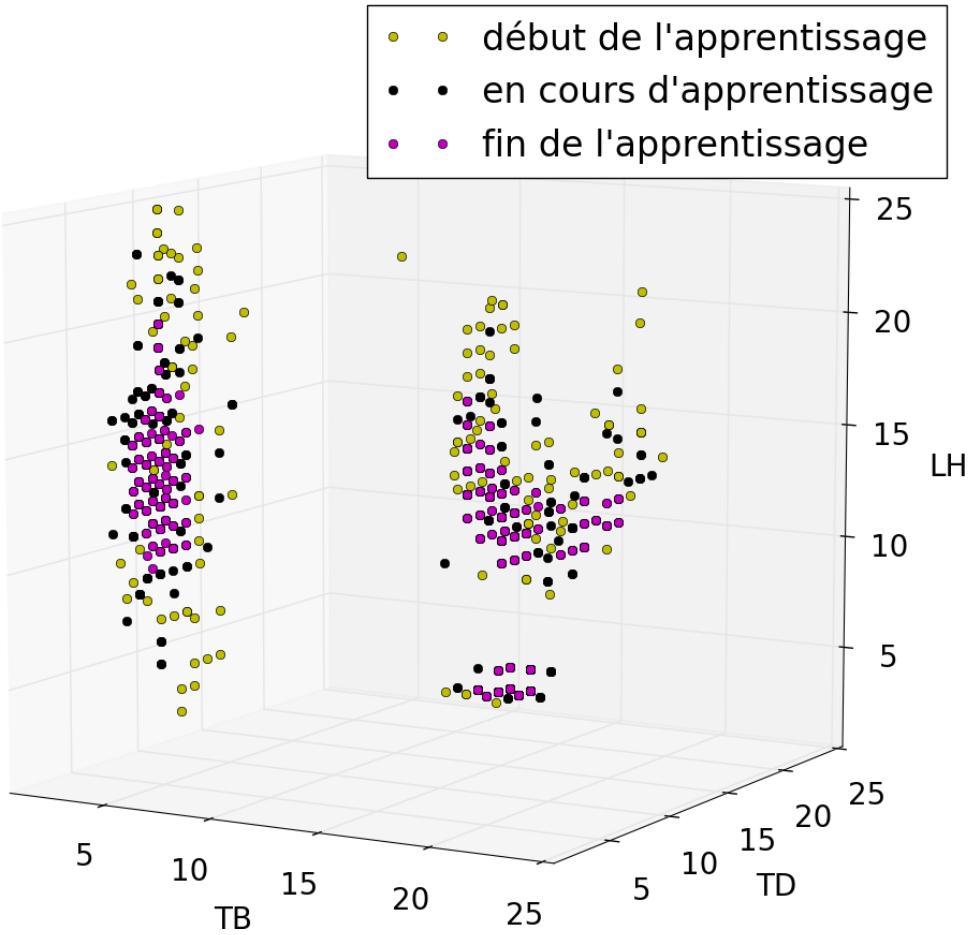


Figure 6.10: **Les gestes moteurs de voyelles produits au cours de l'apprentissage par imitation supervisée** se concentrent progressivement sur une zone de l'espace moteur de plus en plus petite au fur et à mesure que l'apprentissage se fait. Les voyelles sont présentées dans l'espace des paramètres *TongueBody* (TB), *TongueDorsum* (TD) et *LipHeight* (LH) de *VLAM*.

Cette figure montre comment l'ancrage de choix moteurs implémenté par notre mécanisme de renforcement permet à l'agent de résoudre le problème du *many-to-one* en développant progressivement des préférences pour certains gestes. Au début de l'apprentissage, les productions de l'agent sont les mêmes que celles réalisées à la fin de la phase d'apprentissage du système

sensori-moteur de l'agent (voir figure 6.8). En effet, le système moteur de l'agent ne contenant initialement aucune connaissance, ce qui est décrit par des distributions de probabilité quasi uniformes, l'inférence pour la tâche d'imitation présentée à l'équation 6.10 se simplifie : tous les termes correspondant au système moteur ont un poids constant, qui peut être absorbé par l'opérateur de proportionalité, ce qui conduit à l'équation 6.4 qui ne fait intervenir que les modèles internes de l'agent.

Ainsi, au début de l'apprentissage, le système moteur de l'agent n'apporte pas d'information pour le choix des gestes d'imitation, et l'agent utilise simplement les connaissances stockées dans ses modèles internes pour atteindre les cibles proposées par le maître. Ensuite, au fur et à mesure que l'agent continue à imiter le maître, il associe les commandes motrices correspondant aux syllabes communiquées par le maître en mettant à jour ses répertoires de gestes moteurs. De ce fait, confronté ensuite à une même cible, l'agent choisira avec une probabilité supérieure un geste qu'il y a déjà associé par le passé, rendant encore plus plausible le choix de ce même geste dans le futur. Ce mécanisme de renforcement fait que l'agent construit ses prototypes moteurs en ancrant des préférences liées aux réalisations des premiers tirages aléatoires ayant lieu au début de l'apprentissage. C'est ainsi que les premiers choix de gestes effectués par l'agent lui permettent de développer au cours de l'apprentissage des comportements qui lui sont spécifiques, c'est-à-dire de développer des idiosyncrasies qui le distinguent d'autres agents caractérisés par d'autres historiques d'apprentissage.

2.6 Synthèse des différents apprentissages réalisés

Finalement, dans cette section 2 nous avons présenté une manière d'implémenter l'apprentissage de compétences motrices et perceptives. Cet apprentissage que nous proposons repose sur les cinq idées suivantes.

- Nous avons découpé les apprentissages en trois étapes successives inspirées de la chronologie développementale proposée par Kuhl (2004) pour pouvoir illustrer trois principes essentiels. Ce découpage, qui nous est utile d'un point de vue computationnel, est schématiquement en accord avec le décours développemental d'ensemble tel que résumé par la figure 6.2, même si dans la réalité il est clair que le bébé doit traiter tous les problèmes d'apprentissage sur les différents niveaux de manière relativement parallèle.

- L'apprentissage par l'agent de prototypes auditifs de syllabes $P(S'_V\ S'_C \mid O_L\ \pi_{Ag})$, que nous avons choisi d'implémenter sous forme de distributions de probabilité gaussiennes, lui permet de développer des capacités perceptives et rendent possible la catégorisation de stimuli, grâce aux distances implicites définies par ces gaussiennes.

- Un paradigme d'accomodation est suffisant pour que, par imitation, l'agent apprenne progressivement, par accomodation, des modèles internes de plus en plus fins, sans que le maître n'ait besoin de lui procurer de *feedback*. Le fait d'utiliser des représentations probabilistes rend très naturelle l'inversion de la transformation articulatori-acoustique, bien qu'elle soit non injective (*many-to-one*) : au lieu de devoir se donner des critères compliqués pour choisir le meilleur antécédant moteur possible à une cible acoustique, l'agent effectue un tirage aléatoire selon la distribution de probabilité sur l'ensemble des candidats possibles.

- Un mécanisme d'ancrage des choix effectués, qui a pour effet de renforcer des préférences, permet à l'agent de développer des idiosyncrasies en focalisant ses répertoires moteurs sur un petit ensemble de gestes, ce qui constitue pour l'agent une manière de simplifier le caractère complexe (*many-to-one*) de la transformation articulatori-acoustique.

- De plus les trois étapes d'apprentissage se font grâce aux mêmes données : les productions du maître. L'agent est plongé dans un bain acoustique et utilise cette information pour se contruire des représentations auditives, des modèles internes de la transformation articulatori-acoustique, et des répertoires de gestes moteurs à associer à chaque classe de stimuli.

3 Principaux résultats

Notre modèle *COSMO-S* et les algorithmes d'apprentissage qui l'accompagnent composent un cadre très riche au sein duquel on peut étudier un grand nombre de questions intéressantes. Dans cette thèse nous avons choisi de mettre en avant deux résultats principaux : tout d'abord nous montrons comment l'apprentissage de compétences motrices par un agent purement perceptif permet de faire émerger une ébauche de phonologie, et dans un second temps nous montrons comment l'usage de compétences motrices apporte de la robustesse à la perception de parole en conditions dégradées.

3.1 L'apprentissage de compétences motrices par un agent purement perceptif permet de faire émerger une ébauche de phonologie

Parmi les questions que notre modèle *COSMO-S* nous permet d'étudier, on s'intéresse ici à l'émergence de la phonologie. La question est de savoir comment, à partir des stimuli constituant le bain audio dans lequel il est plongé, l'agent percevant parvient à faire émerger de ses représentations internes les notions de catégories vocalique et consonantique. Comment apprend-il que /bu/, /du/ et /gu/ ont quelque chose en commun, qui n'est pas présent dans /ba/ ? Comment apprend-il que /ba/, /bi/ et /bu/ ont quelque chose en commun, qui n'est pas présent dans /ga/ ? Dans cette section nous étudions le rôle que peuvent jouer les représentations perceptives et motrices dans la construction des catégories abstraites de phonèmes que sont la voyelle et la consonne.

3.1.1 Phonologie à partir de représentations auditives ?

Dans le modèle *COSMO-S*, les variables S_C et S_V décrivent les représentations internes des caractéristiques acoustiques de la syllabe au début (partie consonne) et à la fin (partie voyelle) du signal. Chercher à savoir si un système auditif est suffisant pour voir émerger une ébauche de phonologie au sens que nous avons défini précédemment revient à regarder si les représentations des syllabes dans les espaces S_V et S_C contiennent suffisamment d'information pour caractériser les syllabes (dans nos travaux il s'agit de /ba-bi-bu-da-di-du-ga-gi-gu/) et pouvoir les regrouper dans des méta-catégories où elles sont caractérisées par leur voyelle (/a/, /i/ ou /u/ dans notre cas) ou leur consonne (/b/, /d/, ou /g/).

Nous proposons donc de regarder, avec les figures 6.11 et 6.12 comment les prototypes auditifs de syllabes $P\left(S'_V\ S'_C \mid [O_L{=}o]\ \pi_{Ag}\right)$ appris par l'agent pour chacun des objets o se projettent dans les espaces S_V et S_C.

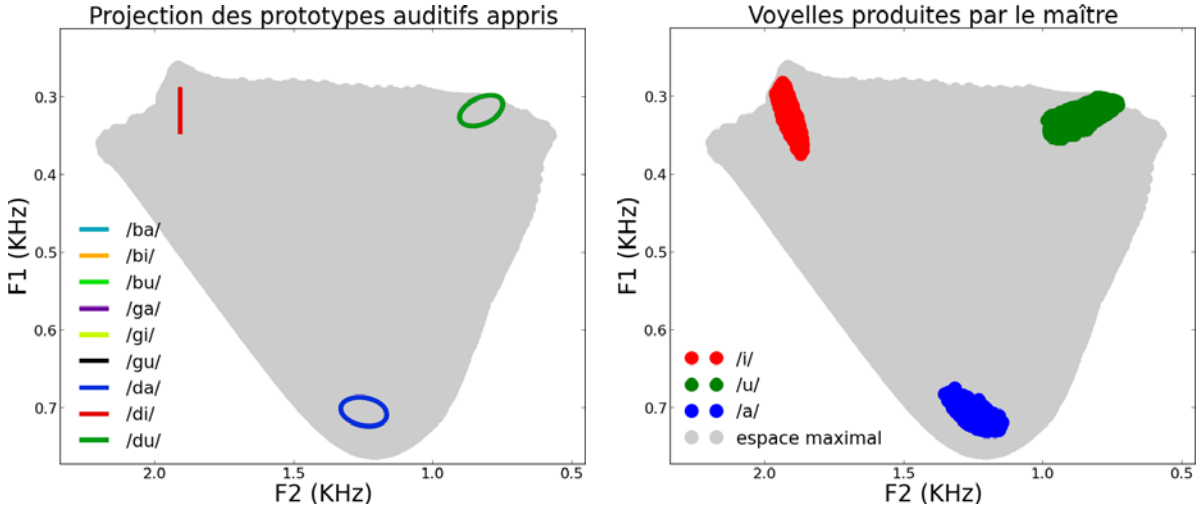


Figure 6.11: **Représentations auditives associées aux syllabes dans l'espace des voyelles** après l'apprentissage, et comparaisons avec les syllabes produites par le maître. La figure de droite montre comment les syllabes produites par le maître au cours de l'apprentissage se projettent dans l'espace acoustique $(F2_V, F1_V)$ des voyelles. La figure de gauche représente une visualisation de la projection des distributions de probabilité gaussiennes $P\left(S'_V\ S'_C \mid [O_L{=}o]\ \pi_{Ag}\right)$ à quatre dimensions dans l'espace acoustique des voyelles. Chaque prototype auditif de syllabe est représenté par une ellipse qui dessine le contour de la zone représentant une masse de probabilité supérieure à 86.5%.

La figure 6.11 montre que lorsque l'on projette les prototypes auditifs de syllabes appris dans l'espace $(F2_V, F1_V)$ des voyelles, on obtient des ellipses qui ont l'air parfaitement superposées pour les syllabes partageant la même voyelle (on a d'ailleurs l'impression de n'y voir que trois couleurs au lieu de neuf). Cela s'explique par le fait que dans les données d'apprentissage des prototypes auditifs, les plosives ont été générées à partir des mêmes ensembles de voyelles. De petites différences numériques existent tout de même au niveau des moyennes et des matrices de covariance des gaussiennes apprises, qui sont bien trop petites pour être visibles sur la figure 6.11, et qui sont dues à des effets de tirages. En revanche, la figure 6.12 montre que les prototypes auditifs de syllabes appris se projettent dans l'espace $(F2_C, F3_C)$ des consonnes selon neuf ellipses bien distinctes.

Finalement, les représentations dans l'espace des voyelles étant quasiment identiques pour trois groupes de syllabes (/ba-da-ga/, /bi-di-gi/ et /bu-du-gu/) notre agent cognitif a donc appris un invariant perceptif lui permettant de caractériser la partie voyelle des syllabes. Des représentations auditives semblent donc suffisantes pour faire émerger la notion de catégorie vocalique. En revanche, dans l'espace acoustique des consonnes, il semble ne rien y avoir qui puisse suggérer un invariant permettant à la notion de catégorie consonantique d'émerger de manière claire.

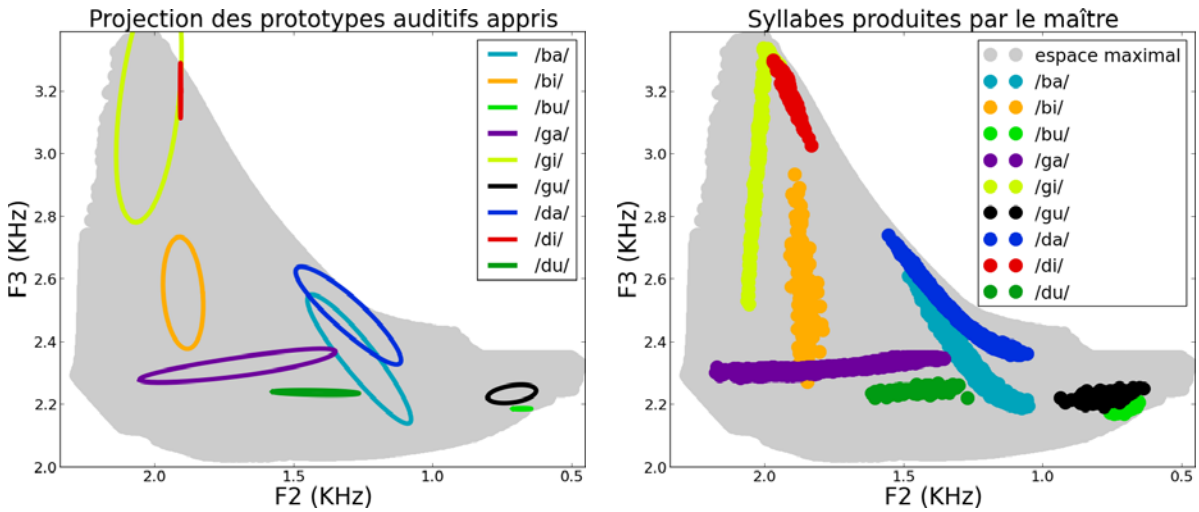


Figure 6.12: **Représentations auditives associées aux syllabes dans l'espace des consonnes** après l'apprentissage, et comparaisons avec les syllabes produites par le maître. La figure de droite montre comment les syllabes produites par le maître au cours de l'apprentissage se projettent dans l'espace acoustique $(F2_C, F3_C)$ des consonnes. La figure de gauche représente une visualisation de la projection des distributions de probabilité gaussiennes $P\,(S'_V\ S'_C \mid [O_L{=}o]\ \pi_{Ag})$ à quatre dimensions dans l'espace acoustique des consonnes. Chaque prototype auditif de syllabe est représenté par une ellipse qui dessine le contour de la zone représentant une masse de probabilité supérieure à 86.5%.

3.1.2 Phonologie à partir de représentations motrices !

Dans la structure de dépendances probabilistes du modèle *COSMO-S*, apparaît côté moteur la variable G'_C qui permet d'exprimer le fait que l'agent, doté d'un vocabulaire d'actions, apprend à associer un type de geste de constriction à chaque syllabe /ba/, /bi/, /bu/, /da/, /di/, /du/, /ga/, /gi/, /gu/. En effet, le mécanisme de renforcement par le biais duquel à une étape de l'apprentissage l'agent choisit avec une probabilité supérieure les gestes qu'il a déjà associés aux objets aux étapes précédentes a également pour effet d'ancrer le choix d'un seul type de geste pour produire chacune des syllabes.

La figure 6.13 montre l'évolution de la proportion de chaque type de geste choisi au cours de l'apprentissage de compétences motrices par imitation.

Initialement, chaque geste est choisi avec des proportions égales, ce qui est dû au fait que, l'agent ne disposant encore d'aucune connaissance, ses représentations internes sont des distributions de probabilité uniformes. Ensuite, le mécanisme de renforcement sur lequel repose notre algorithme d'apprentissage fait que l'agent va associer de plus en plus fortement le choix d'un type de geste privilégié à chaque syllabe. À la fin de l'apprentissage, l'agent a choisi d'associer un unique geste de constriction à chaque syllabe, sauf pour le cas de la syllabe /gi/ à laquelle il associe de manière stable un geste de constriction vélaire avec une probabilité de 70% ou un geste de constriction dentale avec une probabilité de 30%. Notons que ce type de confusion est effectivement plausible et attesté à la fois dans les patterns de confusion chez les locuteurs adultes (voir par exemple l'article de référence de Cooper *et al.* (1952) qui montre que dans la zone du /i/ la perception du lieu d'articulation vélaire tend à s'effacer au profit de la dentale) et dans les patterns développementaux, avec classiquement des confusions entre /di/ et /gi/

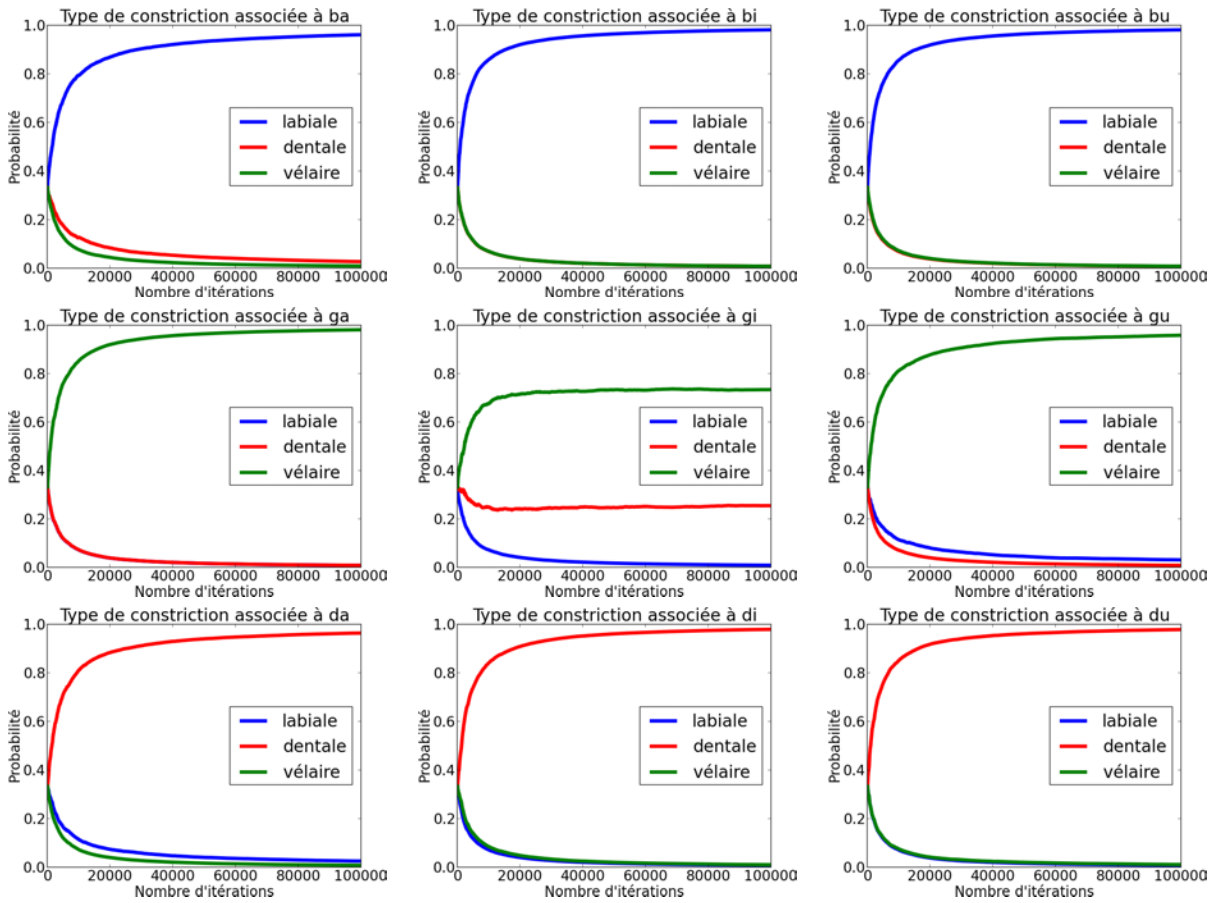


Figure 6.13: **Apprentissage du lieu d'articulation des plosives.** Ces neuf figures présente l'évolution au cours de l'apprentissage de $P\left(G'_C \mid [O_S{=}o]\ \pi_{Ag}\right)$ pour chacune des neuf syllabes o (/ba-bi-bu-ga-gi-gu-da-di-du/). Sur une même figure, chaque courbe représente l'évolution de la proportion selon laquelle l'agent associe chaque type de geste à la syllabe considérée. Ainsi, si l'on considère la valeur de chaque courbe pour une même itération, on lit verticalement une distribution de probabilité $P\left(G'_C \mid [O_S{=}o]\ \pi_{Ag}\right)$.

dans les productions du babillage vers la fin de la première année (voir une revue dans Lalevée (2010)).

