Modèles bayésiens et sélection de modèles de perception pour les systèmes sensori-moteurs

Estelle Gilet

Ecole Doctorale
"Mathématique, sciences et technologie de l'information, informatique"

Master 2ème année de Mathématiques, Informatique
Spécialité *Intelligence, Interaction, Information*

Titre

Modèles bayésiens et sélection de modèles de perception pour les systèmes sensori-moteurs

Estelle GILET

le 19 Juin 2006

Laboratoire GRAVIR, équipe-projet e-Motion
Encadrants : Julien Diard et Pierre Bessière

Jury
C. Berrut
Y. Demazeau
J.-M Vincent
J.-L Schwartz

Table des matières

Chapitre 1

Introduction

Ce stage de MASTER 2 Recherche s'inscrit dans le domaine de la modélisation de la perception pour les systèmes sensori-moteurs. Ce domaine est central à la robotique. Dans ce cadre, la programmation bayésienne des robots est un outil largement répandu. Il permet notamment d'exprimer les processus de perception comme des problèmes inverses. L'utilisation de probabilités permet de plus de gérer le caractère mal-posé de ces problèmes, et de tenir compte de l'incomplétude des modèles en la transformant en incertitude.

Le formalisme de programmation bayésienne a été largement appliqué en robotique, et plus généralement en intelligence artificielle. En effet, ses bases mathématiques en font un bon outil de modélisation du raisonnement rationnel. Une question centrale au domaine consiste à déterminer s'il s'agit également d'un bon outil pour modéliser le raisonnement humain ou animal.

Nous nous intéressons plus spécifiquement à la façon dont des informations sur l'environnement provenant de capteurs différents peuvent être mélangées. Pour répondre à cette question, le modèle dominant, en robotique probabiliste, est la fusion capteur. Ce modèle, mathématiquement simple, consiste à fusionner chacun des capteurs pris indépendamment pour obtenir une information plus précise.

Ce modèle de fusion de capteurs robotique permet-il l'écriture d'algorithmes de fusion d'informations décrivant convenablement le principe de fusion d'informations sensorielles chez l'humain ?

De nombreux travaux ont déjà abordé cette question, en proposant l'application de modèles bayésiens au mélange d'informations chez l'humain, dans des cas intra-modaux (vision-vision ou haptique-haptique, par exemple), ou multi-modaux (visuo-haptique, visuo-acoustique pour la localisation de sources sonores, par exemple). Nous proposons de compléménter ces travaux par une application de la modélisation bayésienne au cas de fusion visuo-acoustique dans une tâche de perception des voyelles.

Plus précisément, dans notre travail, nous fixons deux objectifs.

D'une part, nous proposons plusieurs modèles bayésiens de la perception audiovisuelle des voyelles. Ces modèles sont basés sur des données expérimentales issues de la thèse de Robert-Ribes [13]. Les modèles utilisent tous le schéma de fusion capteur robotique, mais diffèrent sur la nature de l'espace interne dans lequel s'opère la fusion.

D'autre part, une fois les modèles réalisés, nous les comparons aux données expérimentales dans des cas de congruence audiovisuelle, et dans des cas de conflits audiovisuels [12]. Grâce à la modélisation bayésienne, nous réalisons ainsi une sélection de modèles, c'est à dire que nous comparons les modèles de façon quantitative. Cette démarche est originale dans le domaine, où la plupart des travaux précédents se contentent de définir un modèle unique. Ce modèle est alors validé s'il est plus performant qu'un modèle ne contenant aucune connaissance sur le domaine.

Ce document est structuré de la façon suivante : le chapitre 2 fait une brève description du formalisme de la programmation bayésienne des robots avant de présenter un modèle bayésien de perception en robotique : le modèle capteur. Puis, nous faisons une étude sur l'application des modèles bayésiens dans les domaines de la perception humaine. Dans le chapitre 3 sont présentés les trois modèles de perception audiovisuelle que nous avons réalisés. Le chapitre 4 propose une méthode de sélection de modèles que nous avons appliqué dans deux cas, des cas de congruence audiovisuelle et des cas de conflits. Dans le chapitre 5, nous présentons une analyse et une interprétation des résultats obtenus. Enfin, le chapitre 6 propose une synthèse du travail réalisé ainsi que des perspectives de recherche.

Chapitre 2

Etude bibliographique

Dans ce chapitre, nous détaillons tout d'abord la programmation Bayésienne des robots (notée PBR), puis nous proposerons des modèles informatiques de perception robotique avant de présenter des exemples de modélisation de perception humaine.

2.1 Notion de programmation bayésienne

Dans cette section, nous allons passer en revue les différentes étapes de la programmation bayésienne des robots [11] qui sont schématisées Figure 2.1. Cette section est adaptée de [15].

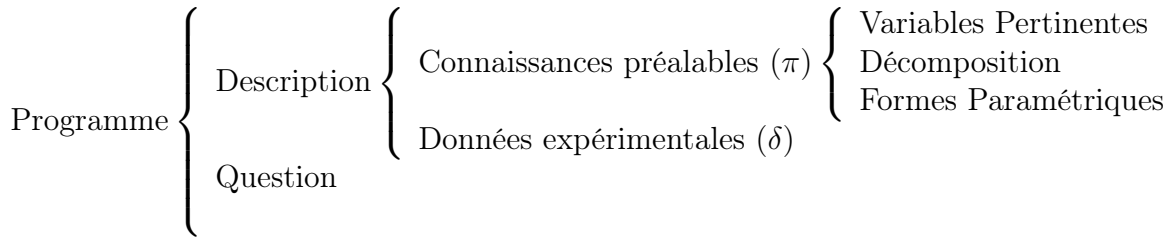

FIG. 2.1 – Structure d'un programme PBR.

Le programme : Le programme est composé d'une phase description et d'une phase question.

La description : La description est dénotée formellement par la distribution de probabilité conjointe $P(V_1 \ . \ . \ . \ V_n \mid \delta \ \pi)$ d'un ensemble de variables $V_1, \ . \ . \ . \ , V_n$. Elle est déterminée au vu des connaissances préalables π spécifiées par le programmeur et d'un ensemble de données expérimentales δ. Ainsi, l'ensemble des connaissances préalables devant être fournies se trouve circonscrit par les connaissances préalables π et le jeu de données δ.

Les connaissances préalables : Les connaissances préalables sont déterminées au cours de la spécification qui est la partie la plus délicate du travail du programmeur. Au cours de cette phase, il doit énoncer clairement et explicitement les connaissances dont il est à l'origine et celles qui résultent d'un processus adaptatif dépendant d'un jeu particulier de données expérimentales. Ces connaissances se subdivisent en trois : le choix des variables pertinentes, l'expression des dépendances entre les variables retenues sous la forme d'un produit de distributions élémentaires, et enfin la forme paramétrique associée à chacune de ces distributions.

Les variables pertinentes : Cette phase consiste à définir l'ensemble des variables $V_1, \ldots, V_n$ apparaissant dans le programme et de spécifier pour chacune d'elles son domaine de variation D_{v_i} et son nombre k_{v_i} d'états possibles. Toutes les autres variables sont ainsi supposées non pertinentes pour le problème considéré.

La décomposition : Comme énoncé précédemment, la description sur les variables $V_1, \ldots, V_n$ a pour but la définition de la distribution conjointe $P(V_1 \ldots V_n \mid \delta\ \pi)$. Ce terme mathématique est une distribution de probabilité sur n dimensions, qui n'est souvent pas facile à spécifier. La règle du produit nous permet de décomposer cette expression, en l'exprimant sous forme de produits de distributions. Une fois la décomposition choisie, une seconde étape permet de simplifier d'avantage l'expression par des hypothèses d'indépendance conditionnelle, ce qui réduit fortement les dimensions des termes sur lesquelles ces hypothèses portent.

Une forme paramétrique : Pour rendre la distribution conjointe effectivement calculable, une forme paramétrique est associée à chacun des termes apparaissant dans la décomposition choisie. Ces choix définissent des a priori sur les valeurs des distributions de probabilité et la manière dont ces valeurs seront modifiées, éventuellement par l'expérience. Quelques formes paramétriques simples sont la loi uniforme, la loi dirac, la loi normale ou gaussienne et la loi de succession de Laplace. Les définitions des formes paramétriques utilisées dans ce travail sont rappelées Annexe A (tirée de [4]).

Données expérimentales : Des données expérimentales représentées par le symbole δ peuvent être nécessaires pour fournir les valeurs des paramètres de certaines formes paramétriques. Par exemple, dans le cas de gaussiennes, nous pouvons tirer de ces données leurs moyennes et écarts type. Cette phase est la phase d'identification ou encore la période d'apprentissage.

Question : Pour la question, les variables $V_1, \ldots, V_n$ sont divisées en trois sous-ensembles : celles dont nous cherchons les valeurs (Cherchées), celles dont nous connaissons les valeurs (Connues) et celles dont nous ne connaissons pas les valeurs sans pour autant les chercher (Inconnues). Répondre à une question consiste donc à calculer le terme :

$$P(\text{Cherchées} \mid \text{Connues}\ \delta\ \pi).$$

2.2 Modèles de perception

2.2.1 Modèles informatiques de perception robotique

Pour percevoir son environnement, un système sensori-moteur, qu'il soit naturel ou artificiel, utilise de multiples sources d'informations sensorielles, comme par exemple la vision, le toucher, l'ouïe... Toutes ces informations doivent être fusionnées afin de fournir un percept robuste et cohérent [7].

La Figure 2.2 propose un schéma présentant la perception comme un problème inverse : soit φ une variable d'environnement, c'est-à-dire une variable caractérisant une ou plusieurs propriétés de l'environnement, soient S_1 et S_2 les variables de perception correspondant à la valeur fournie par chacun des capteurs du robot et soit A une variable d'action permettant au robot d'interagir avec son environnement.

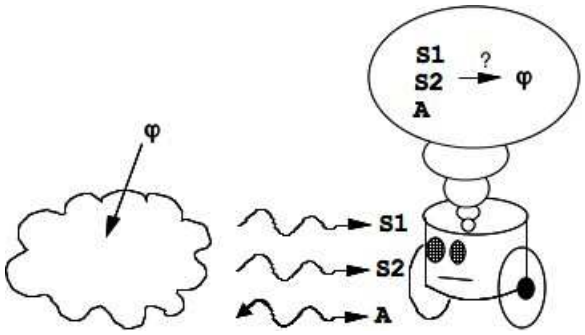

FIG. 2.2 – La perception vue comme un problème inverse.

Les connaissances du robot lui donnent la capacité de prédire l'effet de φ supposé connu, sur les variables sensori-motrices S_1, S_2 et A. Par exemple, sachant la position d'un obstacle, nous pourrons prédire la réponse d'un capteur laser de mesure de proximité. En termes mathématiques, ceci correspond au modèle :

$$P(S_1\ S_2\ A\ \varphi) \quad = \quad P(S_1\ S_2\ A \mid \varphi)P(\varphi).$$

La perception consiste, étant donné ce modèle à déterminer la valeur de φ connaissant les valeurs de S_1, S_2 et A. Nous nous intéressons donc à calculer $P(\varphi \mid S_1\ S_2\ A)$, alors que le modèle donné au robot est basé sur $P(S_1\ S_2\ A \mid \varphi)$: la perception est donc ici vue comme un problème d'inversion. En effet, elle s'attache à calculer l'antécédent d'un élément par une fonction donnée. Cependant, il n'est pas toujours possible d'inverser analytiquement les fonctions directes d'observation. A l'aide de la modélisation par des distributions en probabilité et la règle de Bayes, un problème inverse peut être résolu :

$$\begin{aligned} P(\varphi \mid S_1\ S_2\ A) &= \frac{P(\varphi\ S_1\ S_2\ A)}{P(S_1\ S_2\ A)} \\ &= \frac{1}{Z_1} P(S_1\ S_2\ A \mid \varphi)P(\varphi). \end{aligned}$$

En bayésien, les connaissances d'un sujet confronté à une expérience sont modélisées à l'aide de probabilités. L'utilisation de probabilité permet de résoudre des problèmes mal posés et de tenir compte de l'incomplétude en la transformant en incertitude. Les problèmes mathématiques inverses peuvent également être mal posés : un problème mal posé est un problème pour lequel il n'existe pas une unique solution. Ceci peut être un problème dans une formulation déterministe. Dans une écriture probabiliste, au contraire, s'il existe plusieurs solutions au problème d'inversion, la probabilité résultante aura plusieurs pics de probabilités. En pratique, plusieurs configurations de l'environnement peuvent correspondre à une même observation. A partir d'une telle observation, la perception ne peut donc pas discriminer les différentes possibilités. La perception est donc généralement un problème mal posé.

Dans le reste de ce travail, nous nous plaçons donc dans le cadre de modèles probabilistes de perception, en raison de leur capacité à traiter les problèmes inverses et mal posés.

2.2.2 Modèle de fusion capteur

Ce cadre de travail est largement répandu en robotique. Très couramment, les aspects sensoriels et moteurs sont découplés :

$$P(S_1\ S_2\ A \mid \varphi) \quad = \quad P(S_1\ S_2 \mid \varphi) P(A \mid \varphi).$$

Le terme $P(S_1\ S_2 \mid \varphi)$ est purement sensoriel et le terme $P(A \mid \varphi)$ est purement moteur. Nous nous intéressons ici au premier terme $P(S_1\ S_2 \mid \varphi)$.

Le principe de la fusion de capteur, détaillé dans cette section, est le suivant : pour chaque capteur, une description décrivant le modèle de la réponse du capteur connaissant son environnement est construite. Dans une deuxième phase, ces modèles associés à chacun des capteurs sont fusionnés afin de déterminer les caractéristiques de l'environnement par rapport au robot.

Le terme $P(S_1\ S_2 \mid \varphi)$ est souvent défini par :

$$P(S_1\ S_2\ \varphi) = P(S_1 \mid \varphi) P(S_2 \mid \varphi).$$

Cette décomposition introduit une hypothèse d'indépendance conditionnelle : nous faisons l'hypothèse que les valeurs des capteurs sont indépendantes les unes des autres connaissant la valeur de la variable φ. Ce modèle est appelé modèle de fusion capteur [11].

La question posée généralement dans le cas de fusion de capteur est :

$$P(\varphi \mid [S_1 = s_1]\ [S_2 = s_2]).$$

Cette question est équivalente à quelle est l'environnement sachant que la valeur du capteur S_1 est égale à s_1, que la valeur du capteur S_2 est égale à s_2. L'inférence bayésienne

permet de répondre à cette question à partir de la décomposition.

$$
\begin{aligned}
& P(\varphi \mid [S_1 = s_1]\ [S_2 = s_2]) \\
&= \frac{P(\varphi\ [S_1 = s_1]\ [S_2 = s_2])}{P([S_1 = s_1])P([S_2 = s_2]))} && (2.1) \\
&= \frac{P([S_1 = s_1] \mid \varphi)P([S_2 = s_2] \mid \varphi)P(\varphi)}{P([S_1 = s_1])\ P([S_2 = s_2]))} && (2.2) \\
&= \frac{1}{Z_1} P([S_1 = s_1] \mid \varphi)P([S_2 = s_2] \mid \varphi)P(\varphi) && (2.3)
\end{aligned}
$$

Dans la dérivation ci-dessus, l'égalité (2.1) est obtenue par application de la règle de Bayes et l'égalité (2.2) en remplaçant le numérateur par la décomposition présentée précédemment. Le passage de l'égalité (2.2) à l'égalité (2.3) est justifié par le fait que le dénominateur ayant une valeur fixe peut être noté en une constante de normalisation $\frac{1}{Z_1}$.

Lorsque le terme $P(\varphi)$ est associé à une distribution de probabilité de loi uniforme, il est possible d'obtenir l'équivalence suivante :

$$
\begin{aligned}
& P(\varphi \mid [S_1 = s_1]\ [S_2 = s_2]) \\
&= \frac{1}{Z_2} P([S_1 = s_1] \mid \varphi)P([S_2 = s_2] \mid \varphi). && (2.4)
\end{aligned}
$$

En effet, l'égalité (2.4) est obtenue par inclusion du terme $P(\varphi)$ dans la constante de normalisation. Etant donné que ce terme suit une distribution de probabilité de loi uniforme sa valeur ne varie pas.

Sous cette hypothèse, l'estimation de φ au vu de S_1, S_2 prend le nom d'estimation de maximum de vraisemblance.

La fusion capteur telle que nous venons de la voir offre des avantages multiples : elle permet un rehaussement très important du signal lorsque les capteurs sont en accord. Elle est robuste aux bruits ou à la défaillance de certains capteurs (l'information fournie par chacun des capteurs pris isolément est souvent très médiocre). De plus, elle permet à partir de plusieurs capteurs de piètre qualité, d'obtenir par fusion une information plus précise.

2.2.3 Problématique

Chez l'être humain, la sensation est à la première étape de la chaîne allant, du stimulus des organes sensoriels à la perception. C'est un phénomène psychophysiologique par lequel une stimulation externe a un effet modificateur spécifique (sens) sur l'être vivant et conscient (ce phénomène correspond à la modification des variables S_1 et S_2 dans notre modèle de fusion capteur). La perception, vue comme un problème inverse, est la fonction par laquelle l'esprit se représente l'environnement (variable φ).
Notre problématique :

Le modèle probabiliste robotique de fusion capteur est-il un bon modèle de perception humaine ?

2.3 Modélisation de la perception humaine

Même si l'application du formalisme bayésien est peu courante dans l'étude des interactions multi-sensorielles, c'est un domaine en pleine expansion. Dans cette section, nous avons choisi de présenter différents travaux portant sur la modélisation de fusions sensorielles, comme par exemple les interactions entre la vision et la vision, la vision et le toucher ou encore la vision et l'audition. Dans un premier temps, nous étudions des cas où une seule modalité est en jeu puis, nous nous intéressons à des cas de fusions où deux modalités contribuent à la perception.

1 - Intra-modal : haptique

Le toucher est une modalité sensorielle intéressante parce que les sujets peuvent contrôler le mouvement de leur doigts pour obtenir différentes informations sur l'environnement. Par exemple, nous ne bougeons pas ses mains de la même manière lorsque l'on veut déterminer la texture de l'objet que lorsque l'on veut définir sa taille. Une étude [5] sur la fusion des données sensorielles *force* et *position* pour déterminer la forme d'un objet a montré que le modèle de maximum de vraisemblance permettait de décrire l'intégration intra-modale du toucher : lors de l'expérience, les sujets exploraient librement le stimulus c'est-à-dire la courbure d'un arc en trois dimensions. Généralement, la courbure perçue par les sujets peut être décrite comme une moyenne des deux signaux. Cependant, dans le cas d'arcs très convexes, le poids du signal *position* est plus important que le poids du signal *force* alors que dans le cas d'arcs peu profond, c'est le poids du signal *force* qui est le plus important.

2 - Intra-modal : visuel

La perception d'objet est une tâche importante que nous réalisons souvent et qui requiert que le système visuel obtienne des informations géométriques sur la forme des objets. Généralement, nous percevons très rapidement et avec une bonne fiabilité les formes des objets, malgré leurs complexités provenant par exemple de l'effet de projection, du fouillis de l'arrière plan, ou encore de l'éclairage. Le formalisme bayésien est un cadre intéressant notamment en apportant des connaissances préalables (priors) pour résoudre les ambiguïtés lors de la perception d'objets [9], [10]. Dans notre formulation, il s'agit donc de rendre $P(\varphi)$ différent d'une loi de probabilité uniforme pour apporter des connaissances à priori sur l'environnement et ses propriétés. Par exemple, nous pouvons apporter comme connaissances aux modèles que les visages sont plus probablement convexes que concaves, que les sources lumineuses se situent plus probablement au-dessus de la scène qu'en dessous, que ce sont plus probablement les objets de la scène qui se déplacent plutôt que la source lumineuse, etc.

3 - Multi-modal : visuo-haptique

Nous nous intéressons ici aux travaux de modélisation visuo-haptique initiés par Ernst et Banks [6]. La fusion multi-sensorielle lors de cas non conflictuel permet d'estimer avec une plus grande précision les propriétés des objets de l'environnement (forme, localisation, profondeur...). C'est aussi par exemple le cas, lorsque nous cherchons à estimer la taille d'un objet. Deux modalités peuvent être mises à contribution : la vision et le toucher. Il s'avère que dans certains cas, une modalité domine les autres. C'est fréquemment le cas de la vision lors de l'étude d'une propriété d'un objet. Cette intégration est alors appelée capture visuelle. Cependant, lorsque le niveau de bruit sur la vision est assez important, par exemple lorsque la barre est mal orientée pour le sujet, le toucher joue alors un rôle important [8] : nous observons dans ce cas un mélange des informations sensorielles se comportant comme une fusion capteur.

Nous détaillons maintenant l'expérience d'Ernst et Banks. Le but cette expérience était d'examiner quantitativement l'intégration visuo-haptique. Pour cela, les sujets devaient regarder et/ou toucher une barre horizontale afin de juger sa hauteur (voir Figure 2.3). Le stimulus visuel correspondait à un stéréogramme de points aléatoires simulant la barre. Afin de varier la fiabilité de ce stimulus, l'expérience a été réalisée avec quatre niveaux de bruits (déplacement aléatoire des points du stéréogramme). Deux appareils à retour d'effort (PHANToM) permettaient de générer le stimulus haptique. L'expérience se composait de la manière suivante : des tests sur la perception visuelle seule, sur le toucher seul et sur les deux modalités simultanément. Ainsi avec cet appareillage, il est possible de manipuler indépendamment les indices visuels et haptiques, et surtout de manipuler la qualité de ces indices, et donc leur fiabilité.

Le modèle utilisé pour décrire cette expérience est le modèle de maximum de vraisemblance. Soit $\hat{S}$, l'estimation d'une propriété de l'environnement par un système sensoriel, $\hat{S}_i$ l'estimation pour chaque modalité, φ la propriété physique de l'environnement et f une opération par laquelle le système nerveux réalise l'estimation. L'estimation de l'environnement peut être représentée par :

$$\hat{S}_i \quad = \quad f_i(\varphi).$$

La fluctuation de l'estimation $\hat{S}_i$ est représentée par la variance σ_i^2, dépendant du niveau de bruit.

$$\hat{S} \quad = \quad \sum w_i \hat{S}_i \text{ avec } w_i = \frac{\frac{1}{\sigma_i^2}}{\sum \frac{1}{\sigma_j^2}}$$

Ainsi, nous obtenons la moyenne de l'estimation selon la règle du maximum de vraisemblance en ajoutant l'estimation de chaque modalité pondérée par leur variance. Si cette méthode est appliquée à la fusion visuo-haptique, la variance de l'estimation finale devient :

$$\sigma_{VH}^2 = \quad \frac{\sigma_V^2 \sigma_H^2}{\sigma_V^2 + \sigma_H^2}$$

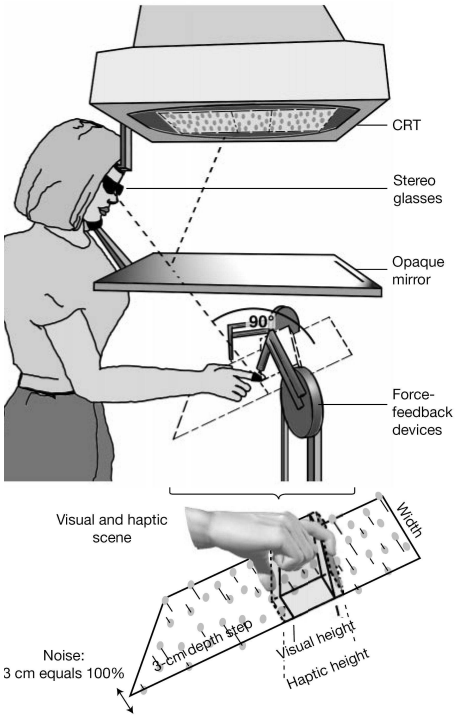

FIG. 2.3 – Appareillage et stimuli de l'expérience d'Ernst et Banks (tirée de [6]).

avec σ_V la variance pour l'estimation de la vision et σ_H pour le toucher. Ainsi, l'estimation finale a une plus petite variance que les estimations de la vision et du toucher comme le montre la Figure 2.4. La formulation présentée ici correspond au modèle robotique de fusion capteur dans le cas où $P(S_1 \mid \varphi)$ et $P(S_2 \mid \varphi)$ sont définis comme des lois gaussiennes.

En utilisant les données de l'expérience sur une seule modalité (vision seule ou haptique seul) et le modèle présenté ci-dessus, il est possible de prédire la réponse du sujet lorsque les informations visuelles et haptiques lui sont présentées simultanément et donc de comparer ces prédictions aux réponses des sujets. Ainsi, les résultats de cette étude montrent que le système nerveux semble combiner les informations visuelles et haptiques de manière similaire à l'intégration proposée par le modèle de maximum de vraisemblance.

Cependant, ce modèle ne peut pas mathématiquement prendre en compte les aspects temporels. Or, il n'y a aucune garantie que les sujets ne s'adaptent pas durant le temps de l'expérience, ce qui pourrait les conduire à donner des réponses différentes selon le moment de l'expérience.

4 - Multi-modal : Visuo-Acoustique pour la localisation spatiale

Prenons un second exemple de fusion sensorielle, cette fois entre la vision et l'audition pour la localisation spatiale [2], [3]. C'est le cas lorsqu'une personne vous appelle dans une foule et que vous cherchez à la localiser. Les deux signaux visuels et auditifs peuvent

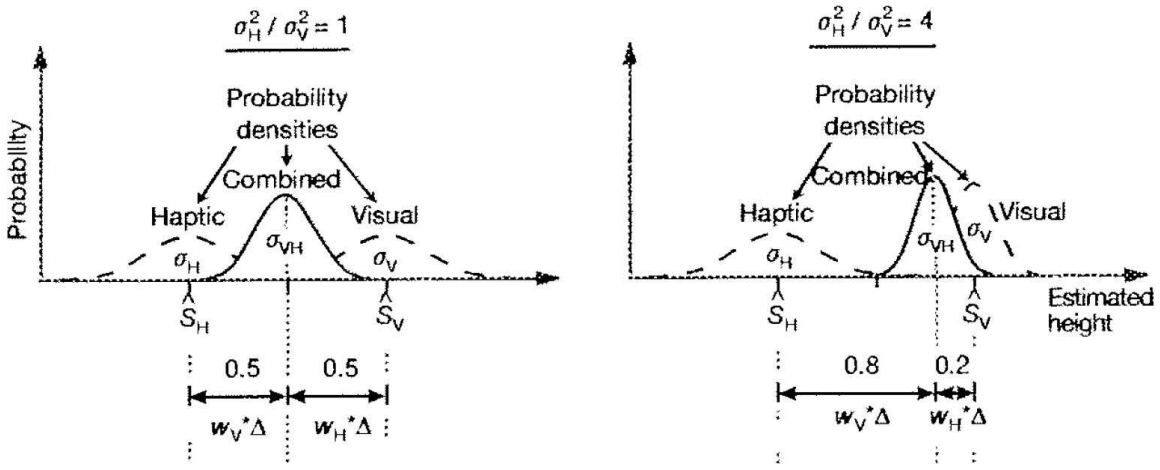

FIG. 2.4 – Maximum de vraisemblance : deux situations. A gauche, les deux informations sensorielles sont aussi fiables l'une que l'autre et l'estimée final est donc la moyenne des estimées fournies par les canaux indépendamment. A droite, l'information visuelle étant plus fiable que l'information haptique, l'estimée finale se rapproche de l'estimée visuelle (Figure tirée de [6]).

fournir des informations sur la localisation et donc indiquer la provenance du stimulus. Bien sûr, cette direction est indiquée avec une plus ou moins bonne précision : généralement, le système visuel est meilleur que le système auditif pour estimer une direction. C'est pourquoi, nous avons une première théorie, la capture visuelle, pour laquelle l'information visuelle pour la localisation spatiale domine complètement la fusion avec l'information auditive. Cependant, certaines études ont aussi montré que l'information auditive est aussi prise en compte selon le modèle de maximum de vraisemblance. En accord avec ce modèle, des expériences ont montré que lorsque les stimuli visuels et auditifs étaient tous les deux présents, les sujets percevaient mieux la direction du signal que lorsqu'un seul stimulus était présent. Cependant, il reste certaines questions non résolues : comment le cerveau connaît la variance de ses estimations pour chaque modalité ? Comment le cerveau sait-il que les estimations proviennent de la même source ou de source différentes ?

Il y a conflit multi-sensoriels lorsque les différentes modalités liées à la perception fournissent des informations contradictoires. Il existe de nombreuses illusions basées sur des conflits multi-sensoriels. Nous en faisons l'expérience très régulièrement en regardant un film au cinéma par exemple : nous avons l'impression que les voix proviennent des lèvres des acteurs bien qu'elles proviennent des haut parleurs placés sur le coté de l'écran. Un ventriloque est une personne qui peut parler sans remuer les lèvres. Il y conflit visuo-acoustique lorsque la localisation spatiale entre les stimulus auditifs et visuels est différente [1]. Comme la localisation visuelle est généralement supérieur à la localisation auditive, c'est cette modalité qui domine et c'est pour cela que nous croyons que la voix du ventriloque sort de la bouche de sa marionnette.