En résumé, l'apprentissage de compétences motrices avec notre mécanisme de renforcement permet progressivement à l'invariance du lieu d'articulation des plosives d'émerger, en tous cas pour huit des neuf syllabes considérées : le cas de la syllabe /gi/ est moins tranché.

3.1.3 Le rôle de la langue de la mère (*motherese*, mamanais) dans l'émergence de la phonologie

Revenons dans cette section sur le cas apparent d'un « défaut d'apprentissage » de notre modèle moteur, avec des syllabes /gi/ produites par le maître avec le geste vélaire adéquat, et l'inférence motrice par l'agent d'une double possibilité de production, soit vélaire soit dentale. Comment peut-on imaginer que soit récupérée l'invariance phonologique dans ce cas précis ? Il faut d'abord noter qu'un processus de généralisation appliqué sur les gestes inférés pour /gi/ pourrait conduire (par exemple par un algorithme de type « *winner take all* ») à associer *in fine* un geste

vélaire unique à tous les exemplaires de la catégorie. Néanmoins, ce processus semble ici fragile, et appelle d'autres éléments de réflexion.

Un premier élément est que la description acoustique des syllabes dans notre modèle est très simplifiée. Rappelons que nous décrivons une syllabe simplement par les deux premiers formants de la voyelle, et les second et troisième formants de la consonne. Ceci fournit un certain nombre de recouvrements observés figure 6.12 (notamment /bu/-/gu/, /di/-/gi/ et /ba/-/da/). On sait en réalité que les indices acoustiques caractérisant les plosives sont multiples, incluant des caractéristiques instantanées sur le bruit d'explosion (*burst* : voir Blumstein *et al.* (1977) ; Stevens (1980)) mais aussi les trajectoires allant du *burst* à la voyelle qui suit (Kewley-Port, 1983) et bien évidemment toute la trajectoire formantique (Sussman *et al.*, 1991), ce qui implique des connaissances également sur le troisième formant de la voyelle, et éventuellement sur le quatrième formant. Avec plus d'information disponible, nous aurions probablement moins d'ambiguïté dans la récupération d'un geste adéquat.

Il faut noter ensuite que lorsque l'information disponible est réduite, on peut voir apparaître effectivement des cas d'apprentissage imparfait. On connaît ainsi le célèbre exemple, rapporté par Mills (1987), des enfants aveugles qui peinent à associer à /m/ ou /n/ le geste articulatoire adéquat, ce qui est interprété par l'auteur comme le fait que la distinction acoustique entre ces deux consonnes nasales respectivement labiale et dentale est réduite. Cette information pauvre est efficacement complétée par l'information visuelle (les deux consonnes sont très bien distinguables visuellement), ce qui permet aux enfants bien voyants d'apprendre le contraste rapidement. Ceci nous conduit au point central de cette section. Une autre manière dont l'ambiguïté peut être levée passe par les actions du maître : le maitre peut choisir d'enrichir l'information qu'il fournit, en tentant de produire des stimuli mieux contrastés. C'est probablement ce qui survient dans la situation de production dite de « *motherese* » (ou « mamanais »), cette version particulière de parole que, typiquement, la maman (les parents) produit pour son bébé. Le mamanais est caractérisé par une intonation appuyée et une voix plus aigue, susceptibles de capter l'intérêt de l'enfant, mais aussi par une articulation extrême, une « hyperarticulation » susceptible de maximiser l'information produite (Fernald et Kuhl (1987) ; voir une présentation et une discussion de l'hypothèse d'hyperarticulation du motherese chez Cristia et Seidl (2013)).

Dans ce qui est présenté dans cette section, nous avons choisi d'implémenter un mécanisme de *motherese* sous forme de surarticulation : l'agent maître cherche à éviter les ambiguïtés en retirant de ses productions les zones de recouvrement. En effet, la figure 6.12 montre que la production de syllabes différentes peut conduire à des valeurs identiques des second et troisième formants : on observe en particulier des ambiguïtés /ba/-/da/, /di/-/gi/ et /bu/-/gu/. En pratique ce principe de surarticulation que nous proposons se traduit, dans l'implémentation de l'algorithme d'apprentissage par imitation des compétences motrices, par le fait que si le maître produit une syllabe qui, au vu de notre discrétisation, tombe dans une zone de l'espace acoustique qui correspond à deux syllabes différentes, alors le maître rejette cette syllabe et fait un autre tirage aléatoire de syllabe à communiquer à l'agent apprenant.

La figure 6.14 montre que, si au cours de son apprentissage l'agent n'est soumis qu'à ce type de stimuli plus simple, il converge vers le bon choix de geste de constriction également dans le cas du /gi/.

Dans un souci d'économie de place, la figure 6.14 ne présente que le cas de la syllabe /gi/, les autres étant par ailleurs identiques à ce qui est présenté à la figure 6.13. Finalement, notre

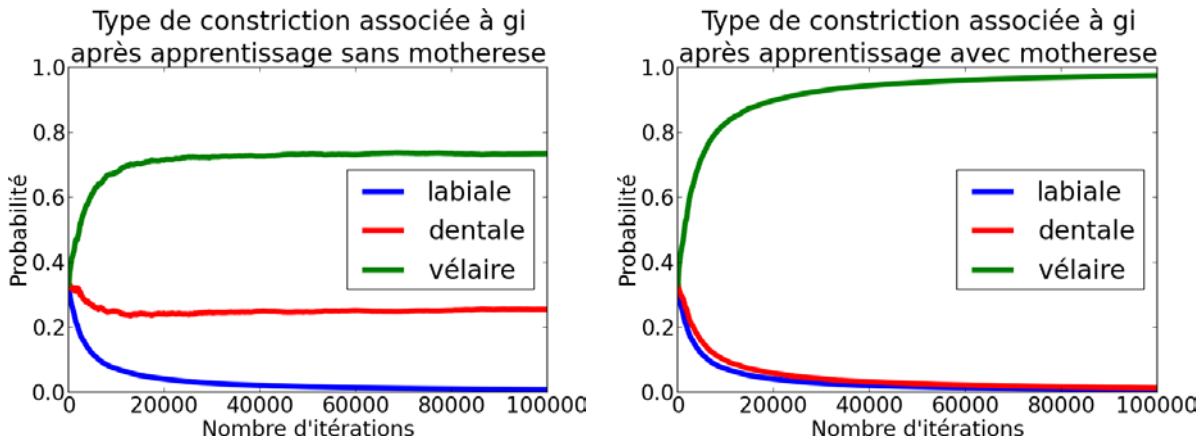


Figure 6.14: **Le rôle du *motherese* dans l'émergence de la phonologie.** Ces deux figures présentent l'évolution du terme $P\left(G'_C \mid [O_S{=}/gi/]\ \pi_{Ag}\right)$ au cours de l'apprentissage, à gauche dans le cas où le maître produit les syllabes normalement, à droite dans le cas où le maître désambiguïse les syllabes du point de vue acoustique en hyperarticulant.

implémentation d'un mécanisme de *motherese* permet à l'agent, en réduisant les ambiguïtés lors de son apprentissage de compétences motrices, de voir émerger plus facilement l'invariance du lieu d'articulation des plosives.

3.2 L'utilisation de connaissances motrices apporte de la robustesse pour la perception de syllabes en conditions dégradées

Une fois que l'agent a appris son système auditif, ses modèles internes et ses répertoires moteurs à partir des productions du maître de la manière décrite à la section 2, nous proposons de comparer les performances prévues par le modèle *COSMO-S* des instances auditive, motrice, et perceptuo-motrice de la tâche de perception de syllabes. Dans un premier temps, nous donnons les résultats des calculs d'inférence pour les instances motrice, auditive et perceptuo-motrice de la tâche de perception. Dans un second temps, nous présentons notre paradigme d'évaluation comparative des performances des modèles, pour, dans un troisième temps, présenter les résultats ainsi obtenus et les analyser.

3.2.1 Inférences probabilistes pour les tâches de perception au sein du modèle *COSMO-S*

La tâche de perception de syllabe consiste pour l'agent à essayer de retrouver la catégorie o d'un percept (s_v, s_c) qu'il reçoit du maître. De la même manière que ce que nous avons présenté à la section 3.3 du chapitre 3, cette tâche de perception s'instancie de trois manières différentes dans le cadre des théories motrice, auditive et perceptuo-motrice. Le calcul des distributions de probabilité correspondantes a la même forme que ce qui a été présenté à la section 4.2.2 du chapitre 4, à ceci près que la structure du modèle *COSMO-S*, un peu plus complexe, fait intervenir trois blocs reliés par les variables de cohérence λ_{SV}, λ_{SC}, λ_{MV} et λ_{MC} : le système moteur, le système auditif, et le système sensori-moteur. Dans les calculs d'inférence que nous allons présenter ici, ces variables de cohérence permettent de choisir quels blocs activer au sein

du modèle complet.

Notre implémentation d'une théorie motrice au sein du modèle *COSMO* consiste à se focaliser sur les objets de la communication O_S considérés du point de vue du locuteur (speaker). Au sein du modèle *COSMO-S*, qui est une instanciation du modèle générique *COSMO* adaptée aux syllabes, effectuer une tâche de perception dans le cadre d'une théorie motrice revient à calculer la distribution de probabilité conditionnelle $P\,(O_S \mid S_V\ S_C\ [\lambda_{MV}{=}1]\ [\lambda_{MC}{=}1])$[39]. Par inférence probabiliste sur la distribution de probablité conjointe du modèle *COSMO-S*, on obtient le résultat suivant :

$$P\,(O_S \mid S_V\ S_C\ [\lambda_{MV}{=}1]\ [\lambda_{MC}{=}1]) \propto \sum_{m'_v}\left(P(M'_V \mid O_S)\sum_{g'}\left(P(G'_C \mid O_S)\sum_{\delta'} P(\Delta'_{MC} \mid M'_V\, G'_C)\ P(S_V \mid [M_V{=}m'_v])\ P(S_C \mid [M_C{=}m'_v{+}\delta'])\right)\right). \tag{6.11}$$

Cette équation montre comment au sein de notre modèle *COSMO-S* la tâche de perception selon une théorie motrice est réalisée de manière identique à ce que nous avons présenté à l'équation 3.20 du chapitre 3, en combinant les modèles internes du système sensori-moteur avec les répertoires moteurs du système moteur de l'agent. Cette tâche de perception suit le principe d'analyse par la synthèse : tous les gestes moteurs susceptibles d'avoir produit le stimulus (S_V, S_C) sont considérés grâce aux sommes sur les variables M'_V, G'_C et Δ'_{MC}. Chacun de ces gestes contribue d'autant plus à rendre plausible la perception de l'objet O_S qu'ils a une probabilité élevée de correspondre au stimulus (s_v, s_c) d'une part, et à l'objet O_S d'autre part.

Notre implémentation d'une théorie auditive de la perception au sein du modèle *COSMO* consiste à se focaliser sur les objets de la communication O_L considérés du point de vue de l'auditeur (listener). Au sein du modèle *COSMO-S*, effectuer une tâche de perception dans le cadre d'une théorie auditive revient à calculer la distribution de probabilité conditionnelle $P\,(O_L \mid S_V\ S_C\ [\lambda_{SV}{=}1]\ [\lambda_{SC}{=}1])$. Par inférence probabiliste sur la distribution de probabilité conjointe du modèle *COSMO-S*, on obtient le résultat suivant :

$$P\,(O_L \mid S_V\ S_C\ [\lambda_{SV}{=}1]\ [\lambda_{SC}{=}1]) \ \propto\ P(S_V\ S_C \mid O_L)\ . \tag{6.12}$$

Cette équation montre comment au sein de notre modèle *COSMO-S* la tâche de perception selon une théorie auditive est réalisée de manière identique à ce que nous avons présenté à l'équation 3.28 du chapitre 3, en sollicitant directement le système auditif de l'agent. Plus précisément, le stimulus (S_V, S_C) a d'autant plus de chances d'être perçu comme objet O_L qu'il est proche dans l'espace acoustique du prototype gaussien $P(S_V\ S_C \mid O_L)$ que l'agent a appris à associer à cet objet.

Notre implémentation d'une théorie perceptuo-motrice de la perception au sein du modèle *COSMO* consiste à se focaliser à la fois sur O_S et O_L, à considérer simultanément le point

[39] Par souci de lisibilité, et vue l'absence d'ambiguïté, nous ommettons dans ce qui va suivre les π_{Ag} en partie droite des distributions de probabilité : tous les calculs se font au sein du modèle de l'agent.

de vue du locuteur et celui de l'auditeur. Au sein du modèle *COSMO-S*, effectuer une tâche de perception dans le cadre d'une théorie perceptuo-motrice revient à calculer la distribution de probabilité conditionnelle $P\left(O_L \mid S_V\ S_C\ [C{=}1]\ [\lambda_{MV}{=}1]\ [\lambda_{MC}{=}1]\ [\lambda_{SV}{=}1]\ [\lambda_{SC}{=}1]\right)$, ou encore la distribution $P\left(O_S \mid S_V\ S_C\ [C{=}1]\ [\lambda_{MV}{=}1]\ [\lambda_{MC}{=}1]\ [\lambda_{SV}{=}1]\ [\lambda_{SC}{=}1]\right)$, ce qui est équivalent. Par inférence probabiliste sur la distribution de probabilité conjointe du modèle *COSMO-S*, on obtient le résultat suivant :

$$P\left(O_L \mid S_V\ S_C\ [C{=}1]\ [\lambda_{MV}{=}1]\ [\lambda_{MC}{=}1]\ [\lambda_{SV}{=}1]\ [\lambda_{SC}{=}1]\right) \propto$$
$$P(S_V\ S_C \mid O_L) \sum_{m'_v} \left(P(M'_V \mid O_S) \sum_{g'} \left(P(G'_C \mid O_S) \sum_{\delta'} P(\Delta'_{MC} \mid M'_V\ G'_C)\ P(S_V \mid [M_V{=}m'_v])\ P(S_C \mid [M_C{=}m'_v{+}\delta']) \right) \right). \tag{6.13}$$

Cette équation montre comment au sein de notre modèle *COSMO-S* la tâche de perception selon une théorie perceptuo-motrice est réalisée de manière identique à ce que nous avons présenté à l'équation 3.35 du chapitre 3, en opérant la fusion bayésienne des réponses des théories motrice et auditive. Ce calcul de la tâche de perception dans le cadre d'une théorie perceptuo-motrice montre comment l'utilisation des variables de cohérence permet de combiner l'information présente dans chacun des trois blocs de notre modèle : le système auditif, le système moteur et le système sensori-moteur.

3.2.2 Méthode de comparaison

Prises ensemble, les équations 6.11, 6.12 et 6.13 permettent de précalculer une fois pour toutes les distributions de probabilité sur les objets reconnus $P(O \mid S_V\ S_C)$ correspondant à tous les stimuli $(S_V\ S_C)$ possibles, pour chacun des trois modèles : moteur, auditif et perceptuo-moteur. Le principe des comparaisons que nous faisons entre ces trois modèles est de faire des tests de perception sous différents niveaux de bruit, et d'analyser l'évolution d'un score de performance global en fonction de ce niveau de bruit.

Pour les besoins de l'évaluation, nous utilisons les données qui ont servi à l'apprentissage, qui sont caractérisées par les distributions de probabilité $P(S_V\ S_C \mid O^{Maître})$ décrites à la section2.2, mais que l'on dégrade en ajoutant un certain niveau de bruit blanc, que l'on implémente en perturbant indépendemment chacun des formants de la syllabe en tirant des perturbations selon une distribution de probabilité gaussienne centrée sur 0 et dont nous allons faire varier l'écart-type.

Le calcul du score global que nous avons choisi pour comparer les performances des différents modèles est similaire à ce qui a été présenté à la section 4.2.3 du chapitre 4 : il s'agit de calculer une matrice de confusion comptabilisant les réponses du modèle considéré en fonction des stimuli fournis par le maître. La seule différence avec ce qui est fait au chapitre 4 tient au fait qu'il y a maintenant neuf syllabes à considérer au lieu des deux objets abstraits considérés précédemment. La figure 6.15 montre la matrice de confusion obtenue par le modèle moteur en l'absence de bruit de communication.

La figure 6.15 présente les résultats d'une évaluation qui est faite sans perturbation, et qui repose donc sur les mêmes données qui ont servi à l'apprentissage. Cette matrice de confusion et les valeurs élevées de probabilités qu'elle contient sur la diagonale permettent de valider nos

	/ba/	/bi/	/bu/	/ga/	/gi/	/gu/	/da/	/di/	/du/
/ba/	**0.82**	3.9e-99	3.9e-99	0.01	3.9e-99	3.9e-99	**0.16**	3.9e-99	3.9e-99
/bi/	3.3e-15	**0.99**	6.3e-19	2.3e-12	**0.005**	1.9e-16	7.1e-15	0.004	1.8e-18
/bu/	1.2e-13	1.8e-13	**0.94**	1.6e-10	1.6e-11	**0.049**	1.7e-12	7.2e-14	0.01
/ga/	0.009	3.5e-99	3.5e-99	**0.98**	3.5e-99	3.5e-99	**0.011**	3.5e-99	3.5e-99
/gi/	1.8e-15	0.013	7.4e-99	2.2e-12	**0.87**	1.8e-16	6.0e-15	**0.116**	7.4e-99
/gu/	9.7e-14	5.3e-13	**0.07**	1.1e-10	3.1e-11	**0.91**	1.2e-12	1.9e-13	0.02
/da/	**0.166**	6.7e-99	6.7e-99	0.006	6.7e-99	6.7e-99	**0.828**	6.7e-99	6.7e-99
/di/	4.4e-16	0.01	2.2e-99	6.8e-13	**0.12**	5.5e-17	1.6e-15	**0.87**	2.1e-99
/du/	1.7e-12	2.7e-12	0.003	1.4e-09	1.7e-10	**0.012**	6.1e-12	9.9e-13	**0.985**

Figure 6.15: **Matrice de confusions** résumant les performances du modèle moteur évalué sur une tâche de perception de syllabes dans les mêmes conditions que celles de l'apprentissage. Chaque ligne correspond, pour une catégorie de syllabe communiquée par le maître, à la distribution de probabilité moyenne sur les syllabes reconnues par l'agent. Sur la diagonale figurent en vert les probabilités de bonne réponse. Pour chaque syllabe, la probabilité de l'erreur majoritaire de classification est mise en évidence avec la couleur rouge.

algorithme d'apprentissage des modèles internes et de répertoires moteurs : bien que la transformation articulatori-acoustique n'ait pas été apprise exhaustivement, suffisamment d'information a été apprise pour assurer de bonnes performances en perception motrice. On retrouve également dans cette matrice les confusions classiques entre les syllabes /ba/-/da/, /bu/-/gu/, et /di/-/gi/.