	/a/	/e/	/i/	/ø/	/y/	/o/	/u/
/a/	210						1
/e/		99	1	74	4	53	10
/i/		3	109	1	86	1	41
/ø/		83	3	105	6	74	15
/y/		6	75		91		36
/o/		17		26		80	12
/u/		2	20	4	23	2	95

TAB. 2.1 – Matrice de confusions auditive avec RSB = -6 dB (données tirées de [13]).

5 - Multi-modal : Visuo-Acoustique pour la reconnaissance de phonèmes

Cet exemple de fusion multi-sensorielle se base sur les mêmes modalités que précédemment, la vue et l'ouïe, mais cette fois pour la reconnaissance de phonèmes. Dans ce paragraphe, nous présentons tout d'abord l'expérience réalisée puis les modèles proposés par Robert-Ribes.

Cette partie est extraite de la présentation du test de perception audiovisuelle réalisé par Robert-Ribes [13]. Lors de l'expérience, les sujets devaient reconnaître des voyelles qu'ils pouvaient soit voir, soit entendre, ou les deux. Le test de perception audiovisuelle est limité aux sept voyelles suivantes : /a/, /e/, /i/, /ø/, /y/, /o/, /u/. Dix stimuli ont été réalisés pour chaque voyelle. Chaque stimulus est composé de 200 ms de signal sonore et de 5 images vidéo synchronisées de manière précise. Les sons ont été bruités avec six niveaux différents de bruit : 12 dB, 6 dB, 0 dB, -6 dB, -12 dB, -18 dB de rapport signal à bruit (noté RSB par la suite). L'expérience comportait trois étapes, passées dans cet ordre : audiovisuelle, visuelle seule et auditive seule. La tâche des sujets consistait à identifier la voyelle que le locuteur avait prononcée.

Le Tableau 2.1 présente la matrice de confusions en auditif pur pour un rapport signal sur bruit de -6 dB. Les colonnes présentent les réponses des sujets aux différents stimuli. Par exemple, nous observons que 209 sujets ont répondu /a/ et 1 sujet a répondu /u/ au stimulus /a/ (informations de la première colonne).

Le Tableau 2.2 présente la matrice de confusions en visuel pur. Nous observons que la voyelle /a/ est très bien reconnue en visuel pur et que les voyelles /e/ et /i/ sont reconnues dans environ 70 % des cas. Les quatre dernières voyelles, /ø/, /y/, /o/, /u/ ne sont pas différenciée en visuel pur.

Enfin, le Tableau 2.3 présente la matrice de confusions en audiovisuel pour un rapport signal sur bruit de -6 dB. Nous observons qu'il y a moins de confusions dans cette matrice que dans chacune des matrices auditives pures et visuelle pure prises séparément. Les matrices de confusions pour les différents niveaux de bruit et pour les stimuli auditifs, visuels et audiovisuels sont présentées Annexe B.

La Figure 2.5 présente les scores de reconnaissance correcte corrigés par rapport au seuil aléatoire selon le niveau de bruit. En annexe C, se trouvent deux méthodes pour analyser

	/a/	/e/	/i/	/ø/	/y/	/o/	/u/
/a/	209						
/e/		150	140				
/i/		58	68	2	4	4	2
/ø/	1	2	2	107	82	76	95
/y/				64	100	23	73
/o/				12	3	91	10
/u/				25	21	16	30

TAB. 2.2 – Matrice de confusions en visuel pur (données tirées de [13]).

	/a/	/e/	/i/	/ø/	/y/	/o/	/u/
/a/	209						
/e/		203	12	3	1	1	
/i/		6	196		4		
/ø/		1		169	12	89	26
/y/			2	4	165		53
/o/	1			34	1	118	12
/u/					27	2	119

TAB. 2.3 – Matrice de confusions en audiovisuel pour RSB = -6 dB (données tirées de [13]).

les résultats que nous exploiterons par la suite : le score d'identification correcte corrigé et le pourcentage d'information transmise par trait.

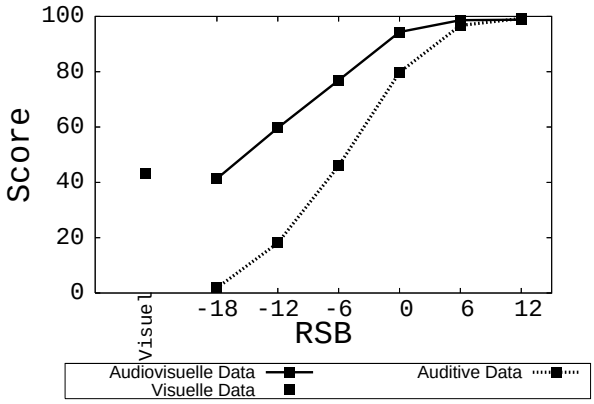

FIG. 2.5 – Scores d'identification correcte corrigés.

Globalement, les données permettent d'observer que la perception des voyelles en audiovisuel est parfaite sans l'ajout de bruit et qu'elle se dégrade progressivement avec les différents niveaux de bruit. Cependant la dégradation est limitée par la perception visuelle : la perception audiovisuelle est soumise à un effet de plancher par la perception visuelle.

Robert-Ribes propose deux modèles : le modèle RD et le modèle RM. Ces deux modèles partagent la même structure en trois étapes :

- 1 : *association* entre un élément de l'espace d'entrée et un élément de l'espace d'intégration ou projection de deux entrées dans un espace de fusion.
- 2 : *intégration* des deux projections dans cet espace de fusion.
- 3 : *classification* de la représentation intégrée.

La différence entre les deux modèles est la nature de l'espace de fusion : motrice pour le modèle RM et acoustique pour le modèle RD. Dans le modèle RM présenté Figure 2.6, l'espace de fusion est supposé tri-dimensionnel basé sur les trois paramètres articulatoires classiques de la phonétiques, c'est-à-dire X et Y les coordonnées horizontale et verticale du point le plus haut de la langue et S l'air intérolabiale.

Dans le modèle RD présenté Figure 2.7, l'espace de fusion est supposé à 20 dimensions. Il correspond à l'espace interne auditif (supposé dominant). L'espace auditif est représenté de la même façon que les signaux acoustiques dans l'expérience, c'est-à-dire avec 20 canaux de 1 Bark dans l'échelle (Bark, dB).

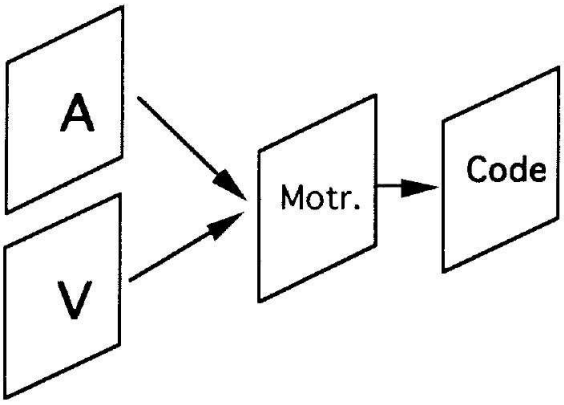

FIG. 2.6 – Schéma du modèle RM (Figure tirée de [13]).

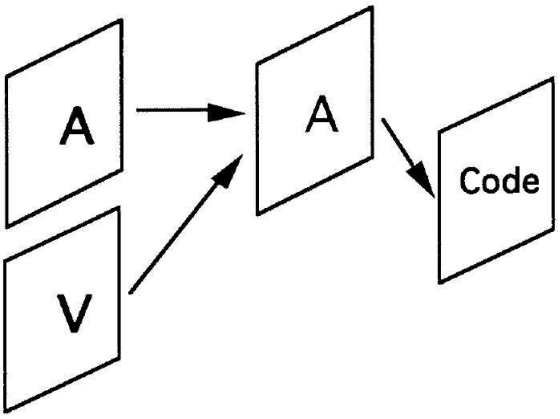

FIG. 2.7 – Schéma du modèle RD (Figure tirée de [13]).

2.4 Modèles bayésiens et sélection de modèles pour la perception audiovisuelle des voyelles

Nous l'avons montré, de nombreux domaines d'étude de la perception humaine sont propices à l'application de modèles bayésiens de fusion capteur. La diversité de ces études semblerait conforter l'hypothèse générale selon laquelle ce modèle mathématique décrit convenablement le principe de fusion d'informations sensorielles chez l'humain. Nous proposons dans notre travail de complémenter les études ci-dessus en proposant une modélisation bayésienne de la perception audiovisuelle des voyelles. Nous utiliserons les données expérimentales issues de la thèse de Robert-Ribes [13] qui décrivent des cas de congruence audiovisuelle, mais aussi sur des données d'expériences dans des cas de conflits audiovisuels, [12] (l'expérience sera décrite Section 4.2). Nous proposons une modélisation bayésienne de ces données expérimentales. Ce travail de modélisation est original dans ce domaine.

De plus, nous proposons d'apporter dans ce travail une autre originalité. En effet, la

modélisation bayésienne est un cadre formel permettant de traiter naturellement des problèmes inverses et mal-posés. C'est également un formalisme approprié à la sélection de modèles. En d'autres termes, nous allons proposer plusieurs modèles bayésiens de perception audiovisuelle des voyelles. Ces modèles feront des hypothèses différentes sur l'espace interne dans lequel s'opère la fusion (de manière similaire à RM et RD). Ces modèles seront présentés Chapitre 3. Ces modèles seront ensuite confrontés aux données expérimentales, dans des cas de congruence (Section 4.1) et dans des cas de fusion (Section 4.2). La comparaison entre les modèles sera effectuée de manière qualitative comme la comparaison des modèles RD et RM ([13]) mais aussi de manière quantitative grâce à la modélisation bayésienne.

Chapitre 3

Les modèles

Ce chapitre propose trois modèles se basant sur des architectures différentes, pour étudier le problème de fusion audiovisuelle en perception de la parole. Ces modèles sont construits en suivant le formalisme de programmation bayésienne. Ils se composent donc de deux phases, une phase description et une phase question. La programmation de ces modèles s'est faite en C++ avec l'aide de la librairie ProBT ©, permettant de faire de l'inférence et du calcul bayésien.

3.1 Modèle M_0

Pour le modèle M_0, nous faisons l'hypothèse que l'espace interne pour la fusion est commun à l'auditif et au visuel, c'est-à-dire que les indices auditifs et visuels sont projetés dans un même espace avant de réaliser la fusion. Nous faisons de plus des hypothèses sur la nature de cet espace interne : nous le supposons mono-dimensionnel et avec la relation d'ordre suivante : /a/, /e/, /i/, /ø/, /o/, /y/, /u/ (/a/ est plus proche de /e/ qu'elle ne l'est de toutes les autres voyelles, /e/ est plus proche de /a/ et de /i/ qu'elle ne l'est de /ø/, /o/, /y/, /u/, ...). Ce modèle présenté dans son intégralité Figure 3.1 est détaillé et commenté dans le reste de cette section.

1 - Les variables pertinentes

Pour la réalisation de ce modèle nous avons utilisé quatre variables.

- A : Variable correspondant au stimulus auditif. Ce stimulus correspond à un signal sonore. A est une variable à 7 états : {/a/, /e/, /i/, /ø/, /o/, /y/, /u/}.
- V : Variable correspondant au stimulus visuel. Il correspond à une courte vidéo composée de 5 images. Cette variable est sur le même domaine que A.
- P : Variable de perception qui correspond à ce que le sujet a perçu, c'est-à-dire à la voyelle qu'il pense avoir vu et/ou entendu. Son domaine est le même que celui de A.
- B : Variable, représentant le niveau de bruit, à 6 états : {-18, -12, -6, 0, 6, 12}.

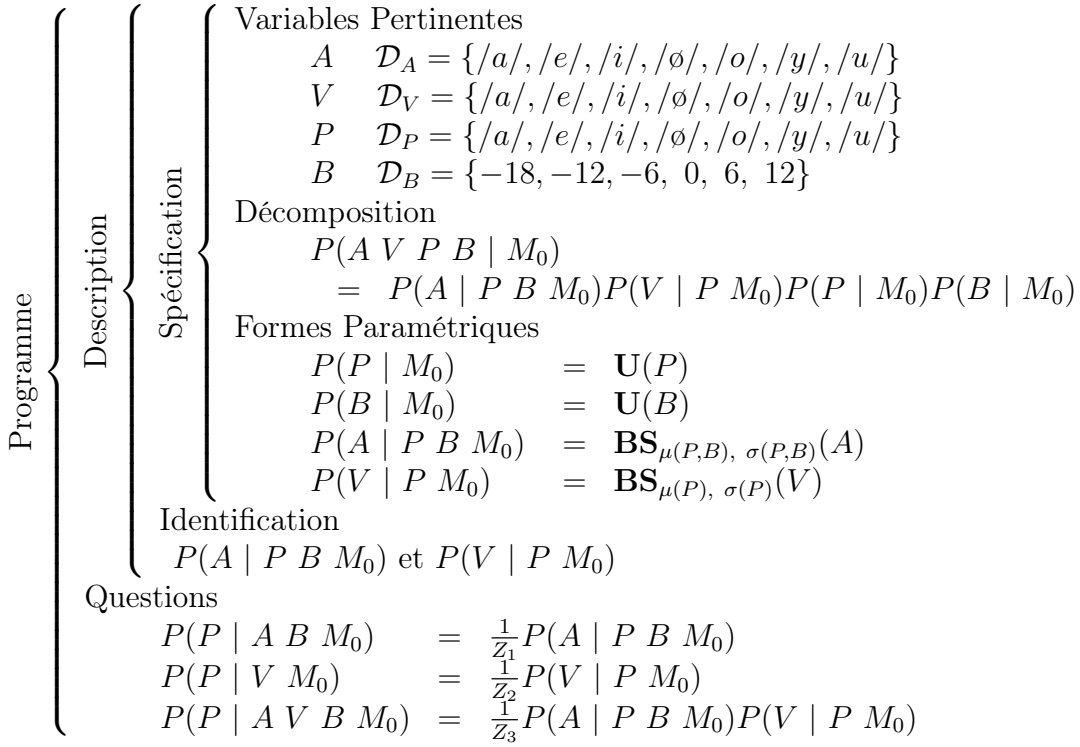

FIG. 3.1 – Résumé du modèle M_0.