À partir de ces matrices de confusion, on définit un indicateur global : le taux de performance, qui se calcule comme étant la moyenne des coefficients diagonaux de la matrice de confusions. Ce taux de performance décrit la précision en reconnaissance du modèle considéré, moyennée sur les différentes catégories de syllabes. À partir des données de la figure 6.15 on peut donc calculer le taux de performances du modèle moteur sur son corpus d'apprentissage, qui est de 91%.

Sachant que le corpus d'évaluation est ambigu (les recouvrements visibles sur la partie droite de la figure 6.12 font que le meilleur score théorique est en dessous des 100%), obtenir un taux de performances à 91% est tout à fait satisfaisant, et permet de valider l'algorithme d'apprentissage par accomodation que nous proposons. L'agent obtient donc un bon score de perception motrice, alors même que le modèle interne qu'il a appris de la transformation articulatori-acoustique est très incomplet, comme le montre la figure 6.4, sur laquelle de nombreuses distributions de probabilité ont une valeur de ΔH supérieure à 0.

3.2.3 Comparaisons des performances des modèles moteur, auditif, et sensori-moteur lors de tests de perception de syllabes dans le bruit

La figure 6.16 montre l'évolution du taux de performance (tel que l'on vient de le définir) pour chacun des différents modèles lorsque l'on fait varier le bruit de l'environnement.

Le modèle auditif montre de meilleures performances que le modèle moteur en conditions normales, mais le modèle moteur prend l'avantage en conditions dégradées. Le modèle perceptuo-moteur quant à lui tire avantageusement parti de la fusion de l'information apportée par les

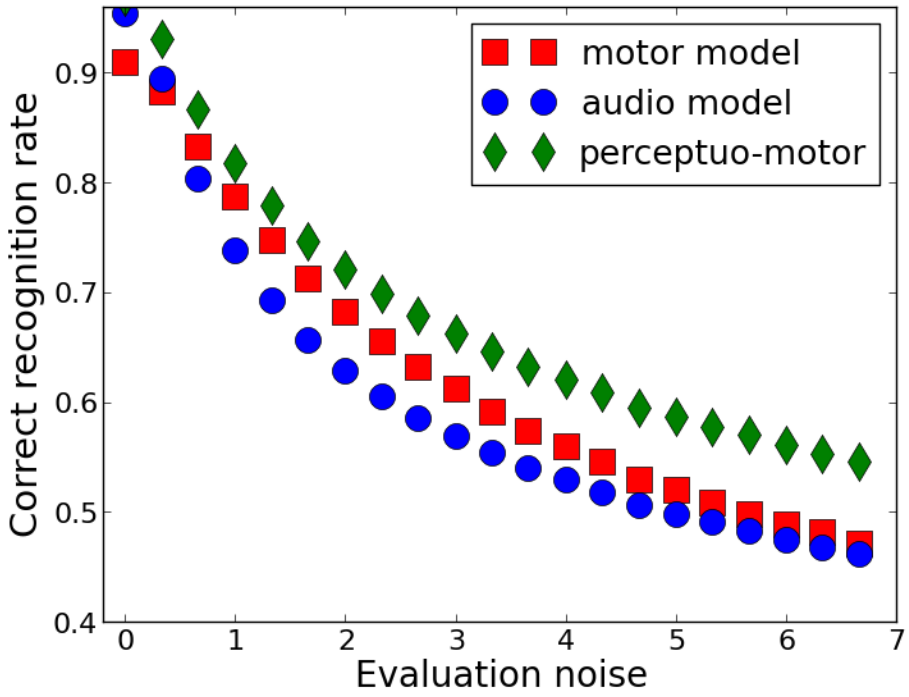


Figure 6.16: **Robustesse au bruit des différents modèles évalués sur des tâches de perception de syllabes**. Le taux de reconnaissance des modèles (en ordonnée) diminue lorsque l'écart-type des gaussiennes (en abscisse) contrôlant le bruit blanc augmente, c'est à dire lorsque le niveau de dégradation augmente. Cette figure est reprise de Laurent *et al.* (2013b).

deux autres et montre de meilleures performances partout.

Ce chapitre généralise donc au cas des syllabes les résultats déjà présentés au chapitre 4 à la section 4.3. L'interprétation que nous proposons de ces résultats se base sur les observations suivantes : alors que le système auditif de l'agent apprend vite et se spécialise rapidement sur les données de son corpus d'apprentissage, le système sensori-moteur de l'agent, plus difficile à apprendre, ne capture qu'un modèle interne imparfait de la transformation articulatori-acoustique. Il est intéressant de remarquer que c'est le fait de ne disposer que d'information partielle qui permet au modèle moteur, moins spécialisé que le modèle auditif, de mieux généraliser ce qu'il a appris, et d'être plus robuste au bruit que ce dernier. Dès lors, le modèle perceptuo-moteur tire avantageusement parti du caractère spécialisé du modèle audio, et des capacités de généralisation du modèle moteur.

4 Conclusion

Dans ce chapitre, nous avons présenté le modèle *COSMO-S*, qui est une extension de *COSMO* permettant de traiter des syllabes de type plosive-voyelle. Nous avons implémenté dans ce cadre l'algorithme d'apprentissage par accomodation déjà introduit au chapitre 4, qui permet à l'agent d'acquérir des compétences motrices en procédant par imitation des stimuli du bain acoustique dans lequel il se trouve. Cet algorithme a plusieurs propriétés remarquables : il permet

un apprentissage moteur à partir d'entrées perceptives uniquement, il permet de concentrer l'apprentissage de la transformation articulatori-acoustique sur les zones de l'espace acoustique pertinentes pour l'agent (celles qui correspondent aux cibles que le maître lui propose), et il permet dans l'espace moteur d'ancrer des choix idiosyncrasiques, ce qui a pour effet pour l'agent de réduire la complexité (le caractère *many-to-one*) de la transformation apprise.

Nous avons montré que si la notion de catégorie vocalique est déjà contenue implicitement dans les connaissances auditives, ce n'est pas le cas de la notion de catégorie consonantique, qui au contraire semble émerger naturellement dans les connaissances motrices, ce qui est facilité lorsque lors de l'apprentissage le maître surarticule (*motherese*). Nous avons également montré une plus grande robustesse du modèle moteur en tâche de perception dans des conditions dégradées par rapport au modèle auditif, ce qui s'explique par les connaissances partielles qui sont accumulées autour des cibles lors de l'apprentissage par accomodation.

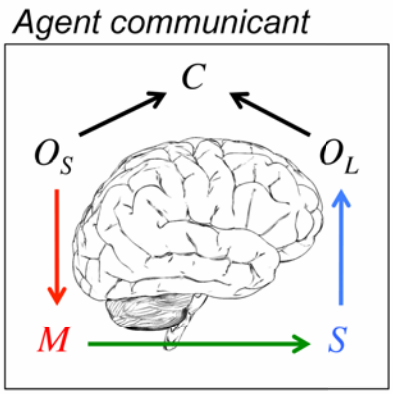
Finalement, alors qu'il ne fait intervenir que peu de variables et que ses formes paramétriques sont définies à partir d'objets mathématiques usuels, le modèle *COSMO-S* procure un cadre unique qui nous a permis dans ce chapitre de nous intéresser à de nombreuses questions : comparer les théories motrice, auditive et perceptuo-motrice, réaliser des tâches de production, de perception et d'imitation, étudier des dynamiques d'apprentissage, développer des idiosyncrasies, faire émerger une ébauche de phonologie, étudier la robustesse aux conditions dégradées.

Chapitre 7

Conclusion générale

1 Contributions principales

La première contribution de ce travail de thèse consiste en un travail de formalisation. L'analyse de la situation de communication a ainsi permis de proposer le modèle d'agent communicant *COSMO* (*Communicating Objects using SensoriMotor Operations*) que nous avons défini au chapitre 3 et que nous rappelons figure 7.1.



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

Figure 7.1: **Le modèle *COSMO*** d'agent communicant, décrit par un réseau bayésien et par la décomposition de sa distribution de probabilité conjointe.

Ce modèle mathématique défini dans un langage probabiliste selon la méthodologie de la Programmation Baysésienne permet, comme le montre la figure 7.2, de formaliser les théories motrice, auditive et perceptuo-motrice de la production et de la perception de la parole ainsi que les hypothèses sur lesquelles elles reposent.

Formaliser ainsi ces différentes théories au sein d'un même cadre intégrateur rend possible des comparaisons systématiques de ces théories qui malheureusement sont souvent étudiées séparément. Cela permet également de structurer de manière synthétique et cohérente des connaissances et hypothèses issues de différentes disciplines.

	Tâche de production inférence de la forme $P(M \mid O)$	Tâche de perception inférence de la forme $P(O \mid S)$
Théorie motrice focalisation sur O_S	$\underbrace{P(M \mid O_S)}_{\text{répertoire moteur}}$	$\propto \sum_M \Big(\underbrace{P(M \mid O_S)}_{\text{décodeur articulatoire}} \times \underbrace{P(S \mid M)}_{\text{modèle inverse}} \Big)$
Théorie auditive focalisation sur O_L	$\propto P(M) \sum_S \Big(\underbrace{P(S \mid M)}_{\text{modèle direct}} \times \underbrace{P(O_L \mid S)}_{\text{cibles acoustiques}} \Big)$	$\underbrace{P(O_L \mid S)}_{\text{classifieur auditif}}$
Théorie perceptuo-motrice C=1, i.e. O_S=O_L	$\propto \underbrace{P(M \mid [O_S{=}O_L])}_{\text{production motrice}} \underbrace{\sum_S \big(P(S \mid M)P(O_L \mid S)\big)}_{\text{production auditive}}$	$\propto \underbrace{P([O_L{=}O_S] \mid S)}_{\text{perception auditive}} \underbrace{\sum_M \big(P(M \mid O_S)P(S \mid M)\big)}_{\text{perception motrice}}$

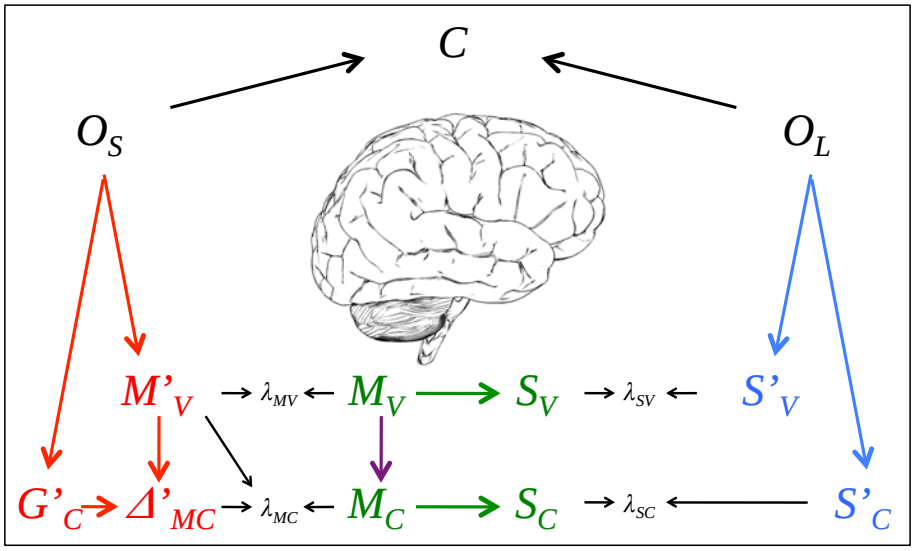
Figure 7.2: **Le modèle *COSMO* permet d'intégrer** dans un cadre mathématique unique les théories motrice, auditive, et perceptuo-motrice de la production et de la perception de la parole en en donnant une formalisation précise.

Notre deuxième contribution réside dans la démonstration de notre thérorème d'indistinguabilité qui est faite au chapitre 3. Ce théorème montre que les voies auditive et motrice de traitement de la parole d'un agent communicant peuvent encoder exactement la même information, les rendant alors indistinguables expérimentalement. En d'autres termes, sous certaines hypothèses, il est imposible de savoir par l'observation extérieure du processus si un agent communicant fait appel à des connaissances auditives ou motrices pour réaliser une tâche de perception de parole. La définition formelle des hypothèses qui conditionnent ce théorème permet également d'identifier des conditions qui rendraient ces théories distinguables expérimentalement.

Ceci nous amène à notre troisième contribution, que nous avons décrite au chapitre 4 : nous proposons un algorithme d'apprentissage permettant d'invalider les hypothèses d'apprentissage parfait sur lesquelles repose le théorème d'indistinguabilité. Nous avons ainsi implémenté l'apprentissage de la voie auditive par association, et celui de la voie motrice grâce à un algorithme original d'apprentissage par accomodation, le tout à partir des mêmes entrées sensorielles. L'algorithme d'apprentissage de compétences motrices par accomodation que nous proposons permet d'obtenir un agent communicant qui n'a pas une connaissance parfaite de la transformation articulatori-acoustique (ce qui serait irréaliste), qui développe des idiosyncrasies de production de parole, et qui pourtant s'adapte au bain acoustique ambiant.

Notre quatrième contribution réside dans la définition du modèle *COSMO-S* : une extension de *COSMO* au traitement des syllabes de type plosive-voyelle que nous avons présentée au chapitre 6. Pour pouvoir définir ce modèle, nous avons rendu explicites des principes de simplification justifiés par la littérature, et qui permettent de se focaliser sur certains aspects de la complexité des objets étudiés (voir le chapitre 5). Le modèle *COSMO-S* (que nous rappelons figure 7.3) est un modèle complet de production et de décodage des syllabes qui, comme le modèle générique *COSMO* qu'il instancie, permet d'implémenter et de comparer les tâches de production et de perception dans le cadre des théories auditive, motrice et perceptuo-motrice.

Les paramètres de ce modèle sont appris à partir de données de syllabes, qui sont générées avec un simulateur de conduit vocal (le modèle *VLAM*) de la manière décrite au chapitre 5, et



$$
\begin{aligned}
&P(O_S\ G'_C\ M'_V\ \Delta'_{MC}\ \lambda_{MV}\ \lambda_{MC}\ M_V\ M_C\ S_V\ S_C\ \lambda_{SV}\ \lambda_{SC}\ S'_V\ S'_C\ O_L\ C)\\
&\quad = P(O_S) \times P(M'_V \mid O_S) \times P(G'_C \mid O_S) \times P(\Delta'_{MC} \mid M'_V\ G'_C) \times\\
&\qquad P(\lambda_{MV} \mid M'_V\ M_V) \times P(\lambda_{MC} \mid M'_V\ \Delta'_{MC}\ M_C) \times\\
&\qquad P(M_V) \times P(S_V \mid M_V) \times P(M_C \mid M_V) \times P(S_C \mid M_C) \times\\
&\qquad P(\lambda_{SV} \mid S_V\ S'_V) \times P(\lambda_{SC} \mid S_C\ S'_C) \times\\
&\qquad P(O_L) \times P(S'_V\ S'_C \mid O_L) \times\\
&\qquad P(C \mid O_S\ O_L)\ .
\end{aligned}
$$

Figure 7.3: **Le modèle *COSMO-S* de traitement des syllabes**, décrit schématiquement par un modèle graphique et par sa distribution de probabilité conjointe.

qui montrent des patrons de variabilité compatibles avec les données réelles. Il s'agit alors de capturer dans ces syllabes à la fois les invariants pour pouvoir ensuite les classifier correctement, et la variabilité pour pouvoir généraliser au-delà du corpus d'apprentissage. Là encore, notre algorithme d'apprentissage par accomodation permet à l'agent de développer des idiosyncrasies motrices de production, et de se construire une représentation de la transformation articulatori-acoustique adaptée au bain acoustique ambiant.

Notre cinquième contribution réside dans l'analyse des simulations informatiques que nous avons réalisées de manière complémentaire dans un cadre théorique abstrait au chapitre 4 et dans le cas des syllabes au chapitre 6. Les observations résultant de nos simulations peuvent être regroupées selon trois axes :

1. Une étude de l'émergence de la phonologie pendant l'apprentissage des syllabes (chapitre 6). Nos algorithmes d'apprentissage montrent des capacités d'apprentissage des notions de catégorie vocalique et consonantique à partir de présentations de stimuli globaux « non découpés » (des syllabes consonne-voyelle). Plus précisément, nous montrons que la notion de catégorie vocalique est implicitement contenue dans les connaissances auditives,

mais pas celle de catégorie consonantique qui apparaît en revanche dans les connaissances motrices.

(2) Une étude de la vitesse de convergence des apprentissages des voies auditive et motrice (chapitres 4 et 6). Nous montrons que les connaissances motrices mènent à des choix idiosyncrasiques de production et sont plus lentes à se mettre en place que les connaissances auditives. Cette dynamique d'évolution observée lors de l'apprentissage permet de faire un parallèle avec le cas de l'adaptation en ligne (voir le chapitre 4). Le modèle auditif serait alors plus facilement évolutif alors que le modèle moteur permettrait de conserver de la stabilité.

(3) Une étude quantitative de la perte de performance des traitements auditifs, moteurs et perceptuo-moteurs dans la perception en conditions bruitées (chapitres 4 et 6). Nous montrons notamment que des connaissances motrices apportent de la robustesse pendant la perception en conditions dégradées, alors que le modèle auditif est plus performant en conditions normales. Cette robustesse peut s'expliquer de deux manières : d'une part le modèle moteur encode naturellement plus de variabilité à cause de la difficulté d'apprendre une transformation articulatori-acoustique complexe, et d'autre part, lors de l'apprentissage par accomodation, le modèle moteur se construit également à partir de connaissances partielles autour des objets du bain acoustique ambiant (alors que cette information est absente du modèle auditif). Notre modèle de fusion perceptuo-motrice obtient quant à lui de meilleures performances que les modèles moteur et auditif à la fois en conditions normales et en conditions dégradées.

Finalement, notre modèle *COSMO* procure un cadre unique qui nous a permis dans cette thèse de nous intéresser à de nombreuses questions : comparer les théories motrice, auditive et perceptuo-motrice, réaliser des tâches de production, de perception et d'imitation, étudier des dynamiques d'apprentissage, développer des idiosyncrasies, faire émerger une ébauche de phonologie, étudier la robustesse aux conditions dégradées.

2 Discussion et perspectives

En rédigeant ce document, nous avons fait le choix, dans les chapitres de contribution 3, 4, 5 et 6, de nous en tenir à la présentation et à l'analyse des principaux résultats. Certains de ces résultats méritent d'être commentés davantage, et peuvent servir de point de départ pour prolonger la réflexion, et proposer des perspectives de recherches futures. C'est l'objectif de la présente section, dans laquelle nos réflexions sont regroupées de manière thématique.

2.1 *COSMO* : un modèle de l'agent communicant, un modèle de la communication ou un modèle d'internalisation ?

Une propriété remarquable de *COSMO* est que la distribution de probabilité conjointe définissant ce modèle peut conduire à des interprétations très différentes en fonction du sens que l'on donne à chacun de ses termes. Nous en décrivons trois.

(1) **Un modèle subjectif de l'agent en situation de communication** qui est résumé par la figure 7.1. Le terme $P(M \mid O_S)$ encode alors la connaissance qu'a l'agent sur les gestes associés aux objets. Le terme $P(S \mid M)$ s'interprète comme un modèle interne qui encode la connaissance (éventuellement incomplète ou partiellement erronée) que l'agent a de son environnement. Le terme $P(S \mid O_L)$ encode la connaissance qu'a l'agent sur les variables perceptives associées aux objets. Le terme $P(C \mid O_S\ O_L)$ implémente un mécanisme de conversion entre les codes phonologiques qui sont différents en production et en perception (Jacquemot *et al.*, 2007).

C'est cette vision que nous adoptons principalement dans cette thèse. La figure 7.2 présente de manière synthétique les équations correspondant aux calculs d'inférence probabiliste permettant de définir rigoureusement l'implémentation de chaque tâche (production ou perception) dans chaque cadre théorique (théories motrice, auditive et perceptuo-motrice). Cette figure (et en particulier les qualificatifs associés à chacun des termes : répertoire moteur, décodeur articulatoire, cibles acoustiques, classifieur auditif, etc) montre que, bien que les connaissances de l'agent encodées dans chaque terme de la distribution de probabilité conjointe restent les mêmes, l'interprétation de chacun de ces termes varie selon la tâche et le cadre théorique considérés. C'est une propriété remarquable des modèles bayésiens que de permettre de bien distinguer la manière dont les connaissances sont encodées dans un modèle de la manière d'utiliser ce modèle par la suite. Cela permet en particulier d'apprendre un modèle direct, et de l'utiliser ensuite, par inférence probabiliste, comme un modèle inverse.

(2) **Un modèle objectif de la situation de communication** qui est présentée de manière conceptuelle à la figure 3.1 et de manière formelle à la figure 3.2. Le terme $P(S \mid M)$ s'interprête alors comme un modèle de l'environnement. Le terme $P(M \mid O_S)$ s'interprète comme un modèle du processus de production du locuteur, vu de l'extérieur, indépendamment de la manière dont il est réalisé. De même, le terme $P(O_L \mid S)$ s'interprète comme un modèle du processus de perception de l'auditeur, et peut également être réalisé de différentes manières. Le terme $P(C \mid O_S\ O_L)$ implémente une mesure du succès de la communication.

C'est cette vision que nous adoptons quand nous décrivons les interactions entre l'agent et l'agent maître, lors de l'apprentissage ou lors de l'évaluation. Par exemple, que ce soit à la section 2.2 du chapitre 6, ou à la section 4.1 du chapitre 3, le modèle du maître est simplement décrit par une distribution de probabilité de la forme $P(M \mid O)$. On ne fait aucune hypothèse à ce niveau quant au processus de production lui même, et en particulier on ne suppose pas que le maître fasse de la production suivant une théorie motrice, auditive ou perceptuo-motrice. En revanche, au moment de l'évaluation des modèles, le processus perceptif $P(O_L \mid S)$ de l'agent est implémenté de trois manières différentes, dans le cadre des théories auditive, motrice et perceptuo-motrice (voir la figure 7.2 pour le cas général, ou la section 3.2.1 du chapitre 6 pour le cas des syllabes), afin de les comparer, sous différents niveaux de dégradation (ce qui est fait en faisant varier le modèle de l'environnement $P(S \mid M)$).

(3) **Un modèle subjectif de la *Théorie de l'Esprit*** que nous avons introduite à la section 2.3.2 du chapitre 2. En effet, selon la tâche considérée, certains termes apparaissant

dans les équations probabilistes (voir figure 7.2) peuvent être interprétés comme des modèles internes que l'agent se serait construits de son interlocuteur. Par exemple, dans l'équation probabiliste qui implémente la tâche de perception dans le cadre d'une théorie motrice, le terme $P(M \mid O_S)$ peut être interprété comme un modèle interne du processus de production du locuteur. C'est ce modèle interne qui permet de récupérer les « gestes intentionnels du locuteur ». De même, dans l'équation probabiliste qui implémente la tâche de production dans le cadre d'une théorie auditive, le terme $P(O_L \mid S)$ peut être interprété comme un modèle interne du processus de perception de l'auditeur. Ce modèle interne permet de fixer comme objectif explicite de la production le fait d'être correctement perçu par l'auditeur.

Ainsi, la structure de notre modèle *COSMO* permet à notre agent communicant de raisonner à partir d'un modèle interne de son interlocuteur.

Dans cette thèse, nous avons décrit le modèle *COSMO* en utilisant majoritairement le point de vue (1), mais nous aurions tout aussi bien pu le faire avec l'interprétation (3). En effet, la question que nous nous sommes posés était celle du contenu informationnel de chacune des branches du modèle, plutôt que celle de leur interprétation. Les algorithmes d'apprentissage que nous proposons pour les modèles internes sont intéressants indépendamment de la question de savoir si l'on apprend un modèle de soi ou de l'autre. Nous avons ainsi proposé une explication de ce que des connaissances motrices peuvent apporter à la perception, peu importe qu'il s'agisse de modèles des différents locuteurs ou d'un processus d'interprétation par comparaison à ses gestes propres.

Par ailleurs, il est intéressant de souligner que l'on n'a pas à choisir entre les interprétations (1) et (3) : le formalisme que nous utilisons permet tout à fait d'étendre le modèle *COSMO* en y faisant figurer à la fois un modèle de soi et des modèles des différents interlocuteurs. Cela peut être implémenté en utilisant des variables de cohérence (Bessière *et al.*, 2013) pour connecter les différentes branches du modèle et choisir lesquelles activer lors de l'inférence probabiliste en fonction de la tâche à réaliser. C'est par exemple ce qui a été fait dans les travaux de la thèse d'Estelle Gilet (Gilet, 2009 ; Gilet *et al.*, 2011) qui intègre différents modèles de scripteurs dans son modèle bayésien de la boucle perception-action pour la lecture et l'écriture.

Finalement, force est de constater que la définition mathématique de notre modèle générique permet d'implémenter un certain nombre des principes mis en avant par Moore (2007) dans le cadre du modèle intégrateur *PRESENCE*, que nous avons décrit à la section 3.2.3. En effet, la construction de notre modèle autour de l'internalisation de la situation de communication, d'un mécanisme d'attention partagée, et des modèles internes, de soi ou de l'interlocuteur, permet à notre modèle *COSMO* de *connaître*, produire (*agir*), *percevoir*, *interpréter*, *mémoriser*, *apprendre*, *imiter*, *communiquer* (en reprenant pour chacun de ces termes la définition proposée par (Moore, 2007, pages 427–428)). Cependant, le modèle *PRESENCE* reste plus général et plus ambitieux que le nôtre, et il serait intéressant de regarder comment le formalisme probabiliste pourrait permettre d'intégrer d'autres pièces du *puzzle* dans le modèle *COSMO*.

2.2 Le débat entre les théories auditive, motrice et perceptuo-motrice

Un des objectifs majeur de cette thèse était de réaliser des comparaisons entre les théories auditive, motrice et perceptuo-motrice de la perception de la parole. Nous avons prouvé un

théorème d'indistinguabilité qui montre que les voies auditive et motrice de traitement de la parole d'un agent communicant peuvent encoder exactement la même information, les rendant alors indistinguables expérimentalement. Pour mettre en défaut les hypothèses d'apprentissage parfait de ce théorème, nous avons proposé des algorithmes d'apprentissage plus réalistes.

Les résultats du chapitre 4 (voir notamment la figure 4.11) montrent que l'on observe le plus de différences entre les modèles lorsque le système moteur est appris à partir d'un nombre d'exemplaires limité, alors que prolonger l'apprentissage semble rapprocher du cas d'indistinguabilité. Limiter l'apprentissage du système moteur nous semble plausible : cet apprentissage étant computationnellement plus complexe que celui du système auditif, il a probablement un coût cognitif également plus élevé, ce qui peut justifier l'idée d'arrêter l'apprentissage une fois que des performances satisfaisantes (selon un critère défini par exemple au niveau du succès de la communication) sont atteintes ; mais nous ne pouvons nier qu'il y ait tout de même là une certaine part d'arbitraire. La question que nous voulons alors poser est la suivante : si l'on prolonge l'apprentissage, quelles conséquences cela a-t-il si l'information contenue dans le système moteur et dans le système auditif est parfaitement redondante ?

Dans ce cas le théorème d'indistinguabilité assure que les théories motrice et auditive font exactement les mêmes prédictions. Ce n'est en revanche pas le cas de la théorie perceptuo-motrice. En effet, la fusion perceptuo-motrice réalisée au sein du modèle *COSMO* grâce à la variable C est calculée, comme le montre la figure 7.2, en faisant le produit renormalisé des distributions de probabilité inférées par les modèles moteur et auditif. Lorsque ces deux sources d'information sont concordantes (ou même parfaitement redondantes dans le cas d'indistinguabilité), cela donne simplement un modèle qui est plus sûr de lui. Dans notre cadre, certes limité à la classification d'un petit nombre d'objets, cela se traduit par des performances qui sont plus élevées. Ainsi, la supériorité de notre modèle perceptuo-moteur évalué sur une tâche de perception, que nous avons observée dans tous nos résultats de simulation aux chapitres 4 et 6, est également présente lorsque les modèles auditif et moteur fournissent exactement la même information. De ce fait, les connaissances motrices permettent d'améliorer les performances en perception (grâce à la fusion) même dans le cas où elles sont exactement identiques aux connaissances auditives.

Considérer ce qui se fait dans le domaine de la parole audio-visuelle nous laisse tout de même penser qu'il n'y a pas de raison que la fusion perceptuo-motrice apporte en toutes circonstances de meilleures performances que celles des modèles qu'elle mélange. Le mécanisme de fusion que nous avons implémenté dans cette thèse sous forme de produit de distributions de probabilité renormalisé (ce qui s'appelle de la fusion de capteurs dans le domaine de la robotique (Lebeltel *et al.*, 2004), ou qui pour Hinton (2002) revient à faire un *product of experts* plutôt qu'une *sum of experts*) a été comparé dans le domaine de la reconnaissance automatique de la parole à d'autres manière de fusionner les connaissances articulatoires et acoustiques par Kirchhoff *et al.* (2002) qui ont montré qu'il s'agit de la méthode permettant d'avoir le taux d'erreurs le plus faible. Il y a toutefois une littérature considérable sur l'optimalité des processus de fusion audio-visuelle (voir par exemple Massaro (1987)), et un enjeu important pour l'avenir nous semble porter sur l'étude comparative des mécanismes de fusion des connaissances motrices et auditives dans la perception de la parole.

Par ailleurs, les travaux de Ernst et Banks (2002) ont montré expérimentalement que des mécanismes d'intégration multimodale peuvent se faire de manières différentes chez des sujets différents, qui ne donnent pas les mêmes poids aux différentes modalités. Dès lors, il pourrait être

intéressant d'étudier expérimentalement si c'est le cas également pour l'intégration perceptuo-motrice dans le cadre de la parole, et de chercher à savoir si les poids relatifs attribués aux réponses des systèmes moteur et auditif changent en fonction du contexte (par exemple en présence de bruit).

2.3 Performances et robustesse de *COSMO*

Comme nous l'avons vu à la section 3.1.2 du chapitre 2, la principale difficulté pour les systèmes de reconnaissance de la parole consiste à traiter la grande variabilité du signal de parole. Pour améliorer les performances d'un modèle, la tendance dans le domaine du *machine learning* est de se contenter d'accumuler toujours plus de données, pour capturer de manière plus précise cette variabilité dans des modèles statistiques. Une autre approche beaucoup moins répandue, qui est défendue par Moore (2007), consiste à intégrer dans la structure des modèles le nécessaire pour décrire les causes de cette variabilité.

C'est ce que nous faisons en proposant le modèle *COSMO*, qui combine un modèle purement descriptif (la branche de perception auditive, apprise par association) avec un modèle génératif (la branche de perception motrice, qui fonctionne selon le principe d'analyse par la synthèse, et qui est apprise par accomodation). Ces deux modèles ont une structure différente, des dynamiques d'évolution différentes, et capturent de la variabilité dans des espaces différents.

Dans la branche de perception motrice, le terme $P(M \mid O_S)$ exprime la variabilité dans le choix des commandes motrices associées aux objets, et le terme $P(S \mid M)$ exprime (d'une manière pouvant être approximative) la variabilité dans la réalisation de ces commandes, et la variabilité liée à d'éventuelles dégradations lors de la transmission du signal par l'environnement. Dans la branche de perception auditive, le terme $P(O_L \mid S)$ capture la variabilité des différentes réalisations des objets considérés, dans sa totalité. Dans le modèle *COSMO-S*, la structure du système moteur étant plus complexe, elle permet de capturer encore davantage de variabilité, aux différents niveaux.

Dans les chapitres 4 et 6, les modèles sont appris avec des données présentant une faible variabilité (en conditions normales) et sont testés sur des données de variabilité plus importante (en conditions dégradées). Le modèle auditif, qui apprend rapidement, encode une variabilité plus faible que celle du modèle moteur, et montre de meilleures performances en conditions normales. En revanche, le modèle moteur, du fait d'une dynamique d'apprentissage plus lente, encode une variabilité supérieure à celle du modèle auditif, ce qui se traduit comparativement par une légère imprécision en conditions normales, mais une robustesse accrue aux conditions dégradées. Quant au modèle perceptuo-moteur, tirant avantageusement parti de la complémentarité des deux niveaux de description des modèle moteur et auditif, il réalise de meilleures performances que ces deux modèles dont il réalise la fusion.

Nous avons montré que le surcroît de variabilité du modèle moteur, dû à une différence de structure et à un apprentissage imparfait, se traduit par une plus grande robustesse aux conditions dégradées. Encoder trop de variabilité dans un modèle peut néanmoins nuire à ses performances. C'est ce que nous avons montré dans les résultats de simulation de Moulin-Frier *et al.* (2012) où nous comparons la robustesse aux dégradations d'un modèle moteur fixé une fois pour toutes, et d'un modèle auditif appris à chaque fois dans les mêmes conditions de dégradations que celles du test. Si le modèle auditif, qui peut s'adapter facilement, capture des

niveaux de variabilité trop élevés, ses performances dégringolent, alors que le modèle moteur, plus difficilement évolutif, permet également de conserver de la stabilité.

Il faut cependant remarquer que le mécanisme de bruitage que nous avons implémenté, consistant à dégrader les variables perceptives dimension par dimension en tirant aléatoirement une perturbation selon une distribution de probabilité gaussienne, n'est pas très réaliste. Pour évaluer l'efficacité de systèmes visant à augmenter l'intelligibilité de vrais signaux de parole, Cooke *et al.* (2013) implémentent du bruit de type *cocktail party* en superposant au signal de parole soit du bruit structuré comme de la parole, soit un signal de parole concurrentiel produit par un autre locuteur, et dont ils contrôlent l'intensité. Dans le cadre du modèle *COSMO-S*, qui se limite à une description utilisant quatre variables perceptives (les premier et second formants de la voyelle, et les second et troisième formants de la consonne), une manière d'implémenter un tel bruit pourrait être de donner en entrée de la tâche de perception, non pas une seule entrée perceptive (s_v, s_c) comme nous l'avons fait dans cette thèse, mais plutôt une distribution de probabilité, définie sur tout l'espace perceptif $S_V \times S_C$, qui contiendrait plusieurs pics représentant les signaux de parole concurrents. Plus fondamentalement, un enjeu majeur de l'apport potentiel du système moteur en conditions adverses nous semble être sa capacité à rétablir, par ses processus génératifs, de l'information manquante supprimée par la compétition entre flux concurrents (voir des études de ce type sur la reconnaissance de lettres manuscrites dans le travail de Gilet (2009)). Ceci est probablement un thème essentiel pour les recherches futures à mener sur le modèle *COSMO*, dans un cadre de traitement de situations plus réalistes et plus complexes.

2.4 L'algorithme d'apprentissage par accommodation

Une de nos contributions majeures réside dans l'algorithme original que nous proposons pour l'apprentissage de compétences motrices par accomodation. Nous nous sommes donné les moyens, au chapitre 4 comme au chapitre 6, de vérifier la convergence de l'algorithme et de comprendre les dynamiques de l'apprentissage. Mais il ne s'agit là que de vérifications empiriques sur des exemples. Une perspective intéressante pourrait être d'étudier plus finement les propriétés mathématiques de notre algorithme d'apprentissage par accomodation et de le tester sur des exemples de référence. Dans cette section, nous nous contentons de le comparer brièvement avec d'autres méthodes d'apprentissage et de représentation de la transformation articulatori-acoustique.

Certaines approches centrées sur le problème de l'inversion, comme c'est le cas des travaux de Ouni *et al.* (2003) ou de Busset (2013), construisent un résumé de la transformation articulatori-acoustique en échantillonnant l'espace moteur avec un pas variable pour avoir un pas constant dans l'espace acoustique. L'avantage de ces approches tient dans le fait qu'elles permettent de représenter finement les non-linéarités : les zones de l'espace moteur plus instables (au sens où elles se projettent dans des zones variées de l'espace acoustique) sont échantillonnées de manière plus fine. De telles approches sont efficaces en termes de codage de l'information : on obtient un résumé compact en représentant finement les zones où il y a de l'information, et en restant plus grossier ailleurs.

L'algorithme d'apprentissage par accomodation que nous avons décrit dans cette thèse tend à faire exactement l'inverse : l'agent accumule des observations dans les zones de l'espace moteur

qui lui permettent de bien atteindre les cibles. Dans le contexte de la théorie quantique de la parole proposée par Stevens (1972), les non-linéarités servent de frontières catégorielles, les objets phonétiques étant plutôt positionnés sur des zones plateaux très stables. Finalement, par opposition à un modèle optimal du point de vue de la théorie de l'information, notre agent cognitif apprend peu de choses sur les non-linéarités, mais apprend bien les zones de stabilité correspondant aux objets phonétiques de son environnement.

Par ailleurs, nous fixons dans notre modèle *COSMO-S* des représentations discrétisées avec un pas uniforme dans l'espace perceptif et dans l'espace moteur. Un prolongement de notre travail pourrait être d'étudier comment faire émerger au cours de l'apprentissage des degrés de discrétisation qui s'adaptent en fonction des stimuli reçus par l'agent. Dans cette perspective, une idée qui nous semble prometteuse serait d'utiliser des outils de type MRBT (*Multi-Resolution Binary Trees*, arbres binaires à résolution multiple, voir Bessière (2002)) qui permettent l'apprentissage de distributions de probabilité en apprenant une discrétisation adaptative de l'espace.

Il est également intéressant de comparer notre algorithme d'apprentissage par accomodation avec les travaux de Moulin-Frier et Oudeyer (2013b), dans lesquels l'apprentissage est guidé par les notions de curiosité et de motivation intrinsèque. Plus précisément, avec l'algorithme proposé par les auteurs, l'agent choisit lui-même les cibles à reproduire de manière à maximiser les progrès réalisés. Cela veut dire en particulier que, une fois que l'agent aura trouvé un geste moteur m permettant d'atteindre une cible s de manière satisfaisante, il ne se donnera plus la cible s comme objectif par la suite (sauf si pour tout le reste de l'espace perceptif les antécédents moteurs sont encore mieux connus). Ceci laisse à penser que l'algorithme de Moulin-Frier et Oudeyer (2013b) permettrait également aux agents apprenant de développer des idiosyncrasies dans leurs choix moteurs. En revanche, cet algorithme, s'il n'est guidé que par de la motivation intrinsèque, va finir par apprendre un modèle interne de la relation articulatori-acoustique qui couvre l'espace perceptif en totalité, ce qui n'est pas compatible avec les données expérimentales sur le développement de la production orofaciale et langagière chez l'enfant.

2.5 L'apprentissage des premiers stades de la communication

Les différents apprentissages que nous avons mis en place au chapitre 6 sont réalisés en trois étapes successives, dont l'ordre est inspiré par la séquence développementale proposée par Kuhl (2004). Cela nous a permis de comparer les systèmes de perception motrice et auditive lorsqu'ils sont appris de manière indépendante, et à partir des mêmes données. Nous avons ainsi pu mieux comprendre comment chacun de ces systèmes capture l'information contenue dans ces données de manière différente et pourquoi ils montrent des performances différentes. Toutefois, cette approche qui consiste à apprendre les branches auditive et motrice de perception par étape et de manière indépendante se heurte à plusieurs limites, que nous allons discuter.

Alors que les théories perceptuo-motrices défendent l'idée que les unités de référence de la parole sont de nature perceptuo-motrice et que les représentations motrices et perceptives sont co-construites lors de l'apprentissage, une première limite de notre choix de procéder par étapes indépendantes tient dans le fait qu'il ne permet pas d'étudier la manière dont les connaissances motrices et auditives influent sur leur développement mutuel. Toutefois, le travail de cette thèse fournit des données et un cadre qui peuvent servir de point de départ pour des travaux

de comparaisons de différentes séquences d'apprentissage.