2 - Décomposition et formes paramétriques

Avec la définition des variables données ci-dessus, la décomposition de la distribution de probabilité conjointe prend la forme suivante :

$$
\begin{aligned}
P(A\ V\ P\ B \mid M_0) &= P(A\ V \mid P\ B\ M_0)P(P \mid B\ M_0)P(B \mid M_0) && (3.1)\\
&= P(A \mid P\ B\ M_0)P(V \mid P\ B\ M_0)P(P \mid B\ M_0)P(B \mid M_0) && (3.2)\\
&= P(A \mid P\ B\ M_0)P(V \mid P\ M_0)P(P \mid M_0)P(B \mid M_0). && (3.3)
\end{aligned}
$$

Les égalités (3.2) et (3.3) sont obtenues par des hypothèses d'indépendance conditionnelle. Pour l'égalité (3.2), nous faisons l'hypothèse que la valeur du stimulus auditif (variable A) est indépendante de la valeur du stimulus visuel (variable V), supposant que l'on connaît la voyelle perçue (variable P). Pour l'égalité (3.3), nous faisons l'hypothèse que seul le stimulus auditif dépend du bruit.

Les formes paramétriques associées aux différents termes sont les suivantes :

$$\begin{aligned} P(P \mid M_0) &= \mathbf{U}(P) \\ P(B \mid M_0) &= \mathbf{U}(B) \\ P(A \mid P\ B\ M_0) &= \mathbf{BS}_{\mu(P,B),\ \sigma(P,B)}(A) \\ P(V \mid P\ M_0) &= \mathbf{BS}_{\mu(P),\ \sigma(P)}(V). \end{aligned}$$

Les termes $P(P \mid M_0)$ et $P(B \mid M_0)$ correspondent à des distributions de probabilité de loi uniforme car nous n'avons aucun a priori sur la valeur de la voyelle perçue et sur le niveau de bruit du stimulus auditif. Les termes $P(A \mid P\ B\ M_0)$ et $P(V \mid P\ M_0)$ sont associés à des courbes en cloches approximant, sur un espace discret, des lois gaussiennes à une dimension (`BellShape` noté **BS**). En effet, nous supposons que l'espace des voyelles est discret, ordonné et mono-dimensionnel. L'ordre a été défini à l'aide de l'arbre de confusions audiovisuelles présenté Figure 3.2.

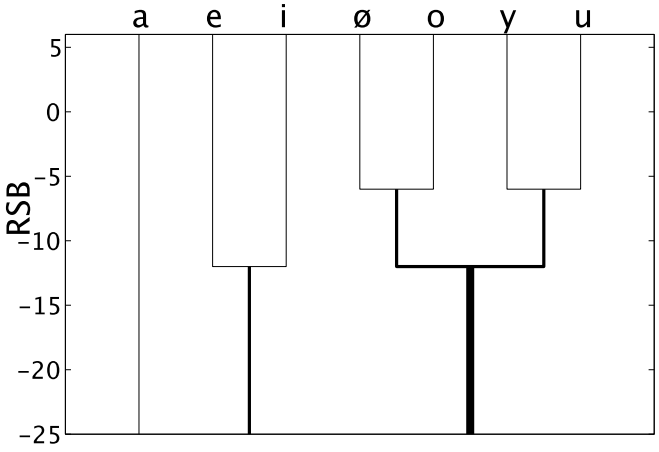

FIG. 3.2 – Arbre de confusions audiovisuelles tiré de [13]. Chaque voyelle est représentée par une branche et l'ordonnée représente le niveau de bruit. Deux branches sont regroupées pour que les stimuli soient identifiés à l'intérieur de chaque branche dans au moins 75 % des cas. Nous en tirons l'agencement suivant entre les voyelles : {/a/, /e/, /i/, /ø/, /o/, /y/, /u/}.

3 - Identification

Après avoir spécifié les connaissances préalables, nous avons fixé les valeurs des formes paramétriques associées aux termes de la décomposition et, en l'occurrence, les moyennes et écart-types des `BellShape` $P(A \mid P\ B\ M_0)$ et $P(V \mid P\ M_0)$. Lors de l'expérience, les sujets devaient déterminer les voyelles vues et/ou entendues. Cela a permis de déterminer les matrices de confusions et les probabilités $P(P \mid A\ B\ M_0)$ et $P(P \mid V\ M_0)$. Ces termes correspondent à quelle est la voyelle perçue pour un stimulus auditif ou visuel donné. Or,

nous cherchons à déterminer les termes $P(A \mid P\ B\ M_0)$ et $P(V \mid P\ M_0)$. Ces termes sont les inverses des termes obtenus lors de l'expérience. Ils correspondent à la probabilité de retrouver le stimulus auditif ou visuel connaissant la variable perçue.

L'égalité (3.7) nous permet de justifier l'équivalence entre les termes $P(P \mid A\ B\ M_0)$ et $P(A \mid P\ B\ M_0)$. De la même manière, nous pouvons montrer l'équivalence entre les termes $P(P \mid V\ M_0)$ et $P(V \mid P\ M_0)$.

$$
\begin{aligned}
P(P \mid A\ B\ M_0) &= \frac{P(P\ A\ B \mid M_0)}{P(A\ B \mid M_0)} && (3.4)\\
&= \frac{\sum_V P(P\ A\ B\ V \mid M_0)}{P(A \mid B\ M_0)P(B \mid M_0)} && (3.5)\\
&= \frac{1}{Z_1}\sum_V P(P \mid M_0)P(A \mid P\ B\ M_0)P(V \mid P\ M_0)P(B \mid M_0) && (3.6)\\
&= \frac{1}{Z_2}P(A \mid P\ B\ M_0) && (3.7)
\end{aligned}
$$

Les données expérimentales tirées de [13] nous donnent donc les tables de probabilité des termes $P(P \mid A\ B\ M_0)$ et $P(P \mid V\ M_0)$, ce qui nous permet par la suite de calculer les moyennes et écart-types des `BellShape`.

Une fois l'identification des termes réalisée, nous obtenons des courbes en cloche approximant des lois gaussiennes telles que celle présentée Figure 3.3. Cette courbe montre que lorsque l'on perçoit la voyelle /ø/ et que le bruit du stimulus auditif est de 0 dB, il y a une forte probabilité que ce stimulus corresponde à la voyelle /ø/ ou encore aux voyelles /o/ et /i/.

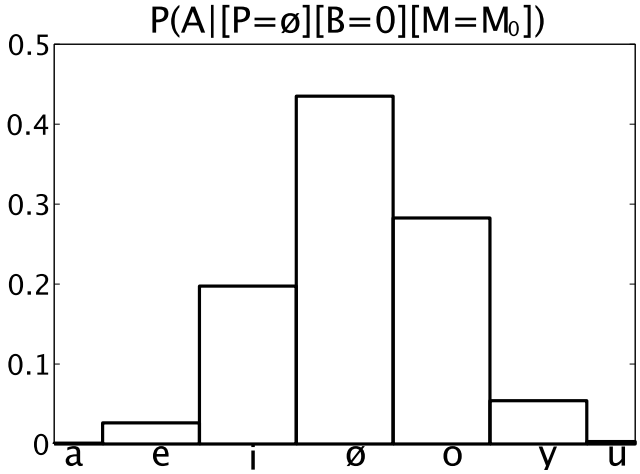

FIG. 3.3 – Courbe en cloche approximant une loi gaussienne.

4 - Questions

Une fois les phases de spécification et d'identification terminées, nous disposons d'une description complètement définie. Durant la phase d'exploitation, nous allons utiliser cette description afin de répondre à des questions. Poser une question consiste à chercher la distribution de probabilité d'un certain nombre de variables de la description connaissant les valeurs d'autres variables. Dans notre cas, nous allons poser trois questions : nous cherchons à connaître la distribution de probabilité de la voyelle perçue par le sujet connaissant tout d'abord le stimulus auditif et son niveau de bruit (équation (3.8)) puis, le stimulus visuel (équation (3.9)) et enfin connaissant le stimulus auditif avec son niveau de bruit et le stimulus visuel (équation (3.10)).

Question dans le cas auditif pur :

$$P(P \mid A\ B\ M_0). \tag{3.8}$$

Question dans le cas visuel pur :

$$P(P \mid V\ M_0). \tag{3.9}$$

Question dans le cas audiovisuel :

$$P(P \mid A\ V\ B\ M_0). \tag{3.10}$$

Dans ce chapitre, nous nous plaçons uniquement dans des cas de non conflits audiovisuels et par conséquent, les stimuli auditifs et visuels doivent correspondre à la même voyelle : dans l'équation (3.10), les variables A et V doivent avoir la même valeur.

L'inférence bayésienne permet de répondre à ces questions à partir de la décomposition choisie précédemment. Nous ne détaillerons dans cette partie que l'inférence dans le cas de la fusion audiovisuelle (équation (3.10)). La même méthode est utilisée pour répondre aux questions dans les cas de l'auditif pur et du visuel pur.

$$\begin{aligned}
P(P \mid A\ V\ B \mid M_0) &= \frac{P(P\ A\ V\ B \mid M_0)}{P(A\ V\ B \mid M_0)} && (3.11)\\
&= \frac{1}{Z_1} P(P\ A\ V\ B \mid M_0) && (3.12)\\
&= \frac{1}{Z_1} P(A \mid P\ B\ M_0) P(V \mid P\ M_0) P(P \mid M_0) P(B \mid M_0) && (3.13)\\
&= \frac{1}{Z_2} P(A \mid P\ B\ M_0) P(V \mid P\ M_0). && (3.14)
\end{aligned}$$

Grâce à ces questions, nous pouvons simuler le modèle dans les mêmes conditions que l'expérience de la thèse de Robert-Ribes [13]. Nous pouvons par exemple poser la question suivante, $P(P \mid [A = /\text{ø}/]\ [V = /\text{ø}/]\ [B = 0]\ M_0)$ qui correspond à : qu'est ce que le

modèle a « perçu » si le stimulus auditif est /ø/, le stimulus visuel est /ø/ et le niveau de bruit est de 0 dB. Nous obtenons la distribution de probabilité présentée Tableau 3.1. Cette simulation nous permet d'évaluer les résultats des modèles dans les mêmes conditions que l'expérience et donc de déterminer les tableaux de scores des modèles.

	/a/	/e/	/i/	/ø/	/o/	/ y/	/u/
$P(P\|[A = /ø/][V = /ø/][B = 0]M_0)$	0	0	0.31	0.49	0.03	0.17	0

TAB. 3.1 – Distribution de probabilité relative à la question $P(P \mid [A = /ø/]\ [V = /ø/]\ [B = 0]\ M_0)$.

Pour chaque niveau de bruit et pour chaque voyelle, nous effectuons 210 tirages (nombre de stimuli présentés à chaque sujet durant le test audiovisuel) suivant la distribution de probabilité obtenue à la question. Cela nous permet par la suite de calculer les scores d'identification correcte corrigés auditifs, visuels et audiovisuels pour chaque niveau de bruit. Afin de pouvoir analyser ces scores, nous avons rajouté sur le même graphe (Figure 3.4) les scores d'identification correcte corrigés des réponses des sujets (courbes *Data*).

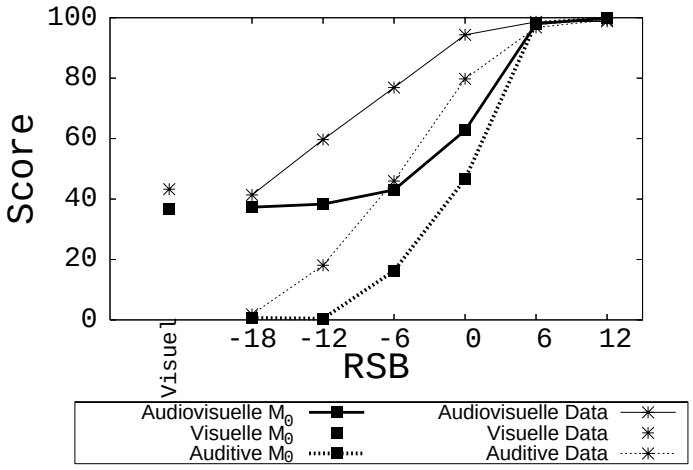

FIG. 3.4 – Scores d'identification correcte corrigés des sujets humains (*Data*) et du modèle M_0.

Nous observons que les scores en auditif pur (en pointillé gras) pour le modèle M_0 sont relativement similaires aux scores en auditif pur (pointillé fin) directement calculés sur les données, pour un niveau de bruit faible (6 et 12 dB) et un niveau élevé (-18 dB). Cependant, nous remarquons que pour des niveaux de bruits moyens (-12 dB, -6 dB, 0 dB), les scores obtenus avec ce modèle sont inférieurs aux scores réalisés par les sujets. Le même constat

est réalisé pour les scores audiovisuels. Nous notons qu'avec notre modèle, la perception audiovisuelle est toujours soumise à l'effet de plancher de la perception visuelle.

En annexe D.1 se trouvent les pourcentages d'information transmise par trait (arrondissement, hauteur, avant-arrière). Nous observons que quelque soit le trait étudié, les probabilités d'identification sont plus faibles pour notre modèle M_0 que pour les sujets. Ceci s'explique par le fait que l'ordre des voyelles ne respecte pas les différents traits des voyelles : par exemple, les voyelles /o/ et /y/ sont voisines alors qu'elles ont un trait de hauteur différent.

3.2 Modèle M_1

Dans cette partie, nous allons commenter la description correspondant au modèle M_1 présentée Figure 3.5. Pour ce modèle, certaines des hypothèses sont les mêmes que précédemment : la fusion est réalisée dans un espace interne mono-dimensionel commun à l'auditif et au visuel et, en revanche, contrairement à M_0, dans M_1, nous supposons qu'il n'y a aucune relation d'ordre entre les voyelles.