Plus précisément, nous pouvons modifier l'algorithme d'apprentissage que nous avons proposé de deux manières : en changeant l'implémentation probabiliste de la tâche d'imitation, et en changeant les systèmes qui sont mis à jour par chaque nouvelle observation. En effet, de même que le modèle *COSMO* permet de réaliser une tâche de perception en utilisant uniquement des connaissances auditives, ou uniquement des connaissances motrices, ou encore un mélange des deux, il est possible de réaliser la tâche d'imitation de différentes manières. Dans le cadre du modèle *COSMO-S*, il est possible de choisir les différents systèmes à activer en conditionnant la distribution de probabilité correspondant à la tâche à réaliser par des termes de la forme $[\lambda{=}1]$. Dans cette thèse, nous avons à la section 2.4.1 implémenté une « imitation réflexe » qui ne repose que sur le système sensori-moteur en calculant la distribution de probabilité conditionnelle $P(M_V\ M_C \mid S_V\ S_C)$. À la section 2.5.1 nous avons implémenté une « imitation influencée par les interprétations phonémiques des gestes moteurs » qui repose sur le système moteur et sur le système sensori-moteur en calculant la distribution de probabilité conditionnelle $P\,(M_V\ M_C \mid S_V\ S_C\ [O_S{=}o]\ [\lambda_{MV}{=}1]\ [\lambda_{MC}{=}1])$. Dans la perspective de faire de l'imitation non-supervisée, on pourrait calculer plutôt la distribution de probabilité conditionnelle $P\,(M_V\ M_C \mid S_V\ S_C\ [\lambda_{MV}{=}1]\ [\lambda_{MC}{=}1])$. De même, on pourrait s'intéresser au calcul de $P\,(M_V\ M_C \mid S_V\ S_C\ [\lambda_{SV}{=}1]\ [\lambda_{SC}{=}1])$, ce qui est une manière d'implémenter une « imitation influencée par les interprétations phonémiques du signal acoustique », ou encore au calcul de $P\,(M_V\ M_C \mid S_V\ S_C\ [\lambda_{MV}{=}1]\ [\lambda_{MC}{=}1]\ [\lambda_{SV}{=}1]\ [\lambda_{SC}{=}1])$, ce qui est une manière d'implémenter une « imitation influencée par les interprétations phonémiques des propriétés perceptuo-motrices du signal de parole ».

Une autre limite de notre approche tient au fait qu'elle repose en partie sur de l'apprentissage supervisé. Alors que l'on peut tout à fait envisager des mécanismes de haut niveau pour décrire l'interaction avec l'agent maître (succès de la communication, attention fixée sur le même objet sémantique), il n'est pas réaliste que le maître transmette une classification phonétique de bas niveau. Le travail réalisé dans cette thèse pourrait ainsi être étendu en s'inspirant par exemple des travaux de Vallabha *et al.* (2007) qui parviennent à faire émerger à la fois le nombre et la structure de catégories vocaliques. Plus fondamentalement, alors que notre modèle *COSMO-S* est limité au niveau phonétique, il est clair que le bébé doit tenter d'apprendre de manière concomitante des connaissances sur tous les niveaux de la communication, conjointement, en alliant des connaissances phonétiques, phonologiques, lexicales, sémantiques, syntaxiques, voire pragmatiques (voir par exemple Peperkamp et Dupoux (2007)). Une perspective intéressante pourrait être alors d'étendre le modèle *COSMO-S* pour regarder quelles sont les catégories qui émergent conjointement dans l'espace acoustique et dans l'espace moteur grâce à l'ajout des niveaux de représentations lexicales et sémantiques, ce qui peut être fait en s'inspirant de Feldman *et al.* (2009) et de Fourtassi et Dupoux (2014).

Enfin, une limite importante de notre modèle et de ce qu'il peut apprendre réside dans le fait qu'il est très pauvre en termes de complexité : la syllabe est résumée par deux états, eux-mêmes caractérisés par deux valeurs de formants. Pour traiter de vrais signaux de parole on pourrait, comme le propose Huang (2012), introduire un module de synthèse pour pouvoir apprendre les paramètres d'un terme de la forme $P(\text{signal} \mid \text{formants})$ qui serait intégré au modèle *COSMO*. Il faudrait alors mettre en place une manière adéquate de gérer les aspects temporels, ce qui peut nécessiter de dupliquer la structure du modèle *COSMO* à chaque pas

de temps. À plus court terme, nos travaux seront prolongés par une autre thèse qui étudiera les aspects dynamiques liés au contrôle moteur de la parole dans la production de séquences simples en inscrivant *GEPPETO*[40], un modèle biomécanique de production (Perrier et Ma, 2008), dans un modèle probabiliste inspiré de *COSMO*. Cette thèse, qui pourra être l'occasion d'étudier la variabilité acoustique liée à la réalisation biomécanique de commandes motrices, fournit également un cadre idéal pour s'intéresser à la notion de point de contrôle et pour regarder si les trajectoires articulatoires permettent de reconstituer des cibles non-atteintes ou de l'information manquante dans le signal acoustique, comme cela a été fait dans la thèse de Gilet (2009) dans le cas de l'écriture.

2.6 Situations d'apprentissage

Une des limites de notre scénario d'apprentissage est que l'agent apprend à partir des productions d'un seul maître, alors qu'en pratique le bébé a plusieurs locuteurs dans son environnement. Que faudrait-il changer à notre approche pour rendre possible de l'apprentissage multi-maîtres ?

Il faudrait tout d'abord se doter de données correspondant à plusieurs locuteurs, par exemple en adaptant successivement le modèle de synthèse *VLAM* à des locuteurs différents, ce qui peut être fait en s'inspirant de la méthode utilisée par Galván-Rodriguez (1997) pour adapter *VLAM* à un locuteur donné à partir d'un petit nombre de formants de voyelles. En revanche, la principale difficulté réside dans le problème de la normalisation : sachant qu'il est par exemple bien connu que la taille et la position du triangle acoustique maximal pour les voyelles varie d'un locuteur à l'autre, comment est-ce que le bébé apprend que des productions qui sont dans des zones de l'espace différentes en fonction des locuteurs correspondent en fait au même objet ?

Nous proposons comme perspective pour des recherches futures d'essayer de résoudre ce problème de la normalisation en deux temps. On peut imaginer d'abord une première phase de normalisation intrinsèque du signal de parole en utilisant des variables perceptives intrinsèquement normalisatrices, dont on a des raisons de penser qu'elles sont exploitées très tôt par le bébé. On peut penser par exemple à des rapports de formants (ou, sur une échelle en bark à des différences entre formants) ou à des rapports entre formants et fréquence fondamentale, en s'appuyant sur des travaux tels que ceux de Miller (1989) et de Ménard (2002), qui montrent que ces variables permettent de capturer des invariants plus robustes aux variations d'âge et de sexe des locuteurs. C'est également le choix qui est fait dans le modèle *DIVA* (Guenther *et al.*, 1998). Ceci pourrait permettre d'amorcer dans une seconde phase un processus de normalisation qui pourrait ensuite se raffiner de manière extrinsèque au signal de parole, grâce à l'apprentissage d'un modèle adapté à chaque locuteur.

Une autre question qu'il serait intéressant d'aborder au sein du modèle *COSMO* concerne l'apprentissage d'une langue étrangère. Avec les algorithmes d'apprentissage que nous proposons, les modèles auditif et moteur concentrent progressivement les pics de leurs distributions de probabilité sur les zones de l'espace qui correspondent aux cibles données par le maître et deviennent moins performants autour. Notre apprentissage permet donc de rendre compte d'une forme de rétrécissement perceptif (voir les notions de *perceptual narrowing* : par exemple Pons *et al.* (2009) ; et de *magnet effect* : Kuhl (1991) ; Kuhl et Iverson (1995)) ainsi que de spécialisation en production. Étrangement, la notion de *motor narrowing* semble ne pas être étudiée, et

[40] L'acronyme *GEPPETO* signifie GEstures shaped by the Physics and by a PErceptually oriented Targets Optimization.

le terme n'existe pas. Notre algorithme d'apprentissage par accomodation permet tout de même de montrer comment on peut se focaliser de plus en plus sur certains gestes moteurs en ancrant ainsi des choix idiosyncrasiques. En effet, puisque la redondance du système moteur fait que l'on peut produire les mêmes unités perceptives de plusieurs manières différentes, différents agents peuvent alors se constituer des prototypes moteurs différents pour le même objet phonétique, sans dégrader les performances en production ni en perception. Dans ce contexte, la question de l'apprentissage d'une langue étrangère consiste à étudier comment évoluent dans les espaces moteur et auditif les prototypes connus pour apprendre de nouveaux objets phonétiques.

Ce sera l'objectif d'une seconde thèse démarrant prochainement que de profiter du cadre probabiliste du modèle *COSMO* et de l'algorithme d'apprentissage par accomodation pour étudier le développement d'idiosyncrasies en production ainsi qu'en perception, et pour étudier la manière dont peuvent émerger dans les espaces auditif et moteur les catégories phonémiques d'une nouvelle langue alors que celles de la langue maternelle sont déjà bien ancrées.

3 Le mot de la fin

Ainsi se termine ce travail qui, bien évidemment, pose plus de questions qu'il ne donne de réponses, mais qui a le mérite essentiel, selon nous, de fournir un cadre de formalisation des questions posées, et de comparaison des réponses possibles, et ce en référence à quelques grandes questions autour desquelles s'organise une partie des recherches actuelles dans le domaine de la perception de la parole : des recherches portant sur les interactions perceptuo-motrices, sur les processus développementaux, sur la nature des mécanismes d'apprentissage et d'adaptabilité, sur l'émergence de la phonologie et ses relations avec les mécanismes de production et de perception, sur le traitement de la variabilité et de la robustesse de la communication dans les conditions les plus adverses.

C'est ainsi, selon nous, entre formalisation des questions, opérationnalisation par des simulations computationnelles, et retour vers l'expérimentation et les données du terrain cognitif naturel fourni par l'humain apprenant et communicant, que se feront les progrès. Nous espérons y avoir un peu contribué...

Bibliographie

James Abbs et Vincent Gracco : Control of complex motor gestures: Orofacial muscle responses to load perturbations of lip during speech. *Journal of Neurophysiology*, 51(4):705–723, 1984. *Cité page 14*

Christian Abry et Louis-Jean Boë : "Laws" for lips. *Speech communication*, 5(1):97–104, 1986. *Cité page 93*

Christian Abry, Anne Vilain et Jean-Luc Schwartz, éditeurs. *Vocalize to localize*, volume 13. John Benjamins Publishing, 2009. *Cité page 24*

Ziad Al Bawab : *An analysis-by-synthesis approach to vocal tract modeling for robust speech recognition.* Thèse de doctorat, Carnegie Mellon University, 2009. *Cité page 28*

Jussi Alho, Fa-Hsuan Lin, Marc Sato, Hannu Tiitinen, Mikko Sams et Iiro Jääskeläinen : Enhanced neural synchrony between left auditory and premotor cortex is associated with successful phonetic categorization. *Frontiers in psychology*, 5, 2014. *Cité page 21*

Jussi Alho, Marc Sato, Mikko Sams, Jean-Luc Schwartz, Hannu Tiitinen et Iiro Jääskeläinen : Enhanced early-latency electromagnetic activity in the left premotor cortex is associated with successful phonetic categorization. *Neuroimage*, 60(4):1937–1946, 2012. *Cité page 21*

Stephen Anderson : Nasal consonants and the internal structure of segments. *Language*, pages 326–344, 1976. *Cité page 10*

Bishnu Atal, JJ Chang, Max Mathews et John Tukey : Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique. *The Journal of the Acoustical Society of America*, 63(5):1535–1555, 1978. *Cité page 95*

Michiel Bacchiani, Andrew Senior et Georg Heigold : Asynchronous, online, GMM-free training of a context dependent acoustic model for speech recognition. *In Proceedings of the European Conference on Speech Communication and Technology*, 2014. *Cité page 29*

Pierre Badin et Gunnar Fant : Notes on Vocal Tract Computation. *In Quarterly Progress and Status Report, Dept for Speech, Music and Hearing, KTH, Stockholm*, pages 53–108, 1984. *2 citations pages 91 et 93*

Leonardo Badino, Claudia Canevari, Luciano Fadiga et Giorgio Metta : Deep-level acoustic-to-articulatory mapping for DBN-HMM based phone recognition. *In IEEE Spoken Language Technology Workshop (SLT)*, pages 370–375. IEEE, 2012. *2 citations pages 30 et 31*

Leonardo Badino, Alessandro D'Ausilio, Luciano Fadiga et Giorgio Metta : Computational validation of the motor contribution to speech perception. *Topics in cognitive science*, 2014. *2 citations pages 30 et 86*

Janet Baker, Li Deng, James Glass, Sanjeev Khudanpur, Chin-Hui Lee, Nelson Morgan et Douglas O'Shaughnessy : Developments and directions in speech recognition and understanding. *Signal Processing Magazine*, 26(3):75–80, 2009. *Cité page 26*

Adrien Baranes et Pierre-Yves Oudeyer : Active learning of inverse models with intrinsically motivated goal exploration in robots. *Robotics and Autonomous Systems*, 61(1):49–73, 2013. *Cité page 69*

Simon Baron-Cohen, Alan Leslie et Uta Frith : Does the autistic child have a "theory of mind"? *Cognition*, 21(1):37–46, 1985. *Cité page 24*

Denis Beautemps, Pierre Badin et Gérard Bailly : Linear degrees of freedom in speech production: Analysis of cineradio-and labio-film data and articulatory-acoustic modeling. *The Journal of the Acoustical Society of America*, 109(5):2165–2180, 2001. *Cité page 92*

Mohamed Benzeghiba, Renato De Mori, Olivier Deroo, Stephane Dupont, Theodora Erbes, Denis Jouvet, Luciano Fissore, Pietro Laface, Alfred Mertins, Christophe Ris *et al.* : Automatic speech recognition and speech variability: A review. *Speech Communication*, 49(10):763–786, 2007. *Cité page 27*

Frédéric Berthommier, Laurent Girin, Louis-Jean Boë *et al.* : A simple hybrid acoustic/-morphologically constrained technique for the synthesis of stop consonants in various vocalic contexts. *In Proceedings of Interspeech*, 2012. *2 citations pages 93 et 107*

Pierre Bessière, Christian Laugier et Roland Siegwart, éditeurs. *Probabilistic Reasoning and Decision Making in Sensory-Motor Systems*, volume 46 de *Springer Tracts in Advanced Robotics*. Springer-Verlag, Berlin, 2008. *2 citations pages 35 et 41*

Pierre Bessière, Emmanuel Mazer, Juan Manuel Ahuactzin-Larios et Kamel Mekhnacha : *Bayesian Programming*. CRC Press, 2013. *6 citations pages 35, 41, 61, 62, 112, et 154*

Pierre Bessière : Procédure de détermination de la valeur à donner à différents paramètres d'un système, 2002. European Patent EP1525520. *Cité page 158*

Thomas Bever et David Poeppel : Analysis by synthesis: A (re-)emerging program of research for language and vision. *Biolinguistics*, 4(2-3):174–200, 2010. *Cité page 23*

Peter Birkholz et Dietmar Jackel : Influence of temporal discretization schemes on formant frequencies and bandwidths in time domain simulations of the vocal tract system. *In Proceedings of Interspeech*, 2004. *Cité page 32*

Peter Birkholz, Dietmar Jackèl et Bernd Kröger : Construction and control of a three-dimensional vocal tract model. *In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages 873–876. IEEE, 2006. *Cité page 32*

Peter Birkholz, Dietmar Jackèl et Bernd Kroger : Simulation of losses due to turbulence in the time-varying vocal system. *IEEE Transactions on Audio, Speech and Language Processing*, 15(4):1218–1226, 2007. *Cité page 32*

Peter Birkholz et Bernd Kröger : Vocal tract model adaptation using magnetic resonance imaging. *In Proceedings of the 7th International Seminar on Speech Production (ISSP)*, pages 493–500, 2006. *Cité page 32*

Peter Birkholz et Bernd Kröger : Simulation of vocal tract growth for articulatory speech synthesis. *In Proceedings of the 16th International Congress of Phonetic Sciences*, pages 377–380, 2007. *Cité page 32*

Sheila Blumstein, Kenneth Stevens et Georgia Nigro : Property detectors for bursts and transitions in speech perception. *The Journal of the Acoustical Society of America*, 61(5): 1301–1313, 1977. *Cité page 142*

LJ Boë, C Abry, D Beautemps, JL Schwartz et R Laboissière : Les sosies vocaliques. inversion et focalisation. *In Actes des XXIIIèmes Journées d'Étude sur la Parole*, pages 257–260, 2000. *Cité page 95*

Louis-Jean Boë : Modelling the growth of the vocal tract vowel spaces of newly-born infants and adults: consequences for ontogenesis and phylogenesis. *In Proceedings of the international congress of phonetic sciences*, volume 3, pages 2501–2504, 1999. *2 citations pages 93 et 94*

Louis-Jean Boë, Pierre Badin, Lucie Ménard, Guillaume Captier, Barbara Davis, Peter Macneilage, Thomas R Sawallis et Jean-Luc Schwartz : Anatomy and control of the developing human vocal tract: A response to lieberman. *Journal of Phonetics*, 41(5):379–392, 2013. *Cité page 94*

Louis-Jean Boë, Jean-Louis Heim, Kiyoshi Honda et Shinji Maeda : The potential neandertal vowel space was as large as that of modern humans. *Journal of Phonetics*, 30(3):465–484, 2002. *Cité page 94*

Louis-Jean Boë, Pascal Perrier et Gérard Bailly : The geometric vocal-tract variables controlled for vowel production : Proposals for constraining acoustic-to-articulatory inversion. *Journal of Phonetics*, 20(1):27–38, 1992. *Cité page 95*

Louis-Jean Boë, Jean-Luc Schwartz, Jean Granat, Jean-Louis Heim, Antoine Serrurier, Pierre Badin, Guillaume Captier et Pierre Bessière : L'émergence de la parole: Aspects historiques et épistémologiques d'une nouvelle réarticulation. *Faits de langues*, 37, 2011. *Cité page 24*

Paul Boersma : *Functional phonology: Formalizing the interactions between articulatory and perceptual drives.* Holland Academic Graphics, 1998. *Cité page 28*

André Bothorel : *Cinéradiographie des voyelles et consonnes du français: recueil de documents synchronisés pour quatre sujets: vues latérales du conduit vocal, vues frontales de l'orifice labial, données acoustiques.* Institut de phonétique de Strasbourg, 1986. *2 citations pages 92 et 93*

Catherine Browman et Louis Goldstein : Towards an articulatory phonology. *Phonology Yearbook*, 3:219–252, 1986. *Cité page 8*

Catherine Browman et Louis Goldstein : Articulatory gestures as phonological units. *Phonology*, 6(02):201–251, 1989. *3 citations pages 8, 20, et 51*

Catherine Browman et Louis Goldstein : Tiers in articulatory phonology, with some implications for casual speech. *Papers in laboratory phonology I: Between the grammar and physics of speech*, pages 341–376, 1990. *Cité page 10*

Catherine Browman et Louis Goldstein : Articulatory phonology: an overview. *Phonetica*, 49(3-4):155–180, 1992. *2 citations pages 8 et 9*

Jana Brunner : *Acoustic compensation and articulo-motor reorganisation in perturbed speech.* Thèse de doctorat, Humboldt Universität Berlin, 2008. *Cité page 97*

Julie Busset : *Inversion acoustique articulatoire à partir de coefficients cepstraux.* Thèse de doctorat, Université de Lorraine, 2013. *3 citations pages 93, 110, et 157*

Daniel Callan, Jeffery Jones, Akiko Callan et Reiko Akahane-Yamada : Phonetic perceptual identification by native-and second-language speakers differentially activates brain regions involved with acoustic phonetic processing and those involved with articulatory–auditory/orosensory internal models. *NeuroImage*, 22(3):1182–1194, 2004. *Cité page 18*

Daniel Callan, Ray Kent, Frank Guenther et Houri Vorperian : An auditory-feedback-based neural network model of speech production that is robust to developmental changes in the size and shape of the articulatory system. *Journal of Speech, Language and Hearing Research*, 43(3):721–736, 2000. *2 citations pages 93 et 94*

Ruth Campbell, Barbara Dodd et Denis Burnham, éditeurs. *Hearing by eye II: Perspectives and directions in research on audiovisual aspects of language processing.* Psychology Press, 1998. *Cité page 20*

Matthieu Camus : *Identification audio pour la reconnaissance de la parole.* Thèse de doctorat, Université Paris Descartes, 2011. *Cité page 27*

Claudia Canevari, Leonardo Badino, Alessandro D'Ausilio, Luciano Fadiga et Giorgio Metta : Modeling speech imitation and ecological learning of auditory-motor maps. *Frontiers in Psychology*, 4:364, 2013a. *2 citations pages 30 et 31*