Programme

Description

Spécification

Variables Pertinentes

$$\begin{array}{ll} A & \mathcal{D}_A = \{/a/, /e/, /i/, /\emptyset/, /o/, /y/, /u/\} \\ V & \mathcal{D}_V = \{/a/, /e/, /i/, /\emptyset/, /o/, /y/, /u/\} \\ P & \mathcal{D}_P = \{/a/, /e/, /i/, /\emptyset/, /o/, /y/, /u/\} \\ B & \mathcal{D}_B = \{-18, -12, -6,\ 0,\ 6,\ 12\} \end{array}$$

Décomposition

$$\begin{aligned} & P(A\ V\ P\ B \mid M_1) \\ & \quad = \ P(A \mid P\ B\ M_1) P(V \mid P\ M_1) P(P \mid M_1) P(B \mid M_1) \end{aligned}$$

Formes Paramétriques

$$\begin{array}{lcl} P(P \mid M_1) & = & \mathbf{U}(P) \\ P(B \mid M_1) & = & \mathbf{U}(B) \\ P(A \mid P\ B\ M_1) & = & \mathbf{L}_{\mu(P,B),\ \sigma(P,B)}(A) \\ P(V \mid P\ M_1) & = & \mathbf{L}_{\mu(P),\ \sigma(P)}(V) \end{array}$$

Identification

$P(A \mid P\ B\ M_1)$ et $P(V \mid P\ M_1)$

Questions

$$\begin{array}{lcl} P(P \mid A\ B\ M_1) & = & \frac{1}{Z_1} P(A \mid P\ B\ M_1) \\ P(P \mid V\ M_1) & = & \frac{1}{Z_2} P(V \mid P\ M_1) \\ P(P \mid A\ V\ B\ M_1) & = & \frac{1}{Z_3} P(A \mid P\ B\ M_1) P(V \mid P\ M_1) \end{array}$$

FIG. 3.5 – Résumé du modèle M_1.

1 - Les variables pertinentes

La définition des variables du modèle M_1 est la même que pour le modèle M_0.

2 - Décomposition et formes paramétriques

La décomposition de la distribution de probabilité conjointe a la même forme que précédemment. Les formes paramétriques associées aux différents termes sont les suivantes :

$$\begin{aligned} P(P \mid M_1) &= \mathbf{U}(P) \\ P(B \mid M_1) &= \mathbf{U}(B) \\ P(A \mid P\ B\ M_1) &= \mathbf{L}_{n_1,n_2,\ldots,n_{kA}}(A) \\ P(V \mid P\ M_1) &= \mathbf{L}_{n_1,n_2,\ldots,n_{kV}}(V). \end{aligned}$$

Comme précédemment, les termes $P(P \mid M_1)$ et $P(B \mid M_1)$ son associés à des distributions de probabilité de loi uniforme.
Dans M_0, nous supposons des espaces ordonnés sur A et V, ce qui permettait de définir $P(A \mid P\ B\ M_1)$ et $P(V \mid P\ M_1)$ comme des approximations de gaussiennes. Dans M_1, ce n'est plus le cas, et nous definissons $P(A \mid P\ B\ M_1)$ et $P(V \mid P\ M_1)$ comme des lois de successions de Laplace (qui sont des « variantes » d'histogrammes). En effet, nous supposons que l'espace des voyelles est discret et qu'il n'y a aucune relation entre les voyelles.

3 - Identification

Comme précédemment, nous utilisons l'équivalence entre $P(A \mid P\ B\ M_1)$ (resp. $P(V \mid P\ M_1)$) et $P(P \mid A\ B\ M_1)$ (resp. $P(P \mid V\ M_1)$) et les matrices de confusions pour déterminer les paramètres des lois de succession de Laplace[1].

Après cette phase d'identification, nous obtenons les lois de succession de Laplace telles que celle présentée Figure 3.6. Ces « histogrammes » montrent que lorsque l'on perçoit la voyelle /ø/ et que le bruit du stimulus auditif est de 0 dB, il y a une très forte probabilité que ce stimulus corresponde à la voyelle /ø/.

4 - Questions

Les questions posées pour ce modèle sont les mêmes que celles du modèle M_0 (équations (3.8), (3.9) et (3.10)). La Figure 3.7 présente le score d'identification correcte corrigé comparé à celui des données.

En annexe D.2 se trouvent les pourcentages d'information transmise par trait (arrondissement, hauteur, avant-arrière).

[1]Par exemple, lors de l'identification du terme $P(V \mid P\ M_1)$ la formule de la loi de succession de Laplace prend la forme suivante : $P(V \mid P\ M_1) = \frac{n_i+1}{N+k}$ avec n_i le nombre de fois où le cas i a été observé, N le nombre total d'observations et k la taille du domaine de la variable P. Cette forme est intéressante parce qu'elle converge vers l'histogramme de fréquences lorsque N est grand, mais sans jamais avoir de probabilité égale à la valeur 0.

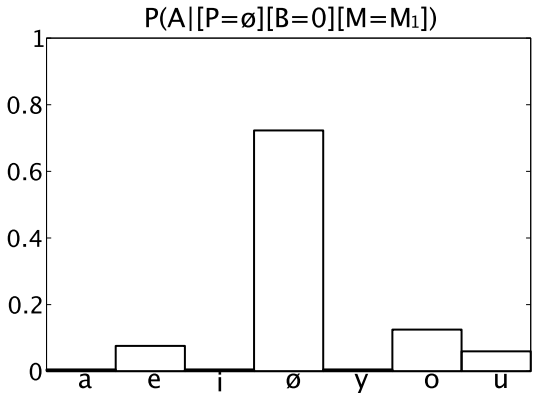

FIG. 3.6: Lois de succession de Laplace identifiées pour $P(A \mid [P = \emptyset][B = 0][M = M_1])$.

Que ce soit pour le score global ou par trait, nous observons que les résultats obtenus pour le modèle M_1 sont quasiment identiques aux résultats obtenus par les sujets lors de l'expérience.

3.3 Modèle M_2

Pour ce modèle, nous avons choisi d'utiliser une structure géométrique susceptible d'expliquer les confusions contenues dans les matrices de confusions. Nous faisons l'hypothèse que l'espace interne est tri-dimensionnel et qu'il traduit des connaissances sur le système articulatoire humain. Chacun des degrés de liberté de cette structure est lié à une dimension articulatoire : t à la dimension avant-arrière des lèvres, u à l'arrondissement des lèvres et v à la hauteur de la mâchoire. La structure géométrique est dégradée par ajout de bruit. La Figure 3.8 présente la structure géométrique de l'espace auditif pour un niveau de bruit de 12 et 6 dB. Les figures représentant cette structure pour des niveaux de bruit égaux à 0, -6, -12 et -18 dB sont présentées Annexe E. Pour l'espace visuel, c'est la structure géométrique de l'espace auditif à 12 dB qui est utilisée.

Ce modèle présenté dans son intégralité Figure 3.9 est détaillé dans le reste de cette section.

1 - Les variables pertinentes

Pour la réalisation de ce modèle nous avons utilisé dix variables :

- A_u : variable correspondant au stimulus auditif selon le degré de liberté lié à la dimension arrondissement. Cette variable peut prendre 6 valeurs entières de 0 à 5.
- A_v : variable correspondant au stimulus auditif selon le degré de liberté lié à la dimension hauteur. Cette variable peut prendre 11 valeurs entières de 0 à 10.
- A_t : variable correspondant au stimuli auditif selon le degré de liberté lié à la dimension avant-arrière. Cette variable peut prendre 6 valeurs entières de 0 à 5.

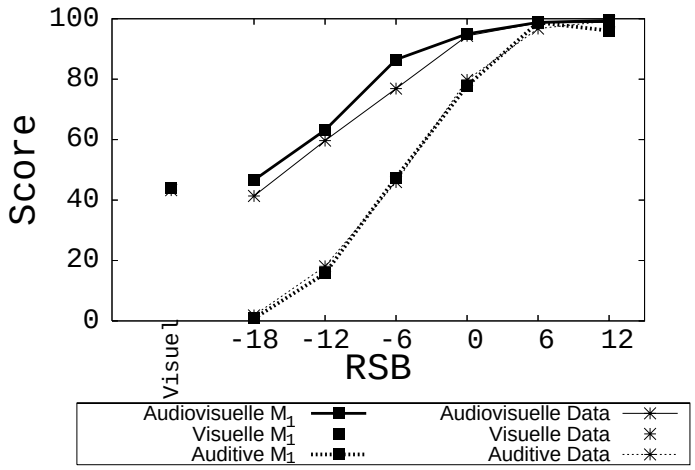

FIG. 3.7: Scores d'identification correcte corrigés des sujets humains (*Data*) et du modèle M_1.

- V_u : variable correspondant au stimuli visuel selon le degré de liberté lié à la dimension arrondissement. Cette variable a le même domaine que A_u.
- V_v : variable correspondant au stimuli visuel selon le degré de liberté lié à la dimension hauteur. Cette variable a le même domaine que A_v.
- V_t : variable correspondant au stimuli visuel selon le degré de liberté lié à la dimension avant-arrière. Cette variable a le même domaine que A_t.
- A : Identique aux modèles précédents.
- V : Identique aux modèles précédents.
- P : Identique aux modèles précédents.
- B : Identique aux modèles précédents.

2 - Décomposition et formes paramétriques

Avec la définition des variables données ci-dessus, la décomposition de la distribution de probabilité conjointe prend la forme suivante :

$$\begin{aligned} &P(A\ V\ A_u\ A_v\ A_t\ V_u\ V_v\ V_t\ P\ B\ M_2) \\ &= \left(\begin{array}{l} P(V_u\ V_v\ V_t \mid P\ M_2)P(A_u\ A_v\ A_t \mid P\ B\ M_2) \\ P(A \mid A_u\ A_v\ A_t\ M_2)P(V \mid V_u\ V_v\ V_t\ M_2)P(P \mid M_2)P(B \mid M_2). \end{array} \right) \end{aligned} \tag{3.15}$$

L'égalité est obtenue avec les mêmes hypothèses d'indépendance conditionnelle que les modèles précédents : nous faisons l'hypothèse que la valeur du stimulus auditif (variables A_u, A_v et A_t) est indépendante de la valeur du stimulus visuel (variables V_u, V_v et V_t), supposant que l'on connaît la voyelle perçue (variable P). Les formes paramétriques associées

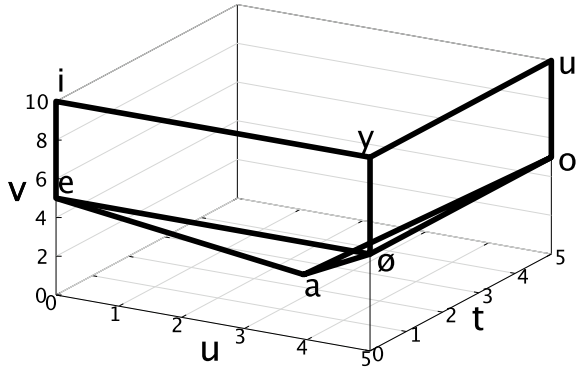

FIG. 3.8: Structure perceptive pour RSB = 12 dB et 6 dB : par exemple, la voyelle /u/ est prononcée en avançant ($t = 5$) et arrondissant ($u = 5$) les lèvres, et en fermant la mâchoire ($v = 10$) (Figure tirée de [13]).

aux différents termes sont les suivantes :

$$\begin{aligned}
P(P \mid M_2) &= \mathbf{U}(P) \\
P(B \mid M_2) &= \mathbf{U}(B) \\
P(A_u\ A_v\ A_t \mid P\ B\ M_2) &= \mathbf{G}_{\mu(P,B),\ \sigma(P,B)}(A_u\ A_v\ A_t) \\
P(V_u\ V_v\ V_t \mid P\ M_2) &= \mathbf{G}_{\mu(P),\ \sigma(P)}(V_u\ V_v\ V_t) \\
P(A \mid A_u\ A_v\ A_t\ M_2) &= \boldsymbol{\delta}_{f(A_u,A_v,A_t)}(A) \\
P(V \mid V_u\ V_v\ V_t\ M_2) &= \boldsymbol{\delta}_{g(V_u,V_v,V_t)}(V).
\end{aligned}$$

Les termes $P(A_u\ A_v\ A_t \mid P\ B\ M_2)$ et $P(V_u\ V_v\ V_t \mid P\ M_2)$ sont associés à des gaussiennes à trois dimensions. Les termes $P(A \mid A_u\ A_v\ A_t\ M_2)$ et $P(V \mid V_u\ V_v\ V_t\ M_2)$ sont la liaison entre les coordonnées sur la structure en trois dimensions et les voyelles correspondantes. Ils sont définis par des ensembles de loi Diracs : pour des valeurs A_u, A_v, A_t données, nous pouvons lire la valeur de la voyelle correspondante et donc prédire A avec certitude. Par exemple, nous avons :

$$P(A \mid [A_u = 0]\ [A_v = 5]\ [A_t = 0]\ M_2) = \begin{cases} 1 \text{ si } A = /e/ \\ 0 \text{ sinon.} \end{cases}$$

3 - Identification

Après avoir spécifié les connaissances préalables de ce modèle, nous avons fixé les valeurs des formes paramétriques associées aux termes de la décomposition et en l'occurrence les moyennes et écart-types des gaussiennes à trois dimensions associées aux termes $P(A_u\ A_v\ A_t \mid P\ B\ M_2)$ et $P(V_u\ V_v\ V_t \mid P\ M_2)$ au vu des données expérimentales.

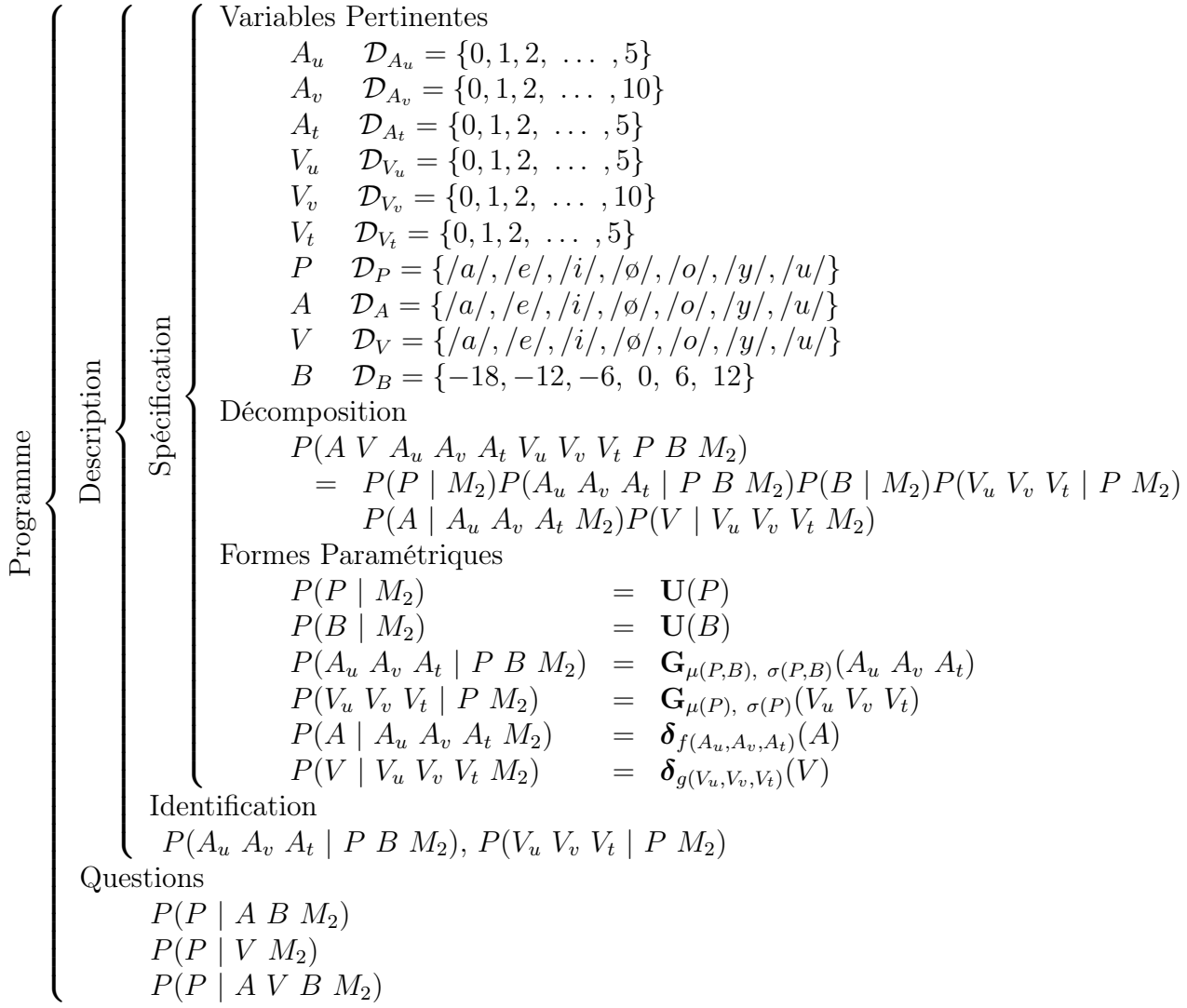

FIG. 3.9: Résumé du modèle M_2.

4 - Questions

De nouveau, nous posons trois questions : nous cherchons à connaître la distribution de probabilité de la voyelle perçue par le sujet connaissant tout d'abord le stimulus auditif et son niveau de bruit (équation (3.16)) puis, le stimulus visuel (équation (3.17)) et enfin connaissant le stimulus auditif avec son niveau de bruit et le stimulus visuel (équation (3.18)).

Question dans le cas auditif pur :

$$P(P \mid A\ B\ M_2). \tag{3.16}$$

Question dans le cas visuel pur :

$$P(P \mid V\ M_2). \tag{3.17}$$

Question dans le cas audiovisuel :

$$P(P \mid A\ V\ B\ M_2). \tag{3.18}$$

L'inférence bayésienne permet de répondre à ces questions à partir de la décomposition choisie précédemment. Nous ne détaillerons dans cette partie que l'inférence dans le cas de la vision pure (équation (3.17)). La même méthode est utilisée pour répondre aux questions dans les cas de l'auditif pur et de l'audiovisuelle.

$$\begin{aligned}
&P(P \mid V\ M_2)\\
&= \frac{P(P\ V \mid M_2)}{P(V \mid M_2)} && (3.19)\\
&= \frac{1}{Z_1} P(P\ V \mid M_2) && (3.20)\\
&= \frac{1}{Z_1} \sum_{A,A_u,A_v,A_t,V_u,V_v,V_t,B} \begin{pmatrix} P(A_u\ A_v\ A_t \mid P\ B\ M_2) P(V_u\ V_v\ V_t \mid P\ M_2) \\ P(A \mid A_u\ A_v\ A_t\ M_2) P(V \mid V_u\ V_v\ V_t\ M_2) \\ P(P \mid M_2) P(B \mid M_2) \end{pmatrix} && (3.21)\\
&= \frac{1}{Z_1} \sum_{A,A_u,A_v,A_t,V_u,V_v,V_t,B} P(P \mid M_2) P(V \mid V_u\ V_v\ V_t\ M_2) P(V_u\ V_v\ V_t \mid P\ M_2) && (3.22)\\
&= \frac{1}{Z_2} \sum_{V_u,V_v,V_t} P(V \mid V_u\ V_v\ V_t\ M_2) P(V_u\ V_v\ V_t \mid P\ M_2) && (3.23)\\
&= \frac{1}{Z_3} \begin{pmatrix} P(V \mid V_u = g_u^{-1}(V)\ V_v = g_v^{-1}(V)\ V_t = g_t^{-1}(V)\ M_2) \\ P(V_u = g_u^{-1}(V)\ V_v = g_v^{-1}(V)\ V_t = g_t^{-1}(V) \mid P\ M_2) \end{pmatrix} && (3.24)
\end{aligned}$$

g_u^{-1}, g_v^{-1} et g_t^{-1} sont les fonctions permettant à une voyelle d'associer les coordonnées u, v et t dans la structure perceptive des voyelles.

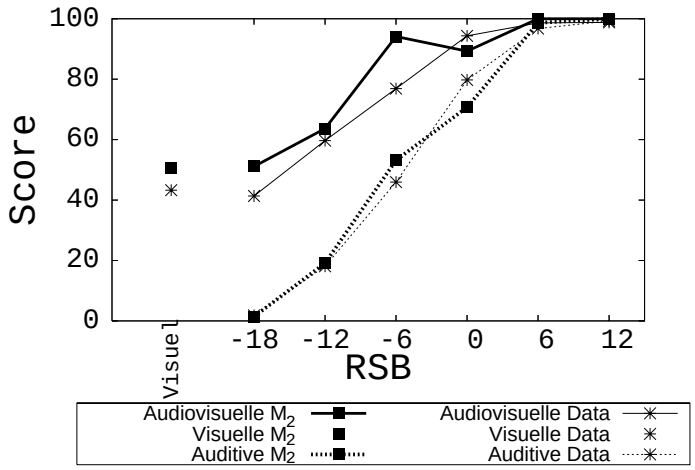

FIG. 3.10: Scores d'identification correcte corrigés des sujets humains (*Data*) et du modèle M_2.

La Figure 3.10 présente le score d'identification correcte corrigé comparé à celui des données.

Nous observons que les scores obtenus par le modèle M_2 sont similaires aux scores des sujets. En annexe D.3 se trouvent les pourcentages d'information transmise par trait (arrondissement, hauteur, avant-arrière). Les pourcentages du modèle M_2 sont globalement similaires à ceux des sujets à l'exception du trait avant-arrière où elles sont supérieures.

3.4 Modèle M_u

Nous avons réalisé un quatrième modèle n'apportant aucune connaissance. Ce modèle a la même structure que les modèles précédents. Il correspond à une distribution uniforme sur l'ensemble des variables du modèle. Afin de pouvoir comparer les modèles M_0, M_1, et M_2, ce modèle servira par la suite de référence.

Ce modèle, présenté dans son intégralité Figure 3.12 est défini sur les mêmes variables que les modèles M_0 et M_1. La décomposition de la distribution de probabilité conjointe est identique et, à chaque terme de la décomposition, nous associons une distribution de probabilité de loi uniforme. Etant donné que M_u n'apporte aucune information, il ne comporte pas de phase d'identification. Il correspond en quelque sorte à n'avoir aucune connaissance sur la perception de la parole humaine. Nous pouvons simuler ce modèle en tirant aléatoirement les réponses dans chaque cas expérimental. Nous obtenons donc des scores très faibles présentés Figure 3.11.

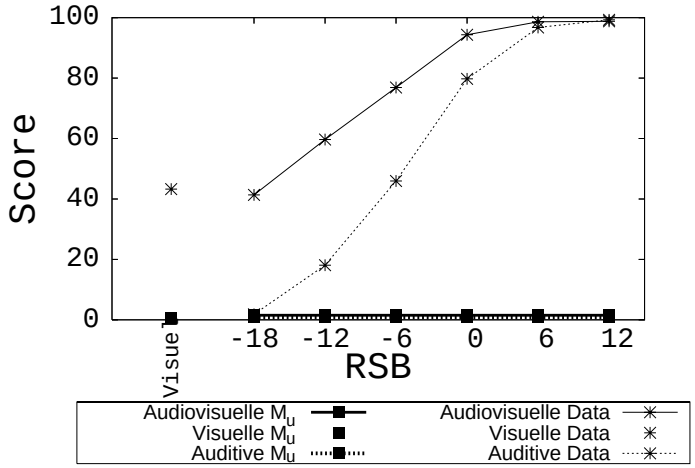

FIG. 3.11: Scores d'identification correcte corrigés des sujets humains (*Data*) et du modèle M_u.

Programme

Description

Spécification

Variables Pertinentes

$A \quad \mathcal{D}_A = \{/a/, /e/, /i/, /\emptyset/, /o/, /y/, /u/\}$

$V \quad \mathcal{D}_V = \{/a/, /e/, /i/, /\emptyset/, /o/, /y/, /u/\}$

$P \quad \mathcal{D}_P = \{/a/, /e/, /i/, /\emptyset/, /o/, /y/, /u/\}$

$B \quad \mathcal{D}_B = \{-18, -12, -6,\ 0,\ 6,\ 12\}$

Décomposition

$$P(A\ V\ P\ B \mid M_u) = P(A \mid P\ B\ M_u)P(V \mid P\ M_u)P(P \mid M_u)P(B \mid M_u)$$

Formes Paramétriques

$$P(P \mid M_u) = \mathbf{U}(P)$$
$$P(B \mid M_u) = \mathbf{U}(B)$$
$$P(A \mid P\ B\ M_u) = \mathbf{U}(A)$$
$$P(V \mid P\ M_u) = \mathbf{U}(V)$$

Identification

Aucun terme à identifier

Questions

$P(P \mid A\ B\ M_u)$

$P(P \mid V\ M_u)$

$P(P \mid A\ V\ B\ M_u)$

FIG. 3.12: Résumé du modèle M_u.

Chapitre 4

Sélection de modèles

Dans le chapitre précédent, nous avons construit trois modèles basés sur des espaces internes différents. Deux modèles sont basés sur des espaces internes mono-dimensionnels, discrets. L'un des espaces est ordonné alors que l'autre ne l'est pas. Le troisième modèle est basé sur un espace à trois dimensions. Jusqu'à présent, pour pouvoir comparer ces modèles, nous avons réalisé une comparaison qualitative des courbes de score. Le but de ce chapitre est de comparer ces modèles de manière quantitative. Pour cela, nous définissons un modèle π englobant tous les modèles définis précédemment. Ceci est possible grâce à la définition d'une variable $M = \{M_0, M_1, M_2, M_u\}$ qui permet de regrouper les modèles individuels dans un modèle π. Ce modèle π étant défini, il nous permet de calculer et de comparer les vraisemblances des données expérimentales selon chaque modèle M_0, M_1, M_2 et M_u. En pratique, nous nous ramènerons toujours à comparer deux modèles entre eux, en utilisant M_u comme référence (calcul des ratios de vraisemblances). Ce chapitre est divisé en deux parties correspondant aux deux cas étudiés. Tout d'abord, nous comparons les modèles dans un cas de non conflit, c'est-à-dire un cas où le stimulus auditif et le stimulus visuel correspondent à la même voyelle. Puis, nous nous plaçons dans un cas où les stimuli auditif et visuel sont contradictoires.

4.1 En cas de non conflit

Afin de comparer les modèles, nous réalisons un nouveau modèle π regroupant les quatre modèles présentés précédemment. Ce modèle présenté dans son intégralité Figure 4.1 est détaillé dans le reste de cette section.

1 - Les variables pertinentes

Pour la réalisation de ce modèle nous avons utilisé cinq variables : soit les variables A, V, P, B définies comme précédemment et soient M la variable correspondant aux quatre modèles étudiés dans le chapitre précédent. M peut donc prendre quatre valeurs : $\{M_0, M_1, M_2, M_u\}$.

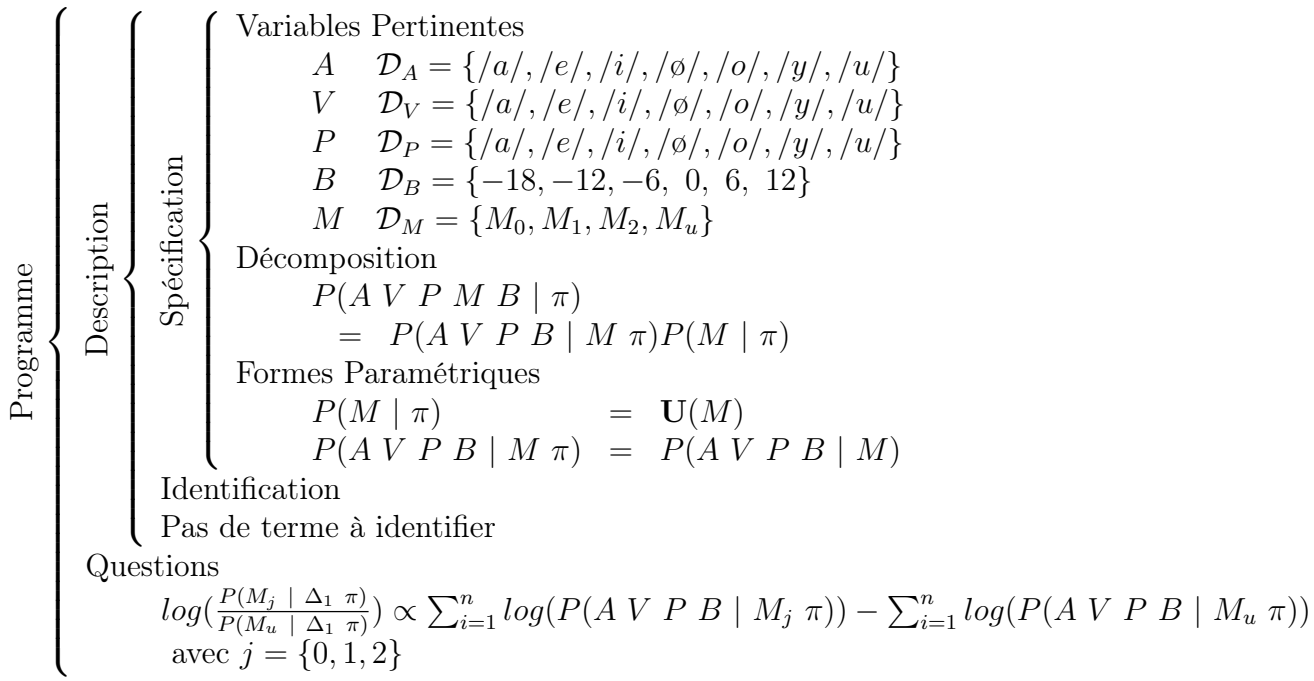

FIG. 4.1: Résumé du modèle π *Sélection sans conflit.*

2 - Décomposition et formes paramétriques

Avec la définition des variables données ci-dessus, la décomposition de la distribution de probabilité conjointe prend la forme suivante :

$$P(A\ V\ P\ M\ B \mid \pi) = P(A\ V\ P\ B \mid M\ \pi)P(M \mid \pi). \tag{4.1}$$

Les formes paramétriques associées aux différents termes sont :

$$\begin{aligned} P(M \mid \pi) &= \mathbf{U}(M) \\ P(A\ V\ P\ B \mid M\ \pi) &= P(A\ V\ P\ B \mid M). \end{aligned}$$

Au terme $P(M \mid \pi)$, nous attribuons une distribution de probabilité de loi uniforme car nous n'avons aucun a priori sur le modèle à appliquer. Les termes $P(A\ V\ P\ B \mid M\ \pi)$ sont des questions probabilistes posées aux modèles précédents.