Claudia Canevari, Leonardo Badino, Luciano Fadiga et Giorgio Metta : Cross-corpus and cross-linguistic evaluation of a speaker-dependent DNN-HMM ASR system using EMA data. *In Workshop on Speech Production in Automatic Speech Recognition*, 2013b. *2 citations pages 30 et 31*

Claudia Canevari, Leonardo Badino, Luciano Fadiga et Giorgio Metta : Relevance-weighted reconstruction of articulatory features in deep-neural-network-based acoustic-to-articulatory mapping. *In Proceedings of Interspeech*, pages 1297–1301, 2013c. *2 citations pages 30 et 31*

Rolf Carlson, Björn Granström et Dennis Klatt : Vowel perception: The relative perceptual salience of selected acoustic manipulations. *Speech Transmission Laboratories (Stockholm) Quarterly Progress Report SR*, 34:19–35, 1979. *Cité page 91*

Claudio Castellini, Leonardo Badino, Giorgio Metta, Giulio Sandini, Michele Tavella, Mirko Grimaldi et Luciano Fadiga : The use of phonetic motor invariants can improve automatic phoneme discrimination. *PloS ONE*, 6(9):e24055, 2011. *Cité page 30*

Ludmilla Chistovich : Auditory processing of speech. *Language and Speech*, 23(1):67–73, 1980. *Cité page 89*

Cecil Coker : A model of articulatory dynamics and control. *Proceedings of the IEEE*, 64 (4):452–460, 1976. *Cité page 91*

Francis Colas, Julien Diard et Pierre Bessière : Common bayesian models for common cognitive issues. *Acta Biotheoretica*, 58:191–216, 2010. *4 citations pages 35, 41, 52, et 112*

Martin Cooke et Daniel PW Ellis : The auditory organization of speech and other sources in listeners and computational models. *Speech communication*, 35(3):141–177, 2001. *Cité page 90*

Martin Cooke, Simon King, Maëva Garnier et Vincent Aubanel : The listening talker: A review of human and algorithmic context-induced modifications of speech. *Computer Speech & Language*, 28(2):543–571, 2014. *Cité page 27*

Martin Cooke, Catherine Mayo et Cassia Valentini-Botinhao : Intelligibility-enhancing speech modifications: the hurricane challenge. *In Proceedings of Interspeech*, pages 3552–3556, 2013. *Cité page 157*

Franklin Cooper, Pierre Delattre, Alvin Liberman, John Borst et Louis Gerstman : Some experiments on the perception of synthetic speech sounds. *The Journal of the Acoustical Society of America*, 24(6):597–606, 1952. *Cité page 140*

Gregory Cooper : The computational complexity of probabilistic inference using bayesian belief networks. *Artificial intelligence*, 42(2):393–405, 1990. *Cité page 44*

William Cooper : *Speech perception and production: Studies in selective adaptation.* Ablex Publishing Corporation, 1979. *Cité page 20*

Alejandrina Cristia et Amanda Seidl : The hyperarticulation hypothesis of infant-directed speech. *Journal of child language*, pages 1–22, 2013. *Cité page 142*

Alessandro D'Ausilio, Ilaria Bufalari, Paola Salmas et Luciano Fadiga : The role of the motor system in discriminating normal and degraded speech sounds. *Cortex*, 48(7):882–887, 2012. *Cité page 21*

Alessandro D'Ausilio, Friedemann Pulvermüller, Paola Salmas, Ilaria Bufalari, Chiara Begliomini et Luciano Fadiga : The motor somatotopy of speech perception. *Current Biology*, 19(5):381–385, 2009. *2 citations pages 21 et 30*

Steven Davis et Paul Mermelstein : Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 28(4):357–366, 1980. *Cité page 26*

Pierre Delattre et Donald Freeman : A dialect study of american r's by x-ray motion picture. *Linguistics*, 6(44):29–68, 1968. *2 citations pages 13 et 15*

Pierre Delattre, Alvin Liberman et Franklin Cooper : Acoustic loci and transitional cues for consonants. *The Journal of the Acoustical Society of America*, 27(4):769–773, 1955. *Cité page 99*

Sébastien Demange et Slim Ouni : An episodic memory-based solution for the acoustic-to-articulatory inversion problem. *The Journal of the Acoustical Society of America*, 133 (5):2921–2930, 2013. *Cité page 95*

Arthur Dempster, Nan Laird et Donald Rubin : Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38, 1977. *Cité page 26*

Li Deng, Gordon Ramsay et Don Sun : Production models as a structural basis for automatic speech recognition. *Speech Communication*, 22(2):93–111, 1997. *Cité page 28*

Li Deng et Don Sun : A statistical approach to automatic speech recognition using the atomic speech units constructed from overlapping articulatory features. *The Journal of the Acoustical Society of America*, 95(5):2702–2719, 1994. *Cité page 28*

Randy Diehl et Keith Kluender : On the objects of speech perception. *Ecological Psychology*, 1(2):121–144, 1989. *Cité page 13*

Randy Diehl, Andrew Lotto et Lori Holt : Speech perception. *Annual Review of Psychology*, 55(1):149–179, 2004. *4 citations pages 13, 20, 51, et 56*

Barbara Dodd et Ruth Campbell, éditeurs. *Hearing by eye: The psychology of lip-reading.* Lawrence Erlbaum Associates, 1987. *Cité page 20*

Robert Dooling, Catherine Best et Susan Brown : Discrimination of synthetic full-formant and sinewave/ra–la/continua by budgerigars (melopsittacus undulatus) and zebra finches (taeniopygia guttata). *The Journal of the Acoustical Society of America*, 97:1839, 1995. *Cité page 13*

Yi Du, Bradley Buchsbaum, Cheryl Grady et Claude Alain : Noise differentially impacts phoneme representations in the auditory and speech motor systems. *Proceedings of the National Academy of Sciences*, 111(19):7126–7131, 2014. *Cité page 21*

Virginie Ducey-Kaufmann : *Le cadre de la parole et le cadre du signe: Un rendez-vous développemental.* Thèse de doctorat, Université Stendhal - Grenoble III, 2007. Département Parole et Cognition. *Cité page 24*

Cornelia Eckers et Bernd Kröger : Semantic, phonetic, and phonological knowledge in a neurocomputational model of speech acquisition. *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung*, pages 244–251, 2012. *Cité page 32*

Marc Ernst et Martin Banks : Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, 415(6870):429–433, 2002. *Cité page 155*

Carol Espy-Wilson et Suzanne Boyce : Acoustic differences between "bunched"and "retroflex"variants of american english/r. *The Journal of the Acoustical Society of America*, 95:2823, 1994. *Cité page 15*

Luciano Fadiga, Laila Craighero, Giovanni Buccino et Giacomo Rizzolatti : Speech listening specifically modulates the excitability of tongue muscles: a tms study. *European Journal of Neuroscience*, 15(2):399–402, 2002. *Cité page 21*

Gunnar Fant : Acoustic theory of speech production.'sgravenhage: Mouton. *The Netherlands*, 1960. *Cité page 91*

Gunnar Fant : *Acoustic theory of speech production: with calculations based on X-ray studies of Russian articulations*, volume 2. Walter de Gruyter, 1971. *Cité page 91*

Edda Farnetani et Daniel Recasens : Coarticulation and connected speech processes. *In The handbook of phonetic sciences*, pages 371–404. Blackwell Oxford, 1997. *Cité page 97*

Naomi Feldman, Thomas Griffiths et James Morgan : Learning phonetic categories by learning a lexicon. *In Proceedings of the 31st annual conference of the cognitive science society*, pages 2208–2213. Cognitive Science Society (CD-ROM) Austin, TX, 2009. *Cité page 159*

Anne Fernald et Patricia Kuhl : Acoustic determinants of infant preference for motherese speech. *Infant behavior and development*, 10(3):279–293, 1987. *Cité page 142*

Cécile Fougeron : La phonologie articulatoire: une introduction. *In* N. Nguyen, S. Wauquier-Gravelines et J. Durand, éditeurs : *Phonologie et phonétique: Forme et substance*, Cognition et Traitement de l'Information, pages 265–290. Hermès, Paris, France, 2005. *2 citations pages 8 et 9*

Abdellah Fourtassi et Emmanuel Dupoux : A rudimentary lexicon and semantics help bootstrap phoneme acquisition. *In Proceedings of the 18th Conference on Computational Natural Language Learning (CoNLL)*, 2014. *Cité page 159*

Carol Fowler : An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics*, 14(1):3–28, 1986. *Cité page 12*

Carol Fowler : Compensation for coarticulation reflects gesture perception, not spectral contrast. *Perception & Psychophysics*, 68(2):161–177, 2006. *Cité page 12*

Carol Fowler, Julie Brown, Laura Sabadini et Jeffrey Weihing : Rapid access to speech gestures in perception: Evidence from choice and simple response time tasks. *Journal of Memory and Language*, 49(3):396–413, 2003. *Cité page 12*

Carol Fowler et Dawn Dekle : Listening with eye and hand: Cross-modal contributions to speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, 17(3):816–828, 1991. *Cité page 12*

Bruce Francis et Murray Wonham : The internal model principle of control theory. *Automatica*, 12(5):457–465, 1976. *Cité page 22*

Joe Frankel, Korin Richmond, Simon King et Paul Taylor : An automatic speech recognition system using neural networks and linear dynamic models to recover and model articulatory traces. *In Sixth International Conference on Spoken Language Processing*, volume 4. International Speech Communication Association, 2000. *2 citations pages 28 et 29*

Joe Frankel, Mirjam Wester et Simon King : Articulatory feature recognition using dynamic bayesian networks. *Computer Speech & Language*, 21(4):620–640, 2007. *Cité page 29*

Bernard Gabioud : Articulatory models in speech synthesis. *In* E. Keller, éditeur : *Fundamentals of speech synthesis and speech recognition*, pages 215–230. John Wiley and Sons Ltd., 1994. *2 citations pages 93 et 104*

Bruno Galantucci, Carol A Fowler et Michael T Turvey : The motor theory of speech perception reviewed. *Psychonomic bulletin & review*, 13(3):361–377, 2006. *2 citations pages 12 et 96*

Vittorio Gallese et Alvin Goldman : Mirror neurons and the simulation theory of mind-reading. *Trends in cognitive sciences*, 2(12):493–501, 1998. *Cité page 24*

Arturo Galván-Rodriguez : *Études dans le cadre de l'inversion acoustico-articulatoire : Amélioration d'un modèle articulatoire, normalisation du locuteur et récupération du lieu de constriction des occlusives.* Thèse de doctorat, Institut National Polytechnique de Grenoble, 1997. *2 citations pages 104 et 160*

James Gibson : *The ecological approach to visual perception.* Lawrence Erlbaum Associates, 1986. *Cité page 12*

Estelle Gilet : *Modélisation Bayésienne d'une boucle perception-action : application à la lecture et à l'écriture.* Thèse de doctorat, Université Joseph Fourier – Grenoble, Grenoble, France, 2009. *4 citations pages 2, 154, 157, et 160*

Estelle Gilet, Julien Diard et Pierre Bessière : Bayesian action–perception computational model: Interaction of production and recognition of cursive letters. *PLoS ONE*, 6(6):e20387, 2011. *4 citations pages 2, 35, 112, et 154*

Anne-Lise Giraud et David Poeppel : Speech perception from a neurophysiological perspective. *In The human auditory cortex*, pages 225–260. Springer, 2012. *Cité page 90*

Louis Goldstein et Carol Fowler : Articulatory phonology: A phonology for public language use. *Phonetics and phonology in language comprehension and production: Differences and similarities*, pages 159–207, 2003. *Cité page 8*

Krystyna Grabski : *LES CARTES SENSORIMOTRICES DE LA PAROLE: Corrélats neurocognitifs et couplage fonctionnel des systèmes de perception et de production des voyelles du Français.* Thèse de doctorat, Université de Grenoble, 2012. Département Parole et Cognition. *2 citations pages 21 et 39*

Krystyna Grabski, Jean-Luc Schwartz, Laurent Lamalle, Coriandre Vilain, Nathalie Vallée, Monica Baciu, Jean-François Le Bas et Marc Sato : Shared and distinct neural correlates of vowel perception and production. *Journal of Neurolinguistics*, 26(3):384–408, 2013. *Cité page 21*

Alex Graves, Abdel-rahman Mohamed et Geoffrey Hinton : Speech recognition with deep recurrent neural networks. *In Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on*, pages 6645–6649. IEEE, 2013. *2 citations pages 26 et 29*

Thomas Griffiths, Nick Chater, Charles Kemp, Amy Perfors et Joshua Tenenbaum : Probabilistic models of cognition: exploring representations and inductive biases. *Trends in cognitive sciences*, 14(8):357–364, 2010. *Cité page 35*

Mirko Grimaldi, Barbara Gili Fivela, Francesco Sigona, Michele Tavella, Paul Fitzpatrick, Laila Craighero, Luciano Fadiga, Giulio Sandini et Giorgio Metta : New technologies for simultaneous acquisition of speech articulatory data: 3D articulograph, ultrasound and electroglottograph. *Proceedings Language Teching*, 2008. *2 citations pages 30 et 31*

Rick Grush : The emulation theory of representation: motor control, imagery, and perception. *Behavioral and Brain Sciences*, 27(3):377–396, 2004. *Cité page 23*

Frank Guenther : Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production. *Psychological review*, 102(3):594–621, 1995. *3 citations pages 69, 93, et 94*

Frank Guenther : Cortical interactions underlying the production of speech sounds. *Journal of Communication Disorders*, 39(5):350–365, 2006. *5 citations pages 15, 16, 20, 51, et 93*

Frank Guenther, Carol Espy-Wilson, Suzanne Boyce, Melanie Matthies, Majid Zandipour et Joseph Perkell : Articulatory tradeoffs reduce acoustic variability during american english/r/production. *The Journal of the Acoustical Society of America*, 105 (5):2854–2865, 1999. *2 citations pages 93 et 94*

Frank Guenther, Michelle Hampson et Dave Johnson : A theoretical investigation of reference frames for the planning of speech movements. *Psychological Review*, 105:611–633, 1998. *6 citations pages 14, 20, 51, 93, 94, et 160*

Robert Hagiwara : Three types of american/r/. *UCLA Working Papers in Phonetics*, pages 55–62, 1994. *Cité page 15*

Morris Halle et Kenneth Stevens : Analysis by synthesis. *In Proceedings of the Seminar on Speech Compression and Processing*, volume 2, 1959. *Cité page 17*

Morris Halle et Kenneth Stevens : Speech recognition: A model and a program for research. *IRE Transactions on Information Theory*, 8(2):155–159, 1962. *Cité page 17*

Richard Harshman, Peter Ladefoged et Louis Goldstein : Factor analysis of tongue shapes. *The Journal of the Acoustical Society of America*, 62(3):693–707, 1977. *Cité page 92*

Jeff Hawkins et Sandra Blakeslee : *On intelligence.* Times Books, 2004. *2 citations pages 33 et 34*

Katrina Hayward : *Experimental phonetics.* Longman London, 2000. *Cité page 101*

Donald Hebb : *The organization of behavior: A neuropsychological theory.* Psychology Press, 2002. *Cité page 32*

John Heinz et Kenneth Stevens : On the relations between lateral cineradiographs, area functions, and acoustic spectra of speech. *In Proc. 5th Int. Congress of Acoustics*, volume 44, 1965. *Cité page 93*

Hynek Hermansky : Perceptual linear predictive (PLP) analysis of speech. *the Journal of the Acoustical Society of America*, 87(4):1738–1752, 1990. *Cité page 26*

Hynek Hermansky et Nelson Morgan : RASTA processing of speech. *IEEE Transactions on Speech and Audio Processing*, 2(4):578–589, 1994. *Cité page 26*

Gregory Hickok et David Poeppel : Towards a functional neuroanatomy of speech perception. *Trends in cognitive sciences*, 4(4):131–138, 2000. *Cité page 21*

Gregory Hickok et David Poeppel : Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language. *Cognition*, 92(1):67–99, 2004. *Cité page 21*

Gregory Hickok et David Poeppel : The cortical organization of speech processing. *Nature Reviews Neuroscience*, 8(5):393–402, 2007. *Cité page 21*

Geoffrey Hinton : Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800, 2002. *Cité page 155*

Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath *et al.* : Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *Signal Processing Magazine, IEEE*, 29(6):82–97, 2012. *Cité page 26*

Guangpu Huang : *Articulatory Phonetic Features for Robust Speech Recognition.* Thèse de doctorat, Nanyang Technological University, 2012. *2 citations pages 28 et 159*

Guangpu Huang et Meng Joo Er : A novel neural-based pronunciation modeling method for robust speech recognition. *In IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 517–522. IEEE, 2011. *Cité page 28*

Guangpu Huang et Meng Joo Er : Model-based articulatory phonetic features for improved speech recognition. *In IEEE International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2012. *Cité page 28*

Thomas Hueber, Elie-Laurent Benaroya, Gérard Chollet, Bruce Denby, Gérard Dreyfus et Maureen Stone : Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips. *Speech Communication*, 52(4):288–300, 2010. *2 citations pages 28 et 29*

Marco Iacoboni : Understanding others: Imitation, language, empathy. *In* Susan Hurley et Nick Chater, éditeurs : *Perspectives on imitation: from mirror neurons to memes*, volume 1, pages 255–282. MIT Press, 2005. *Cité page 33*

Charlotte Jacquemot, Emmanuel Dupoux et Anne-Catherine Bachoud-Lévi : Breaking the mirror: Asymmetrical disconnection between the phonological input and output codes. *Cognitive Neuropsychology*, 24(1):3–22, 2007. *2 citations pages 40 et 153*

Edwin Jaynes : *Probability Theory: The Logic of Science.* Cambridge University Press, 2003. *2 citations pages 35 et 36*

Ian Jolliffe : *Principal component analysis.* Wiley Online Library, 2005. *Cité page 92*

Matt Jones et Bradley Love : Bayesian fundamentalism or enlightenment? on the explanatory status and theoretical contributions of bayesian models of cognition. *Behavioral and Brain Sciences*, 34(04):169–188, 2011. *Cité page 35*

Michael Jordan et David Rumelhart : Forward models: Supervised learning with a distal teacher. *Cognitive science*, 16(3):307–354, 1992. *Cité page 18*

Bing-Hwang Juang, Stephen E Levinson et M Mohan Sondhi : Maximum likelihood estimation for multivariate mixture observations of markov chains (corresp.). *Information Theory, IEEE Transactions on*, 32(2):307–309, 1986. *Cité page 26*

Mitsuo Kawato, Kazunori Furukawa et Ryoji Suzuki : A hierarchical neural-network model for control and learning of voluntary movement. *Biological cybernetics*, 57(3):169–185, 1987. *Cité page 22*

Scott Kelso et Betty Tuller : Intrinsic time in speech production: theory, methodology, and preliminary observations. *In* Eric Keller et Myrna Gopnik, éditeurs : *Motor and sensory processes of language*, pages 203–222. Erlbaum Hillsdale, NJ, 1987. *Cité page 10*

Diane Kewley-Port : Measurement of formant transitions in naturally produced stop consonant–vowel syllables. *The Journal of the Acoustical Society of America*, 72(2):379–389, 1982. *Cité page 99*

Diane Kewley-Port : Time-varying features as correlates of place of articulation in stop consonants. *The Journal of the Acoustical Society of America*, 73(1):322–335, 1983. *2 citations pages 99 et 142*

Simon King, Joe Frankel, Karen Livescu, Erik McDermott, Korin Richmond et Mirjam Wester : Speech production knowledge in automatic speech recognition. *The Journal of the Acoustical Society of America*, 121(2):723–742, 2007. *Cité page 28*

Katrin Kirchhoff : *Robust speech recognition using articulatory information.* Thèse de doctorat, University of Bielefeld, 1999. *Cité page 28*

Katrin Kirchhoff, Gernot Fink et Gerhard Sagerer : Combining acoustic and articulatory feature information for robust speech recognition. *Speech Communication*, 37(3):303–319, 2002. *2 citations pages 28 et 155*

Keith Kluender, Randy Diehl et Peter Killeen : Japanese quail can learn phonetic categories. *Science*, 237(4819):1195–1197, 1987. *Cité page 13*

Teuvo Kohonen : Self-organized formation of topologically correct feature maps. *Biological cybernetics*, 43(1):59–69, 1982. *Cité page 32*

Bernd Kröger et Peter Birkholz : A gesture-based concept for speech movement control in articulatory speech synthesis. *In Verbal and Nonverbal Communication Behaviours*, pages 174–189. Springer, 2007. *Cité page 32*