3 - Questions

Nous cherchons à déterminer le modèle le plus pertinent connaissant les données expérimentales. Soit (Δ_1) l'ensemble des données expérimentales :

$$\Delta_1 = \{\delta_{1_i}\}_{i=1}^{n} = \{\langle a_i, v_i, b_i, p_i \rangle\}_{i=1}^{n}.$$

La question correspond à :

$$P(M \mid \Delta_1\ \pi) \tag{4.2}$$

L'inférence bayésienne nous permet de répondre à cette question en appliquant la règle de Bayes :

$$\begin{aligned} P(M \mid \Delta_1\ \pi) = & \quad \frac{P(\Delta_1 \mid M\ \pi)P(M \mid \pi)}{P(\Delta_1 \mid \pi)} && (4.3)\\ = & \quad \frac{1}{Z_1}P(\Delta_1 \mid M\ \pi) && (4.4)\\ = & \quad \frac{1}{Z_1}\prod_{i=1}^{n} P(\delta_{1_i} \mid M\ \pi) && (4.5) \end{aligned}$$

4 - Résultats

Pour des raisons de calcul numérique, il est difficile d'estimer $\prod_{i=1}^{n} P(\delta_{1_i} \mid M\ \pi)$. Une technique courante consiste à calculer à la place le logarithme du ratio des vraisemblances de deux modèles. Nous calculons donc, en prenant M_u comme référence de ce calcul :

$$\begin{aligned} & log(\frac{P(M = M_j \mid \Delta_1\ \pi)}{P(M = M_u \mid \Delta_1\ \pi)}), \text{pour } j = \{0, 1, 2\}\\ = & \quad log(P(M = M_j \mid \Delta_1\ \pi)) - log(P(M = M_u \mid \Delta_1\ \pi))\\ \propto & \quad log(\prod_{i=1}^{n} P(\delta_{1_i} \mid M_j\ \pi)) - log(\prod_{i=1}^{n} P(\delta_{1_i} \mid M_u\ \pi))\\ \propto & \quad \sum_{i=1}^{n} log(P(A\ V\ P\ B \mid M_j\ \pi)) - \sum_{i=1}^{n} log(P(A\ V\ P\ B \mid M_u\ \pi)). \end{aligned}$$

Le Tableau 4.1 présente les résultats obtenus par chacun des modèles. Pour chaque ligne la mesure varie entre $-\infty$ (pour le modèle théorique le plus mauvais) et 8582 pour le modèle théorique qui prédit parfaitement les données expérimentales. La mesure vaut 0 pour le modèle qui tire les voyelles selon une distribution de probabilité uniforme dans tous les cas. Les lignes représentent les différents niveaux de bruit et les colonnes les trois modèles que nous comparons. La dernière ligne du tableau (*Total*) présente le score total de chaque modèle quelque soit le niveau de bruit.

Nous remarquons que le modèle M_1 est le modèle le plus pertinent. Quelque soit le niveau de bruit, le modèle M_0 est moins bon que le modèle M_1. Ces résultats nous permettent de douter de l'efficacité d'ordonner l'espace des voyelles lorsqu'il est mono-dimensionnel. De plus, nous observons que le modèle M_2, modèle avec l'espace interne à trois dimensions, est plus pertinent que M_0 mais moins pertinent que M_1.

	M=M_0	M=M_1	M=M_2
RSB= 12	2 841	4 491	3 673
RSB = 6	2 914	4 475	4 149
RSB = 0	2 986	4 052	3 498
RSB = -6	1 593	2 999	2 460
RSB = -12	988	2 092	1 617
RSB = -18	1 033	1 503	1 156
Total	12 355	19 612	16 553

TAB. 4.1 – Comparaison de l'adéquation des modèles sur les données de Robert-Ribes.

4.2 En cas de conflits

Un conflit visuo-acoustique très connu est l'effet McGurk. Pour montrer l'effet McGurk, on présente une vidéo montrant une personne prononçant un phonème (par exemple /ga/) alors que la bande sonore diffuse l'enregistrement d'un autre phonème (par exemple /ba/). Plusieurs cas se présentent alors, le cas d'une réponse « auditive », le cas d'une réponse « visuelle » et plus couramment, le cas d'un percept « fusion », qui n'est semblable ni à la catégorie auditive ni à la catégorie visuelle mais constitue une sorte de compromis entre les deux modalités ($/ba/_A + /ga/_V \rightarrow /da/_{AV}$).

Ici, nous étudions des cas de conflits visuo-acoustique dans le cas de reconnaissance de voyelles. Nous utilisons les données tirées d'une expérience de Lisker et Rossi [12]. Les auteurs ont étudié le pourcentage d'arrondissement perçu par les sujets : une image et/ou un son sont présentés aux différents sujets qui doivent déterminer si la voyelle est arrondie ou non. La Figure 4.2 présente le pourcentage d'arrondissement perçu par les sujets dans le cas de stimuli audiovisuels congruents comparé à des stimuli auditif pur et visuel pur.

La Figure 4.3 résume les réponses dans le cas de conflits audiovisuels. Les sujets devaient de même déterminer si la voyelle était arrondie ou non mais cette fois lorsque l'image présentée ne correspondait pas au son présenté. Les auteurs ont observé que certains sujets sont plus sensibles à l'auditif (leurs réponses ne se basent que sur les stimuli auditifs) alors que d'autres sont plus sensibles aux stimuli visuels.

Pour comparer les modèles M_0, M_1, M_2 et M_u dans cette situation, nous définissons un modèle π_2, qui, comme précédemment, englobe M_0, M_1, M_2 et M_u que nous enrichissons par une variable R et un terme $P(R \mid P)$ permettant de modéliser le degré d'arrondissement des voyelles. Ce terme est déterminé de la façon suivante. Pour chaque voyelle nous déterminons le pourcentage d'arrondissement suivant les données expérimentales. Nous prenons le pourcentage fourni par les sujets dans les cas de stimuli audiovisuels non conflictuels. Par exemple, si l'on prend la voyelle /ø/, nous obtenons la distribution de probabilité présentée Figure 4.4.

Cette section présente un modèle (Figure 4.5) se basant sur les quatre modèles M_0, M_1, M_2, M_u et sur la variable d'arrondissement R.

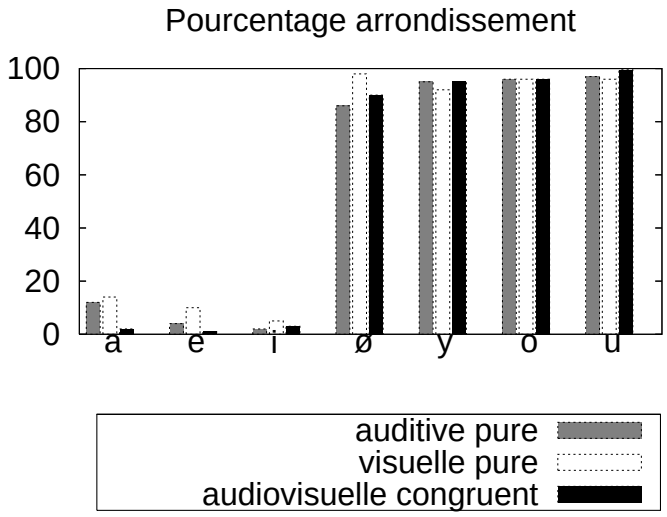

FIG. 4.2: Jugement sur l'arrondissement dans le cas de stimuli audiovisuels comparé aux réponses en auditif pur et visuel pur.

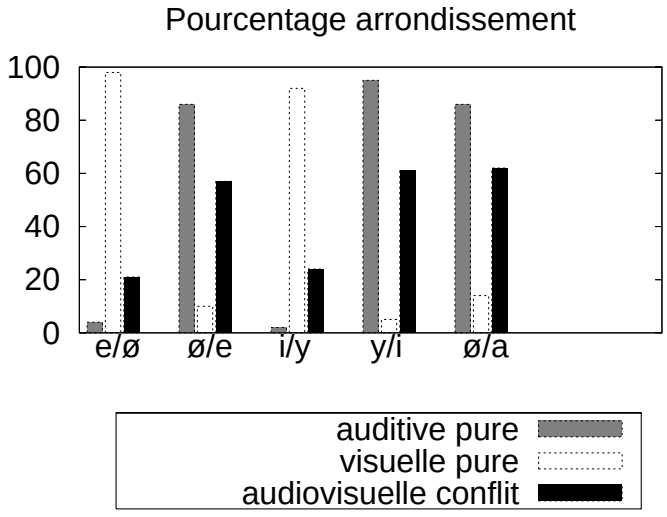

FIG. 4.3: Jugement sur l'arrondissement dans le cas conflits audiovisuels comparé aux réponses en auditif pur et visuel pur.

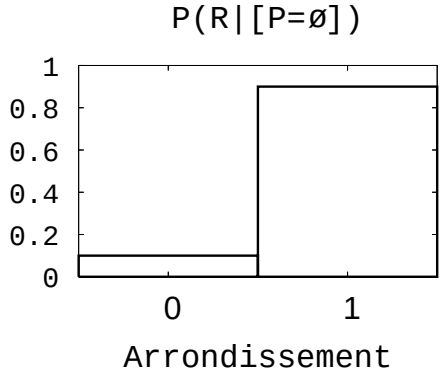

FIG. 4.4: Distribution de probabilité sur l'arrondissement pour la voyelle /ø/.

1 - Variables pertinentes

Pour la réalisation de ce modèle nous avons utilisé cinq variables : Soit les variables A, V, P, M définies comme précédemment. Nous rajoutons une variable R correspondant au pourcentage d'arrondissement de la variable perçue. Cette variable peut prendre 2 valeurs : 0 correspond à une variable non arrondie et 1 à une variable arrondie.

2 - Décomposition et formes paramétriques

Avec la définition des variables données ci-dessus, la décomposition de la distribution de probabilité conjointe prend la forme suivante :

$$P(A\ V\ P\ M\ B\ R \mid \pi_2) = P(A\ V\ P\ B\ R \mid M\ \pi_2)P(M \mid \pi_2). \tag{4.6}$$

Les formes paramétriques sont définies comme pour le modèle π, c'est-à-dire une distribution uniforme associée au terme $P(M \mid \pi_2)$ et une question probabiliste associée au terme $P(A\ V\ P\ B\ R \mid M\ \pi_2)$

3 - Questions

Comme précédemment, nous cherchons à déterminer le modèle le plus pertinent connaissant les données expérimentales. Soit (Δ_2) l'ensemble des données expérimentales :

$$\Delta_2 = \{\delta_{2_i}\}_{i=1}^{n} = \{\langle a_i, v_i, r_i, p_i\rangle\}_{i=1}^{n}$$

La question correspond à :

$$P(M \mid \Delta_2\ \pi_2). \tag{4.7}$$

Programme

Description

Spécification

Variables Pertinentes

$A \quad \mathcal{D}_A = \{/a/, /e/, /i/, /\emptyset/, /o/, /y/, /u/\}$

$V \quad \mathcal{D}_V = \{/a/, /e/, /i/, /\emptyset/, /o/, /y/, /u/\}$

$P \quad \mathcal{D}_P = \{/a/, /e/, /i/, /\emptyset/, /o/, /y/, /u/\}$

$B \quad \mathcal{D}_B = \{-18, -12, -6,\ 0,\ 6,\ 12\}$

$M \quad \mathcal{D}_M = \{M_0, M_1, M_2, M_u\}$

$R \quad \mathcal{D}_R = \{0, 1\}$

Décomposition

$$P(A\ V\ P\ M\ B\ R \mid \pi_2) = P(A\ V\ P\ B\ R \mid M\ \pi_2) P(M \mid \pi_2)$$

Formes Paramétriques

$$P(M \mid \pi_2) = \mathbf{U}(M)$$

$$P(A\ V\ P\ B\ R \mid M\ \pi_2) = P(A\ V\ P\ B\ R \mid M)$$

Identification

Pas de terme à identifier

Questions

$$log(\tfrac{P(M_j \mid \Delta_2\ \pi_2)}{P(M_u \mid \Delta_2\ \pi_2)}) \propto \sum_{i=1}^{n} log(P(A\ V\ R\ B \mid M_j\ \pi_2)) - \sum_{i=1}^{n} log(P(A\ V\ R\ B \mid M_u\ \pi_2))$$

avec $j = \{0, 1, 2\}$

FIG. 4.5: Résumé du modèle π_2 *Sélection avec conflits.*

Nous déterminons le modèle le plus pertinent en prenant M_u comme référence de ce calcul,

$$log(\frac{P(M = M_j \mid \Delta_2\ \pi_2)}{P(M = M_u \mid \Delta_2\ \pi_2)}), \text{pour } j = \{0, 1, 2\}$$
$$\propto \sum_{i=1}^{n} log(P(A\ V\ R\ B \mid M_j\ \pi_2)) - \sum_{i=1}^{n} log(P(A\ V\ R\ B \mid M_u\ \pi_2))$$

Le Tableau 4.2 présente les résultats obtenus par chacun des modèles. Pour chaque ligne la mesure varie entre$-\infty$ (pour le modèle théorique le plus mauvais) et 458 pour le modèle théorique qui prédit parfaitement les données expérimentales. Les lignes représentent les différents cas de conflits et les colonnes les trois modèles que nous comparons. La dernière ligne du tableau (*Total*) présente le score total de chaque modèle quelque soit le conflit.