Bernd Kröger, Peter Birkholz, Jim Kannampuzha, Emily Kaufmann et Christiane Neuschaefer-Rube : Towards the acquisition of a sensorimotor vocal tract action repository within a neural model of speech processing. *In Analysis of Verbal and Nonverbal Communication and Enactment. The Processing Issues*, pages 287–293. Springer, 2011. *2 citations pages 32 et 33*

Bernd Kröger, Peter Birkholz, Jim Kannampuzha et Christiane Neuschaefer-Rube : Learning to associate speech-like sensory and motor states during babbling. *In Proceedings of the 7th International Seminar on Speech Production. Belo Horizonte, Brazil*, pages 67–74, 2006. *Cité page 32*

Bernd Kröger, Jim Kannampuzha, Cornelia Eckers, Stefan Heim, Emily Kaufmann et Christiane Neuschaefer-Rube : The neurophonetic model of speech processing ACT: structure, knowledge acquisition, and function modes. *In Cognitive Behavioural Systems*, pages 398–404. Springer, 2012. *Cité page 31*

Bernd Kröger, Jim Kannampuzha et Emily Kaufmann : Associative learning and self-organization as basic principles for simulating speech acquisition, speech production, and speech perception. *EPJ Nonlinear Biomedical Physics*, 2(1):1–28, 2014. *Cité page 33*

Bernd Kröger, Jim Kannampuzha et Christiane Neuschaefer-Rube : Towards a neurocomputational model of speech production and perception. *Speech Communication*, 51 (9):793–809, 2009. *2 citations pages 32 et 33*

Bernd Kroger, Nick Miller, Anja Lowit et Christiane Neuschaefer-Rube : Defective neural motor speech mappings as a source for apraxia of speech: Evidence from a quantitative neural model of speech processing. *In* Anja Lowit et Raymond Kent, éditeurs : *Assessment of Motor Speech Disorders*, pages 325–346. Plural Publishing, 2011. *Cité page 33*

Bernd J Kröger et Jim Kannampuzha : A neurofunctional model of speech production including aspects of auditory and audio-visual speech perception. *In AVSP*, pages 83–88, 2008. *Cité page 33*

Patricia Kuhl : A new view of language acquisition. *Proceedings of the National Academy of Sciences*, 97(22):11850–11857, 2000. *Cité page 13*

Patricia Kuhl : Early language acquisition: cracking the speech code. *Nature reviews neuroscience*, 5(11):831–843, 2004. *3 citations pages 119, 137, et 158*

Patricia Kuhl et James Miller : Speech perception by the chinchilla: Identification functions for synthetic VOT stimuli. *The Journal of the Acoustical Society of America*, 63:905, 1978. *Cité page 13*

Patricia Kuhl, James Miller *et al.* : Speech perception by the chinchilla: Voiced-voiceless distinction in alveolar plosive consonants. *Science*, 190(4209):69–72, 1975. *Cité page 13*

Patricia Kuhl et Denise Padden : Enhanced discriminability at the phonetic boundaries for the place feature in macaques. *The Journal of the Acoustical Society of America*, 73:1003, 1983. *Cité page 13*

Patricia K Kuhl : Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. *Perception & psychophysics*, 50(2):93–107, 1991. *Cité page 160*

Patricia K Kuhl et Paul Iverson : Chapter 4: Linguistic experience and the "perceptual magnet effect,". *Speech perception and linguistic experience: Issues in cross-language research*, pages 121–154, 1995. *Cité page 160*

Peter Ladefoged, Joseph DeClerk, Mona Lindau et George Papcun : An auditory-motor theory of speech production. *UCLA working papers in phonetics*, 22(48):48–76, 1972. *Cité page 13*

Peter Ladefoged et Keith Johnstone : *A course in phonetics.* Cengage learning, 1982. *Cité page 101*

Claire Lalevée : *Développement du contrôle moteur de la parole: une étude longitudinale d'un enfant francophone âge de 7 à 16 mois, à partir d'un corpus audio-visuel.* Thèse de doctorat, Université de Grenoble, 2010. *Cité page 141*

Raphaël Laurent, Clément Moulin-Frier, Pierre Bessière, Jean-Luc Schwartz et Julien Diard : Integrate, yes, but what and how? a computational approach of sensorimotor fusion in speech. *Behavioral and Brain Sciences*, 36(04):364–365, 2013a. *Cité page 3*

Raphaël Laurent, Jean-Luc Schwartz, Pierre Bessière et Julien Diard : COSMO, un modèle bayésien de la communication parlée : application à la perception des syllabes. *In Actes de la conférence conjointe JEP-TALN-RECITAL*, volume 1 (JEP), pages 305–312, Grenoble, France, 2012. *2 citations pages 4 et 39*

Raphaël Laurent, Jean-Luc Schwartz, Pierre Bessière et Julien Diard : A computational model of perceptuo-motor processing in speech perception: learning to imitate and categorize synthetic CV syllables. *In Proceedings of Interspeech*, pages 2797–2801, Lyon, France, 2013b. *2 citations pages 4 et 147*

Olivier Lebeltel, Pierre Bessière, Julien Diard et Emmanuel Mazer : Bayesian robot programming. *Autonomous Robots*, 16:49–79, 2004. *4 citations pages 35, 41, 52, et 155*

Willem Levelt : Monitoring and self-repair in speech. *Cognition*, 14(1):41–104, 1983. *Cité page 15*

Willem Levelt : The perceptual loop theory not disconfirmed: A reply to mackay. *Consciousness and Cognition*, 1(3):226–230, 1992. *Cité page 15*

Willem Levelt : *Speaking: From intention to articulation*, volume 1. MIT press, 1993. *2 citations pages 15 et 91*

Willem Levelt, Ardi Roelofs et Antje Meyer : A theory of lexical access in speech production. *Behavioral and Brain Sciences*, 22:1–75, 1999. *Cité page 15*

Alvin Liberman, Franklin Cooper, Donald Shankweiler et Michael Studdert-Kennedy : Perception of the speech code. *Psychological review*, 74(6):431–461, 1967. *2 citations pages 23 et 96*

Alvin Liberman, Pierre Delattre, Franklin Cooper et Louis Gerstman : The role of consonant-vowel transitions in the perception of the stop and nasal consonants. *Psychological Monographs: General and Applied*, 68(8):1–13, 1954. *Cité page 12*

Alvin Liberman, David Isenberg et Brad Rakerd : Duplex perception of cues for stop consonants: Evidence for a phonetic mode. *Perception & Psychophysics*, 30(2):133–143, 1981. *Cité page 20*

Alvin Liberman et Ignatius Mattingly : The motor theory of speech perception revised. *Cognition*, 21(1):1–36, 1985. *6 citations pages 11, 12, 20, 23, 24, et 51*

Johan Liljencrants et Björn Lindblom : Numerical simulation of vowel quality systems: The role of perceptual contrast. *Language*, 48(4):839–862, décembre 1972. *Cité page 19*

Mona Lindau : The story of/r. *The Journal of the Acoustical Society of America*, 67:S27, 1980. *Cité page 15*

Björn Lindblom : Phonetic universals in vowel systems. *Experimental phonology*, pages 13–44, 1986. *Cité page 19*

Björn Lindblom : Explaining phonetic variation: A sketch of the H&H theory. *In Speech production and speech modelling*, pages 403–439. Springer, 1990a. *Cité page 33*

Björn Lindblom, James Lubker et Thomas Gay : Formant frequencies of some fixed-mandible vowels and a model of speech motor programming by predictive simulation. *The Journal of the Acoustical Society of America*, 62:S15, 1977. *Cité page 14*

Börn Lindblom : On the notion of 'possible speech sound'. *Journal of phonetics*, 18(2):135–152, 1990b. *Cité page 19*

Karen Livescu, James Glass et Jeff Bilmes : Hidden feature models for speech recognition using dynamic bayesian networks. *In Proceedings of Interspeech*. Citeseer, 2003. *Cité page 28*

Andrew Lotto et Keith Kluender : General contrast effects in speech perception: Effect of preceding liquid on stop consonant identification. *Perception & Psychophysics*, 60(4):602–619, 1998. *Cité page 13*

Andrew Lotto, Keith Kluender et Lori Holt : Perceptual compensation for coarticulation by japanese quail (coturnix coturnix japonica). *The Journal of the Acoustical Society of America*, 102:1134, 1997. *Cité page 13*

Liang Ma : *La coarticulation en français et en chinois: étude expérimentale et modélisation.* Thèse de doctorat, Université de Provence-Aix-Marseille I, 2008. *Cité page 97*

Peter MacNeilage : The frame/content theory of evolution of speech production. *Behavioral and Brain Sciences*, 21(04):499–511, 1998. *Cité page 119*

Shinji Maeda : An articulatory model of the tongue based on a statistical analysis. *The Journal of the Acoustical Society of America*, 65(S1):S22–S22, 1979. *Cité page 92*

Shinji Maeda : Improved articulatory models. *The Journal of the Acoustical Society of America*, 84(S1):S146–S146, 1988. *Cité page 93*

Shinji Maeda : Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal tract shapes using an articulatory model. *In* W.J. Hardcastle et A. Marchal, éditeurs : *Speech production and speech modeling*, pages 131–149. Kluwer Academic, 1990. *3 citations pages 92, 93, et 104*

Shinji Maeda et Kiyoshi Honda : From EMG to formant patterns: the implication of vowel spaces. *Phonetica*, 51:17–29, 1994. *Cité page 93*

Virginia Mann : Influence of preceding liquid on stop-consonant perception. *Perception & Psychophysics*, 28(5):407–412, 1980. *Cité page 13*

Dominic Massaro : *Speech perception by ear and eye: A paradigm for psychological inquiry.* Psychology Press, 1987. *2 citations pages 12 et 155*

Ingo Meister, Stephen Wilson, Choi Deblieck, Allan Wu et Marco Iacoboni : The essential role of premotor cortex in speech perception. *Current Biology*, 17(19):1692–1696, 2007. *Cité page 21*

Lucie Ménard : *Production et perception des voyelles au cours de la croissance du conduit vocal : variabilité, invariance et normalisation.* Thèse de doctorat, Grenoble 3, 2002. *2 citations pages 94 et 160*

Lucie Ménard, Jean-Luc Schwartz, Louis-Jean Boë et Jérôme Aubin : Articulatory–acoustic relationships during vocal tract growth for french vowels: Analysis of real data and simulations with an articulatory model. *Journal of Phonetics*, 35(1):1–19, 2007. *Cité page 94*

Paul Mermelstein : Articulatory model for the study of speech production. *The Journal of the Acoustical Society of America*, 53(4):1070–1082, 1973. *Cité page 91*

Nicholas Metropolis et Stanislaw Ulam : The monte carlo method. *Journal of the American statistical association*, 44(247):335–341, 1949. *Cité page 44*

Christine Meunier : Phonétique acoustique. *In* P. Auzou, éditeur : *Les dysarthries*, pages 164–173. Solal, 2007. *Cité page 102*

Christopher Miall : Connecting mirror neurons and forward models. *Neuroreport*, 14(17):2135–2137, 2003. *Cité page 18*

James Miller : Auditory-perceptual interpretation of the vowel. *The journal of the Acoustical society of America*, 85:2114, 1989. *Cité page 160*

James Miller, Craig Wier, Richard Pastore, William Kelly et Robert Dooling : Discrimination and labeling of noise–buzz sequences with varying noise-lead times: An example of categorical perception. *The Journal of the Acoustical Society of America*, 60:410, 1976. *Cité page 13*

Anne Mills : The development of phonology in the blind child. *In* Barbara Dodd et Ruth Campbell, éditeurs : *Hearing by eye: the psychology of lipreading*, pages 145–161. Lawrence Erlbaum Associates, Londres, 1987. *Cité page 142*

Vikramjit Mitra, Hosung Nam, Carol Espy-Wilson, Elliot Saltzman et Louis Goldstein : Articulatory information for noise robust speech recognition. *IEEE Transactions on Audio, Speech and Language Processing*, 19(7):1913–1924, 2011. *Cité page 28*

Vikramjit Mitra, Hosung Nam, Carol Espy-Wilson, Elliot Saltzman et Louis Goldstein : Recognizing articulatory gestures from speech for robust speech recognition. *The Journal of the Acoustical Society of America*, 131:2270–2287, 2012. *2 citations pages 28 et 29*

Vikramjit Mitra, Ganesh Sivaraman, Hosung Nam, Carol Espy-Wilson et Elliot Saltzman : Articulatory features from deep neural networks and their role in speech recognition. *In Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014. *2 citations pages 28 et 29*

Kenneth Moll : Cinefluorographic techniques in speech research. *Journal of Speech, Language, and Hearing Research*, 3(3):227–241, 1960. *Cité page 92*

Roger Moore : Spoken language processing: Piecing together the puzzle. *Speech communication*, 49(5):418–435, 2007. *4 citations pages 33, 34, 154, et 156*

Roger Moore : Spoken language processing: where do we go from here? *In* Robert Trappl, éditeur : *Your Virtual Butler*, pages 119–133. Springer, 2013. *2 citations pages 33 et 34*

Riikka Möttönen, Rebekah Dutton et Kate Watkins : Auditory-motor processing of speech sounds. *Cerebral Cortex*, 23(5):1190–1197, 2013. *Cité page 21*

Riikka Möttönen et Kate Watkins : Motor representations of articulators contribute to categorical perception of speech sounds. *The Journal of Neuroscience*, 29(31):9819–9825, 2009. *Cité page 21*

Clément Moulin-Frier : *Rôle des relations perception-action dans la communication parlée et l'émergence des systèmes phonologiques: étude, modélisation computationnelle et simulations*. Thèse de doctorat, Université de Grenoble, 2011. Département Parole et Cognition. *5 citations pages 2, 20, 24, 25, et 36*

Clément Moulin-Frier, Raphaël Laurent, Pierre Bessière, Jean-Luc Schwartz et Julien Diard : Adverse conditions improve distinguishability of auditory, motor and perceptuo-motor theories of speech perception: an exploratory Bayesian modeling study. *Language and Cognitive Processes*, 27(7/8):1240–1263, 2012. *5 citations pages 2, 4, 39, 85, et 156*

Clément Moulin-Frier et Pierre-Yves Oudeyer : Exploration strategies in developmental robotics: a unified probabilistic framework. *In Development and Learning and Epigenetic Robotics (ICDL), 2013 IEEE Third Joint International Conference on*, pages 1–6. IEEE, 2013a. *Cité page 69*

Clément Moulin-Frier et Pierre-Yves Oudeyer : The role of intrinsic motivations in learning sensorimotor vocal mappings: a developmental robotics study. *In Proceedings of Interspeech*, 2013b. *Cité page 158*

Clément Moulin-Frier, Jean-Luc Schwartz, Julien Diard et Pierre Bessière : Emergence of articulatory-acoustic systems from deictic interaction games in a "vocalize to localize" framework. *In* Anne Vilain, Jean-Luc Schwartz, Christian Abry et Jacques Vauclair, éditeurs : *Primate communication and human language: Vocalisations, gestures, imitation and deixis in humans and non-humans*, Advances in Interaction Studies, pages 193–220. John Benjamins Pub. Co., Amsterdam / Philadelphia, PA, 2011. *Cité page 2*

Clément Moulin-Frier, Jean-Luc Schwartz, Julien Diard et Pierre Bessière : A unified theoretical bayesian model of speech communication. *In* Vincent Duffy, éditeur : *Proceedings of the first conference on Applied Digital Human Modeling*, Advances in Human Factors and Ergonomics Series, pages 457–466, Boca Raton, Florida, 2010. CRC Press, Taylor & Francis Group. *4 citations pages 2, 20, 51, et 99*

Mohammad Ali Nazari, Pascal Perrier et Yohan Payan : The Distributed Lambda Model (DLM): A 3-D, Finite-Element Muscle Model Based on Feldman's Lambda Model; Assessment of Orofacial Gestures. *Journal of Speech, Language, and Hearing Research*, 56(6):1909–1923, 2013. *Cité page 92*

Daniel Neiberg, Gopal Ananthakrishnan et Mats Blomberg : On acquiring speech production knowledge from articulatory measurements for phoneme recognition. *In Proceedings of Interspeech*, pages 1387–1390. Citeseer, 2009. *Cité page 29*

Alfonso Nieto-Castanon, Frank Guenther, Joseph Perkell et Hugh Curtin : A modeling investigation of articulatory variability and acoustic stability during american english/r/production. *The Journal of the Acoustical Society of America*, 117(5):3196–3212, 2005. *2 citations pages 93 et 94*

Sven Öhman : Coarticulation in VCV utterances : Spectrographic measurements. *The Journal of the Acoustical Society of America*, 39(1):151–168, 1966. *6 citations pages 10, 97, 103, 110, 112, et 120*

Kimbrough Oller, Rebecca Eilers, Rebecca Neal et Heidi Schwartz : Precursors to speech in infancy: the prediction of speech and language disorders. *Journal of Communication Disorders*, 32(4):223–245, 1999. *Cité page 32*

Darryl Ong et Maureen Stone : Three-dimensional vocal tract shapes in/r/and/l/: A study of MRI, ultrasound, electropalatography, and acoustics. *Phonoscope*, 1(1):1–13, 1998. *Cité page 15*

Mari Ostendorf : Moving beyond the 'beads-on-a-string' model of speech. *In Proceedings of the IEEE ASRU Workshop*, pages 79–84, 1999. *Cité page 28*

Slim Ouni : *Modélisation de l'espace articulatoire par un codebook hypercubique pour l'inversion acoustico-articulatoire.* Thèse de doctorat, université Henri Poincaré – Nancy 1, 2001. *Cité page 95*

Slim Ouni, Yves Laprie *et al.* : A study of the French vowels through the main constriction of the vocal tract using an acoustic-to-articulatory inversion method. *In 15th International Congress of Phonetic Sciences (ICPhS)*. Citeseer, 2003. *Cité page 157*

Yohan Payan et Pascal Perrier : Synthesis of VV sequences with a 2D biomechanical tongue model controlled by the equilibrium point hypothesis. *Speech communication*, 22(2):185–205, 1997. *Cité page 92*

Sharon Peperkamp et Emmanuel Dupoux : Learning the mapping from surface to underlying representations in an artificial language. *Laboratory phonology*, 9:315–338, 2007. *Cité page 159*

Amy Perfors, Joshua Tenenbaum, Thomas Griffiths et Fei Xu : A tutorial introduction to bayesian models of cognitive development. *Cognition*, 120(3):302–321, 2011. *Cité page 35*

Joseph Perkell : *A physiologically-oriented model of tongue activity in speech production.* Thèse de doctorat, Massachusetts Institute of Technology., 1974. *Cité page 92*

Joseph Perkell, Frank Guenther, Harlan Lane, Melanie Matthies, Ellen Stockmann, Mark Tiede et Majid Zandipour : The distinctness of speakers' productions of vowel contrasts is related to their discrimination of the contrasts. *The Journal of the Acoustical Society of America*, 116(4):2338–2344, 2004a. *2 citations pages 93 et 94*

Joseph Perkell, Melanie Matthies, Harlan Lane, Frank Guenther, Reiner Wilhelms-Tricarico, Jane Wozniak et Peter Guiod : Speech motor control: Acoustic goals, saturation effects, auditory feedback and internal models. *Speech communication*, 22(2):227–250, 1997. *2 citations pages 13 et 14*

Joseph Perkell, Melanie Matthies et Mario Svirsky : Articulatory evidence for acoustic goals for consonants. *The Journal of the Acoustical Society of America*, 96:3326, 1994. *Cité page 14*

Joseph Perkell, Melanie Matthies, Mario Svirsky et Michael Jordan : Trading relations between tongue-body raising and lip rounding in production of the vowel/u: A pilot "motor equivalence"study. *The Journal of the Acoustical Society of America*, 93:2948, 1993. *Cité page 14*

Joseph Perkell, Melanie Matthies, Mark Tiede, Harlan Lane, Majid Zandipour, Nicole Marrone, Ellen Stockmann et Frank Guenther : The distinctness of speakers' /s/–/ʃ/

contrast is related to their auditory discrimination and use of an articulatory saturation effect. *Journal of Speech, Language, and Hearing Research*, 47(6), 2004b. *2 citations pages 93 et 94*

Joseph S Perkell : *Physiology of speech production: Results and implications of a quantitative cineradiographic study.* Numéro 53. MIT Press, 1969. *Cité page 91*

Pascal Perrier : Control and representations in speech production. *ZAS Papers in Lingustics*, 40:109–132, 2005. *2 citations pages 14 et 17*

Pascal Perrier : Gesture planning integrating knowledge of the motor plant's dynamics: A literature review from motor control and speech motor control. *In* Daniel Pape & Pascal Perrier Susanne Fuchs, Melanie Weirich, éditeur : *Speech Planning and Dynamics*, Speech Production and Perception, pages 191–238. Peter Lang Publishers, 2012. *2 citations pages 22 et 23*

Pascal Perrier, Louis-Jean Boë et Rudolph Sock : Vocal tract area function estimation from midsagittal dimensions with ct scans and a vocal tract castmodeling the transition with two sets of coefficients. *Journal of Speech, Language, and Hearing Research*, 35(1):53–67, 1992. *Cité page 93*

Pascal Perrier et Liang Ma : Speech planning for V1CV2 sequences: Influence of the planned sequence. *In Proceedings of the 8th International Seminar on Speech Production, ISSP8*, pages 69–72, 2008. *Cité page 160*

Pascal Perrier, Yohan Payan, Stéphanie Isabelle Buchaillard, Mohammad Ali Nazari, Matthieu Chabanas *et al.* : Biomechanical models to study speech. *Faits de Langues*, 37:155–171, 2011. *Cité page 92*

Pascal Perrier, Yohan Payan, Majid Zandipour et Joseph Perkell : Influences of tongue biomechanics on speech movements during the production of velar stop consonants: A modeling study. *The Journal of the Acoustical Society of America*, 114(3):1582–1599, 2003. *Cité page 92*

Martin Pickering et Simon Garrod : An integrated theory of language production and comprehension. *Behavioral and Brain Sciences*, 36(04):329–347, 2013. *Cité page 29*

David Pisoni : Identification and discrimination of the relative onset time of two component tones: implications for voicing perception in stops. *The Journal of the Acoustical Society of America*, 61:1352, 1977. *Cité page 13*

David Poeppel, William Idsardi et Virginie van Wassenhove : Speech perception at the interface of neurobiology and linguistics. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1493):1071–1086, 2008. *Cité page 23*

David Poeppel et Philip Monahan : Feedforward and feedback in speech perception: Revisiting analysis by synthesis. *Language and Cognitive Processes*, 26(7):935–951, 2011. *Cité page 23*

Louis Pols : Analysis and synthesis of speech using a broad-band spectral representation. *Auditory analysis and perception of speech*, pages 23–36, 1975. *Cité page 91*

Ferran Pons, David Lewkowicz, Salvador Soto-Faraco et Núria Sebastián-Gallés : Narrowing of intersensory speech perception in infancy. *Proceedings of the National Academy of Sciences*, 106(26):10598–10602, 2009. *Cité page 160*

R. Porter et J. Lubker : Rapid reproduction of vowel-vowel sequences : Evidence for a fast and direct acoustic-motoric linkage in speech. *Journal of Speech, Language, and Hearing Research*, 23(3):593–602, 1980. *Cité page 20*

Gerasimos Potamianos, Chalapathy Neti, Juergen Luettin et Iain Matthews : Audiovisual automatic speech recognition. *In* Gerard Bailly, Pascal Perrier et Eric Vatikiotis-Bateson, éditeurs : *Audiovisual Speech Processing.* Cambridge University Press, 2012. *2 citations pages 28 et 29*

William Powers : *Behavior: The control of perception.* Aldine de Gruyter, Chicago, 1973. *Cité page 33*

Friedemann Pulvermüller : Brain mechanisms linking language and action. *Nature Reviews Neuroscience*, 6(7):576–582, 2005. *Cité page 33*

Friedemann Pulvermüller et Luciano Fadiga : Active perception: sensorimotor circuits as a cortical basis for language. *Nature Reviews Neuroscience*, 11(5):351–360, 2010. *Cité page 21*

Friedemann Pulvermüller, Martina Huss, Ferath Kherif, Fermin Moscoso del Prado Martin, Olaf Hauk et Yury Shtyrov : Motor cortex maps articulatory features of speech sounds. *Proceedings of the National Academy of Sciences*, 103(20):7865 –7870, 2006. *Cité page 21*

Lawrence Rabiner : A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989. *Cité page 26*

Lucile Rapin : *Hallucinations auditives verbales et trouble du langage intérieur dans la schizophrénie: traces physiologiques et bases cérébrales.* Thèse de doctorat, Université de Grenoble, 2011. *Cité page 39*

Bruno Repp : Trading relations among acoustic cues in speech perception are largely a result of phonetic categorization. *Speech Communication*, 2(4):341–361, 1983. *Cité page 20*

Bruno Repp : Categorical perception: Issues, methods, findings. *Speech and language: Advances in basic research and practice*, 10:243–335, 1984. *Cité page 20*

Korin Richmond, Phil Hoole et Simon King : Announcing the electromagnetic articulography (day 1) subset of the mngu0 articulatory corpus. *In Proceedings of Interspeech*, pages 1505–1508, 2011. *Cité page 31*

Giacomo Rizzolatti et Michael Arbib : Language within our grasp. *Trends in neurosciences*, 21(5):188–194, 1998. *2 citations pages 24 et 33*

Giacomo Rizzolatti, Luciano Fadiga, Vittorio Gallese et Leonardo Fogassi : Premotor cortex and the recognition of motor actions. *Cognitive Brain Research*, 3(2):131–141, 1996a. *Cité page 21*

Giacomo Rizzolatti, Luciano Fadiga, Massimo Matelli, Valentino Bettinardi, Eraldo Paulesu, Daniela Perani et Ferruccio Fazio : Localization of grasp representations in humans by PET: 1. observation versus execution. *Experimental brain research*, 111(2):246–252, 1996b. *Cité page 21*

Matthias Rolf, Jochen J Steil et Michael Gienger : Goal babbling permits direct learning of inverse kinematics. *Autonomous Mental Development, IEEE Transactions on*, 2(3):216–229, 2010. *Cité page 69*

Raymond Romand : Introduction au fonctionnement du système auditif. *In* Pierre Escudier et Jean-Luc Schwartz, éditeurs : *La parole, des modèles cognitifs aux machines communicantes*, pages 101–133. Hermès, 2000. *2 citations pages 88 et 89*

Richard Rose, Juergen Schroeter et Man Mohan Sondhi : An investigation of the potential role of speech production models in automatic speech recognition. *In International Conference on Spoken Language Processing (ICSLP)*, 1994. *Cité page 28*

Philip Rubin, Thomas Baer et Paul Mermelstein : An articulatory synthesizer for perceptual research. *The Journal of the Acoustical Society of America*, 70(2):321–328, 1981. *Cité page 10*

Frank Rudzicz : Correcting errors in speech recognition with articulatory dynamics. *In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 60–68. Association for Computational Linguistics, 2010a. *Cité page 27*

Frank Rudzicz : Learning mixed acoustic/articulatory models for disabled speech. *In Proceedings of the Workshop on Machine Learning for Assistive Technologies at the 24th annual conference on Neural Information Processing Systems (NIPS)*, pages 70–78, 2010b. *Cité page 27*

Frank Rudzicz : Articulatory knowledge in the recognition of dysarthric speech. *IEEE Transactions on Audio, Speech and Language Processing*, 19(4):947–960, 2011a. *Cité page 27*

Frank Rudzicz : *Production knowledge in the recognition of dysarthric speech.* Thèse de doctorat, University of Toronto, 2011b. *Cité page 27*

Haşim Sak, Andrew Senior et Françoise Beaufays : Long Short-Term Memory based Recurrent Neural Network architectures for large vocabulary speech recognition. *arXiv preprint arXiv:1402.1128*, 2014. *Cité page 29*

Elliot Saltzman et Scott Kelso : Skilled actions: A task-dynamic approach. *Psychological Review*, 94(1):84–106, 1987. *Cité page 10*

Vittorio Sanguineti, Rafael Laboissière et David Ostry : A dynamic biomechanical model for neural control of speech production. *The Journal of the Acoustical Society of America*, 103(3):1615–1627, 1998. *Cité page 92*

Marc Sato, Krystyna Grabski, Arthur Glenberg, Amelie Brisebois, Anahita Basirat, Lucie Menard et Luigi Cattaneo : Articulatory bias in speech categorization: Evidence from use-induced motor plasticity. *cortex*, 47(8):1001–1003, 2011. *Cité page 21*

Marc Sato, Pascale Tremblay et Vincent Gracco : A mediating role of the premotor cortex in phoneme segmentation. *Brain and language*, 111(1):1–7, 2009. *Cité page 21*

Christophe Savariaux, Pascal Perrier et Jean-Pierre Orliaguet : Compensation strategies for the perturbation of the rounded vowel [u] using a lip tube: A study of the control space in speech production. *The Journal of the Acoustical Society of America*, 98(5):2428–2442, 1995. *Cité page 14*

Christophe Savariaux, Pascal Perrier, Jean-Pierre Orliaguet et Jean-Luc Schwartz : Compensation strategies for the perturbation of french [u] using a lip tube. II. perceptual analysis. *The Journal of the Acoustical Society of America*, 106(1):381–393, 1999. *Cité page 14*

Odette Scharenborg : Reaching over the gap: A review of efforts to link human and automatic speech recognition research. *Speech Communication*, 49(5):336–347, 2007. *Cité page 29*

Odette Scharenborg, Vincent Wan et Roger Moore : Towards capturing fine phonetic variation in speech using articulatory features. *Speech Communication*, 49(10):811–826, 2007. *Cité page 28*

Otto Schmidbauer : Robust statistic modelling of systematic variabilities in continuous speech incorporating acoustic-articulatory relations. *In Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pages 616–619. IEEE, 1989. *Cité page 28*

Jean Schoentgen et Sorin Ciocea : Direct calculation of the vocal tract area function from measured formant frequencies. *In EUROSPEECH*, 1995. *Cité page 91*

Manfred Schroeder, Bishnu Atal et J.L. Hall : Objective measure of certain speech signal degradations based on masking properties of human auditory perception. *Frontiers of speech communication research*, pages 217–229, 1979. *Cité page 89*

Jean-Luc Schwartz, Christian Abry, Louis-Jean Boë et Marie Cathiard : Phonology in a theory of perception-for-action-control. *Phonetics, phonology and cognition*, pages 255–280, 2002a. *2 citations pages 18 et 19*

Jean-Luc Schwartz, Anahita Basirat, Lucie Ménard et Marc Sato : The Perception-for-Action-Control Theory (PACT): A perceptuo-motor theory of speech perception. *Journal of Neurolinguistics*, 25(5):336–354, 2012a. *4 citations pages 18, 19, 20, et 51*

Jean-Luc Schwartz, Denis Beautemps, Yonnel Arrouas et Pierre Escudier : Auditory analysis of speech gestures. *The Auditory Processing of Speech*, pages 239–252, 1992. *2 citations pages 89 et 90*

Jean-Luc Schwartz, Louis-Jean Boë et Christian Abry : Linking dispersion-focalization theory and the maximum utilization of the available distinctive features principle in a perception-for-action-control theory. *Experimental approaches to phonology*, pages 104–124, 2007. *2 citations pages 18 et 19*

Jean-Luc Schwartz, Louis-Jean Boë, Pierre Badin et Thomas Sawallis : Grounding stop place systems in the perceptuo-motor substance of speech: On the universality of the labial-coronal-velar stop series. *Journal of Phonetics*, 40:20–36, 2012b. *3 citations pages 94, 105, et 108*

Jean-Luc Schwartz, P Teissier et Pierre Escudier : La parole multimodale : deux ou trois sens valent mieux qu'un. *In* Joseph Mariani, éditeur : *Traitement automatique du langage parlé 2: reconnaissance de la parole*, pages 141–178. Hermès, 2002b. *Cité page 29*

Jihène Serkhane, Jean-Luc Schwartz et Pierre Bessière : Simulating vocal imitation in infants, using a growth articulatory model and speech robotics. *In Proceedings of the International Congress of Phonetic Sciences*, pages 2241–2245, 2003. *Cité page 94*

Jihène Serkhane : *Un bébé androïde vocalisant: Etude et modélisation des mécanismes d'exploration vocale et d'imitation orofaciale dans le développement de la parole.* Thèse de doctorat, Institut Polytechnique de Grenoble, 2005. *Cité page 90*

Jihène Serkhane, Jean-Luc Schwartz, Louis-Jean Boë, Barbara Davis et Christine Matyear : Infants' vocalizations analyzed with an articulatory model: A preliminary report. *Journal of Phonetics*, 35(3):321–340, 2007. *Cité page 94*

Kristina Simonyan et Barry Horwitz : Laryngeal motor cortex and control of speech in humans. *The Neuroscientist*, 17(2):197–208, 2011. *Cité page 91*

Jeremy Skipper, Howard Nusbaum et Steven Small : Lending a helping hand to hearing: Another motor theory of speech perception. *Action to language via the mirror neuron system*, pages 250–285, 2006. *Cité page 18*

Jeremy Skipper, Virginie van Wassenhove, Howard Nusbaum et Steven Small : Hearing lips and seeing voices: How cortical areas supporting speech production mediate audiovisual speech perception. *Cerebral Cortex*, page bhl147, 2007. *Cité page 18*

Hagen Soltau, Florian Metze et Alex Waibel : Compensating for hyperarticulation by modeling articulatory properties. *In Proceedings of the International Conference on Speech and Language Processing*, pages 841–844, 2002. *Cité page 27*

Roger Sperry : Neural basis of the spontaneous optokinetic response produced by visual inversion. *Journal of comparative and physiological psychology*, 43(6):482, 1950. *Cité page 18*

Ian Stavness, Mohammad Nazari, Flynn Cormac, Pascal Perrier, Yohan Payan, John Lloyd et Sidney Fels : Coupled Biomechanical Modeling of the Face, Jaw, Skull, Tongue, and Hyoid Bone. *In* Nadia Magnenat-Thalmann, Osman Ratib et Hon Fai Choi, éditeurs : *3D Multiscale Physiological Human*, pages 253–274. Springer London, 2014. *Cité page 92*

Richard Stern et Nelson Morgan : Hearing is believing: Biologically-inspired feature extraction for robust automatic speech recognition. *IEEE Signal Processing Magazine*, 29(34-43):170, 2012. *Cité page 26*

Kenneth Stevens : Toward a model for speech recognition. *The Journal of the Acoustical Society of America*, 32(1):47–55, 1960. *2 citations pages 17 et 23*

Kenneth Stevens : The quantal nature of speech: Evidence from articulatory-acoustic data. *In* EE David et PB Denes, éditeurs : *Human communication: A unified view*, pages 51–66. McGraw-Hill, 1972. *3 citations pages 19, 60, et 158*

Kenneth Stevens : Acoustic correlates of some phonetic categories. *The Journal of the Acoustical Society of America*, 68(3):836–842, 1980. *2 citations pages 98 et 142*

Kenneth Stevens : On the quantal nature of speech. *Journal of Phonetics*, 17(1):3–45, 1989. *2 citations pages 60 et 65*

Kenneth Stevens et Sheila Blumstein : Invariant cues for place of articulation in stop consonants. *The Journal of the Acoustical Society of America*, 64(5):1358–1368, 1978. *Cité page 98*

Kenneth Stevens et Morris Halle : Remarks on analysis by synthesis and distinctive features. *Models for the perception of speech and visual form*, pages 88–102, 1967. *Cité page 17*

Kenneth Stevens et Arthur House : Development of a quantitative description of vowel articulation. *The Journal of the Acoustical Society of America*, 27(3):484–493, 1955. *Cité page 91*

Kenneth Stevens et Dennis Klatt : Role of formant transitions in the voiced-voiceless distinction for stops. *The Journal of the Acoustical Society of America*, 55:653, 1974. *Cité page 13*

Sebastian Stüker, Tanja Schultz, Florian Metze et Alex Waibel : Multilingual articulatory features. *In Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages 144–147. IEEE, 2003. *Cité page 28*

Jiping Sun, Xing Jing et Li Deng : Data-driven model construction for continuous speech recognition using overlapping articulatory features. *In Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP)*, volume 1, pages 437–440, 2000. *Cité page 28*

Harvey Sussman, David Fruchter, Jon Hilbert et Joseph Sirosh : Linear correlates in the speech signal: The orderly output constraint. *Behavioral and brain sciences*, 21(02):241–259, 1998. *6 citations pages 97, 99, 100, 101, 107, et 109*

Harvey Sussman, Helen McCaffrey et Sandra Matthews : An investigation of locus equations as a source of relational invariance for stop place categorization. *The Journal of the Acoustical Society of America*, 90(3):1309–1325, 1991. *3 citations pages 99, 107, et 142*

Mark Tiede, Hani Yehia et Eric Vatikiotis-Bateson : A shape-based approach to vocal tract area function estimation. *In 4th Speech Production Seminar/ETRW*, pages 41–44, 1996. *Cité page 92*

Michael Tomasello, Malinda Carpenter, Josep Call, Tanya Behne et Henrike Moll : Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences*, 28(5):675–691, 2005. *Cité page 24*

George Trager et Bernard Bloch : The syllabic phonemes of english. *Language*, pages 223–246, 1941. *Cité page 10*

Gautam Vallabha, James McClelland, Ferran Pons, Janet Werker et Shigeaki Amano : Unsupervised learning of vowel categories from infant-directed speech. *Proceedings of the National Academy of Sciences*, 104(33):13273–13278, 2007. *Cité page 159*

Andrew Viterbi : Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *Information Theory, IEEE Transactions on*, 13(2):260–269, 1967. *Cité page 26*

Erich Von Holst : Relations between the central nervous system and the peripheral organs. *The British Journal of Animal Behaviour*, 2(3):89–94, 1954. *Cité page 18*

Kate Watkins, Antonio Strafella et Tomas Paus : Seeing and hearing speech excites the motor system involved in speech production. *Neuropsychologia*, 41(8):989–994, 2003. *Cité page 21*

John Westbury, Michiko Hashi et Mary Lindstrom : Differences among speakers in lingual articulation for american english /r/. *Speech Communication*, 26(3):203–226, 1998. *Cité page 15*

Margaret Wilson et Günther Knoblich : The case for motor involvement in perceiving conspecifics. *Psychological bulletin*, 131(3):460, 2005. *Cité page 33*

Stephen Wilson, Ayşe Pinar Saygin, Martin Sereno et Marco Iacoboni : Listening to speech activates motor areas involved in speech production. *Nature neuroscience*, 7(7):701–702, 2004. *Cité page 21*

Ralf Winkler, Liang Ma et Pascal Perrier : A model of optimal speech production planning integrating dynamical constraints to achieve appropriate articulatory timing. *In* Susanne Fuchs, Melanie Weirich, Daniel Pape et Pascal Perrier, éditeurs : *Proceedings of Cognitive and Physical Models of Speech Production (CPMSP2)*, pages 44–48, Berlin, Allemagne, 2010. *Cité page 110*

Daniel Wolpert, Kenji Doya et Mitsuo Kawato : A unifying computational framework for motor control and social interaction. *Philosophical Transactions of the Royal Society of London*, 358(1431):593–602, 2003. *Cité page 24*

Alan Wrench : The MOCHA-TIMIT articulatory database. Rapport technique, University of Edinburgh, 1999. URL `http://www.cstr.ed.ac.uk/research/projects/artic/mocha.html`. *Cité page 31*

Zong Liang Wu, Jean-Luc Schwartz et Pierre Escudier : Physiologically plausible modules for the detection of articulatory-acoustic events. *Advances in Speech, Hearing and Language Processing*, 3:479–495, 1996. *Cité page 89*

Eric Young et Murray Sachs : Representation of steady-state vowels in the temporal aspects of the discharge patterns of populations of auditory-nerve fibers. *The Journal of the Acoustical Society of America*, 66(5):1381–1403, 1979. *Cité page 90*

Victoria Young et Alex Mihailidis : Difficulties in automatic speech recognition of dysarthric speakers and implications for speech-based applications used by the elderly: A literature review. *Assistive Technology*, 22(2):99–112, 2010. *Cité page 27*

Fang Zheng, Guoliang Zhang et Zhanjiang Song : Comparison of different implementations of MFCC. *Journal of Computer Science and Technology*, 16(6):582–589, 2001. *Cité page 26*

Olek Zienkiewicz et Robert Taylor : *The Finite Element Method: Its Basis And Fundamentals.* Butterworth-Heinemann, 6ème édition, 2005. *Cité page 92*

Geoffrey Zweig : *Speech Recognition with Dynamic Bayesian Networks.* Thèse de doctorat, University of Califormia, 1998. *2 citations pages 28 et 29*