	M=M_0	M=M_1	M=M_2
e/ø	84	-26	64
ø/e	144	101	105
i/y	116	137	65
y/i	151	53	78
ø/a	-46	-72	-889
Total	449	193	- 577

TAB. 4.2 – Comparaison de l'adéquation des modèles sur les données de Lisker et Rossi.

4.3 Avec et sans conflits

Dans cette section, nous cherchons à déterminer le modèle le plus pertinent connaissant l'ensemble des données expérimentales. Soit Δ l'ensemble des données expérimentales regroupant les données expérimentales de la thèse de Robert-Robes et les données expérimenatales de l'expérience de Lisker et Rossi :

$$\Delta = \{\delta_k\}_{k=1}^{p} = \{\delta_{1_i}\}_{i=1}^{n} + \{\delta_{2_t}\}_{t=1}^{m} \tag{4.8}$$

Nous cherchons donc à déterminer :

$$log(\frac{P(M = M_j \mid \Delta)}{P(M = M_u \mid \Delta)}), \text{pour } j = \{0, 1, 2\} \tag{4.9}$$

Par dérivation, nous obtenons :

$$
\begin{aligned}
log(\frac{P(M = M_j \mid \Delta)}{P(M = M_u \mid \Delta)}) &= log(\frac{\prod_k P(\delta_k \mid M = M_j)}{\prod_k P(\delta_k \mid M = M_u)}) \\
&= \sum_k log(P(\delta_k \mid M = M_j)) - \sum_k log(P(\delta_k \mid M = M_u)) \\
&= \sum_i log(P(\delta_{1_i} \mid M = M_j\ \pi)) + \sum_t log(P(\delta_{2_t} \mid M = M_j\ \pi_2)) \\
&- \sum_i log(P(\delta_{1_i} \mid M = M_u\ \pi)) - \sum_t log(P(\delta_{2_t} \mid M = M_u\ \pi_2))
\end{aligned}
$$

δ_1 correspond aux données expérimentales dans les cas de congruence et δ_2 aux données expérimentales dans les cas de conflits.

En normalisant les données des deux Tableaux 4.1 et 4.2 par rapport à la proportion des données expérimentales, nous obtenons le Tableau 4.3. Nous attribuons le même « poids » à chacune des données expérimentales des deux expériences. Ce tableau nous permet donc de comparer les trois modèles dans les cas de non conflits et de conflits.

	M=M_0	M=M_1	M=M_2
« Non conflit normalisé »	1.40	2.22	1.88
« Conflits normalisés »	0.90	0.39	-1.15
Total	2.3	2.61	0.73

TAB. 4.3 – Total des scores sur l'ensemble des données expérimentales.

Dans le Tableau 4.2, nous remarquons que tous les modèles ont des scores très faibles dans le dernier cas de conflit ø/a. Par conséquent, nous avons choisi de reprendre le Tableau 4.3 en ne tenant pas compte de ce résultat. Les scores obtenus sont présentés Tableau 4.4.

	M=M_0	M=M_1	M=M_2
« Non conflit normalisé »	1.40	2.22	1.88
« Conflits normalisés »	1.24	0.66	0.78
Total	2.64	2.88	2.66

TAB. 4.4 – Total des scores sur l'ensemble des données expérimentales moins la dernière ligne de conflit.

Chapitre 5

Discussion

Dans le chapitre 3, nous avons étudié quatre modèles de perception audiovisuelle des voyelles. Pour chaque modèle, nous avons simulé une expérience de test audiovisuel ce qui nous a permis de calculer pour chacun les scores d'identification correcte corrigés. Les modèles reproduisent les données expérimentales étant donné que ces scores sont similaires aux scores obtenus par les sujets. Cela nous amène à conclure que la modélisation bayésienne basée sur la fusion capteur est propice au domaine visuo-acoustique pour la reconnaissance de phonèmes.

Le chapitre 4 propose une méthode de sélection de modèles afin de déterminer le plus pertinent au vu des données des expériences de perception audiovisuelle. Les trois principaux modèles que nous avons confronté sont basés sur des espaces internes différents. La sélection de modèles nous permet d'effectuer une comparaison quantitative des modèles et donc de réfléchir sur l'espace interne le plus pertinent.

La fin de cette section est consacrée à l'analyse des résultats de la sélection de modèles.

1 - En cas de non conflit Nous avons analysé les performances des modèles de deux manières.

D'une part, nous avons simulé chacun des modèles sur les conditions expérimentales, ce qui nous a permis de recréer les courbes de scores d'identification correcte corrigés. Ainsi, nous avons pu juger, qualitativement, de l'adéquation des modèles aux données. En effet, nous avons noté que chacun des modèles généraient des courbes de scores assez proches des courbes de scores des sujets.

D'autre part, pour rendre cette comparaison qualitative plus précise, nous avons ensuite appliqué une méthode de sélection de modèles bayésienne. Elle consiste à quantifier l'adéquation d'un modèle aux données par un rapport de probabilités. Ce rapport divise les probabilités des données selon ce modèle par les probabilités des données selon le modèle « uniforme ». Ce rapport quantifie précisément l'adéquation aux données : il est d'autant plus grand que le modèle « colle » mieux aux données.

Nous observons (Tableau 4.1) que M_1 est le plus performant selon ce critère, M_2 est le deuxième meilleur modèle, et M_0 est le moins bon. Cet ordre est conservé quelque soit le niveau de bruit considéré.

Or, le modèle M_1 est celui qui fait le moins d'hypothèses sur la nature de l'espace interne : dans M_1, il est supposé discret et non-ordonné. Nous constatons donc que l'hypothèse de fusion des informations mono-modales par un mécanisme de fusion capteur est suffisant pour expliquer correctement les données. En d'autres termes, les hypothèses supplémentaires sur l'espace interne, faites par M_0 et M_2, n'apportent pas d'amélioration sur l'explication des données. Au contraire même, les performances de ces deux modèles sont dégradées par rapport à M_1.

Nous observons de plus que les probabilités d'identification trait par trait sont bonnes pour M_1, c'est-à-dire qu'elles collent très bien aux données. Ceci peut paraître contre intuitif. En effet, nous analysons ici les réponses des sujets et des modèles en cas d'erreur sur la reconnaissance de la voyelle prononcée. Lorsqu'ils se trompent, les sujets le font selon des schémas d'erreur particuliers. Par exemple, lorsqu'un /i/ est prononcé, dans des conditions de bruit, les sujets reconnaissent essentiellement soit un /i/, soit un /e/. Ces deux voyelles sont similaires vis-à-vis du trait d'arrondissement. Cette proximité entre les deux voyelles n'est a priori pas traduite dans le modèle M_1, qui ne suppose aucune relation entre les voyelles. M_1 ne devrait donc pas être capable de reproduire les erreurs des sujets dans ce cas. Or, ces erreurs sont en fait déjà présentes dans les données expérimentales des cas auditifs purs et visuels purs. Ces données sont très bien traduites dans les paramètres du modèle M_1. La fusion capteur permet ensuite de reproduire les schémas de confusions audiovisuelles, sans faire d'hypothèses supplémentaires sur la proximité des voyelles.

Nous observons enfin que le score global de M_0 est le plus faible des trois modèles. Ce résultat pourrait être dû à l'arbitraire dans l'ordre choisi pour l'espace des voyelles. En effet, nous avons choisi un ordre basé sur l'arbre de confusions audiovisuelles établi expérimentalement par Robert-Ribes. Cependant, nous remarquons que cet arbre n'implique pas un ordre unique sur les voyelles : l'arbre comporte des sous-branches qui pourraient être permutées. Ainsi, il y a trois sous-branches indépendantes : d'une part, /a/ seul, /e/ et /i/ ensuite, et /ø/, /y/, /o/, /u/ enfin. Il existe peut-être une permutation de ces branches plus pertinente que les autres, et qui permettrait à M_0 de voir son score global remonter.

2 - En cas de conflits Le Tableau 4.2 présente les résultats obtenus par chacun des modèles dans des cas de conflits audiovisuels. Les modèles sont basés sur les données de l'expérience de Robert-Ribes et testés sur le test de perception audiovisuelle de Lisker et Rossi.

Nous observons que le modèle M_1, modèle très performant dans les cas de non conflits, a des résultats assez faibles dans les cas de conflits. Dans ce modèle, nous avons supposé qu'il n'y avait aucune relation entre les voyelles. Or, nous pouvons supposer que dans les cas de conflits, c'est la relation entre les voyelles qui est en jeu. Ceci pourrait expliquer les faibles scores du modèle M_1. Dans les modèles M_0 et M_2, il existe une relation d'ordre entre les voyelles. Nous observons que le modèle M_0 est le modèle le plus performant dans le test de perception dans le cas de conflits.

Nous observons Tableau 4.2 que les trois modèles sont mauvais lors du cas de conflit ø/a. C'est même le seul cas expérimental où les trois modèles sont plus mauvais que le modèle uniforme M_u, de référence. Ce cas de conflit concerne une présentation visuelle d'un /a/ avec le signal auditif d'un /ø/. Or, le /a/ est très bien reconnu visuellement. Mathématiquement, cela se traduit dans nos trois modèles par des probabilités très « piquées », assez proche de lois Dirac. Ceci empêche une généralisation dans les cas de conflits, c'est-à-dire la possibilité d'avoir des réponses intermédiaires, entre le /a/ et le /ø/.

Nous observons de plus un score très mauvais pour le modèle M_2 dans le cas de ce conflit ø/a. Nous avons basé l'espace interne utilisé par M_2 sur une structure géométrique tri-dimensionnelle. Les dimensions de cette structure traduisent trois paramètres articulatoires : « avant-arrière », « arrondissement » et « hauteur ». Quand un sujet prononce un /a/, la bouche est ouverte, les deux autres dimensions ne sont plus pertinentes (difficile de juger de l'arrondissement des lèvres d'un sujet bouche ouverte). Par convention, nous avons placé la voyelle /a/ à des positions fixes sur les coordonnées u et t, mais ce choix est arbitraire. Ce choix a cependant une grande influence sur le résultat du conflit ø/a, car il conditionne les possibilités de réponses intermédiaires. Ceci pourrait expliquer le très mauvais résultat de M_2 dans ce cas expérimental.

3 - Comparaison cas de non conflit et en cas de conflits En donnant le même « poids » aux deux expériences de Robert-Ribes et Lisker et Rossi, nous avons proposé une comparaison globale des trois modèles. Cette comparaison indique une performance finale pour le modèle M_1 légèrement supérieure aux performances de M_0 et M_2, qui obtiennent des scores comparables si l'on omet le cas de conflit ø/a.

Plus précisément, M_1 est de loin le plus performant hors des cas de conflits, et M_0 et M_2 sont meilleurs pour les cas de conflits. M_1 « colle » le mieux au premier jeu de données, mais est le moins capable de généraliser. Nous comprenons du coup l'intérêt de comparer les modèles sur plusieurs jeux de données expérimentales. Dans le cas contraire, nous aurions pu conclure, au vu de la première expérience que le modèle M_0 était le plus pertinent et, au vu de la seconde expérience, que M_1 était le plus pertinent. Nous en concluons que les deux expériences sont trop spécifiques individuellement pour pouvoir conclure sur la validité de chacun des modèles.

Chapitre 6

Conclusion

6.1 Synthèse

Dans ce travail, nous avons réalisé une étude bibliographique permettant de conforter l'hypothèse générale selon laquelle le modèle mathématique de fusion capteur décrit convenablement le principe de fusion d'informations sensorielles chez l'humain.

Dans la suite du travail, nous avons réalisé deux objectifs.

Le premier était de créer des modèles de perception audiovisuelle de reconnaissance des voyelles. Ces modèles sont basés sur le formalisme bayésien et la fusion capteur. Les trois modèles proposés sont basés sur différents espaces internes : deux espaces monodimensionnels avec ou non une relation d'ordre sur les voyelles, et un espace tri-dimensionnel basé sur des dimensions articulatoires. Les scores de reconnaissance correcte nous ont permis de valider qualitativement chacun des modèles.

Le second objectif était d'appliquer la méthode de sélection bayésienne de modèles pour pouvoir discuter de la pertinence de chacun de ces modèles, et notamment de l'espace interne le plus probant. Cette comparaison de modèles originale est un guide pour l'amélioration des modèles. De plus, la sélection de modèles a été appliquée à la fois sur les données expérimentales sur lesquelles les modèles sont basés et sur les données d'une autre expérience.

6.2 Perspectives

Cette section présente les pistes de recherche que nous dégageons suite à notre travail. Nous présentons d'abord les pistes théoriques liées à la modélisation et à la sélection bayésienne de modèles, puis nous proposons de nouvelles expériences afin de confirmer ou d'infirmer certaines de nos prédictions expérimentales.

6.2.1 Enjeux théoriques

Amélioration des modèles

Une première perspective de travail est de chercher à améliorer les trois modèles présentés. Par exemple, pour le modèle M_0, il semblerait intéressant de modifier l'ordre des voyelles fixé par l'arbre de confusions audiovisuelles. Nous pouvons, par exemple, utiliser l'arbre de confusions auditives ou visuelles pour déterminer la relation d'ordre. Nous pouvons également utiliser les permutations possibles entre les branches de ces arbres.

Une seconde perspective est l'amélioration du modèle M_2. Nous avons observé, Tableau 4.2, que ce modèle est assez bon lors des cas de conflits sauf pour le dernier cas de conflit ø/a. Ce score très faible s'explique par la mauvaise position de /a/ dans la structure que nous avons utilisée. En effet, pour cette voyelle, seul le trait de hauteur est connu. Une amélioration consisterait donc à laisser libre la position de /a/ sur les traits d'arrondissement et avant-arrière et d'intégrer sur toutes les positions possibles. Techniquement, cela pourrait être exprimé naturellement dans le formalisme bayésien, par des distributions uniformes pour la position de /a/ selon les dimensions arrondissement et avant-arrière : $P(A_u\ A_t \mid [A = /a/]\ M_2) = \mathbf{U}(A_u, A_t)$.

Espace interne moteur

Une perspective de recherche est de proposer un nouveau modèle, qui, comme M_2, aurait un espace interne tri-dimensionel. Cependant, cet espace interne serait de nature différente. Plus particulièrement, M_2 se base sur des dimensions traduisant des aspects articulatoires perçus : l'arrondissement des lèvres, par exemple, est bien un indice perçu par le sujet. Nous pourrions baser notre nouveau modèle sur des dimensions articulatoires de nature plus motrices, comme par exemple la position du point le plus haut de la langue et l'aire intérolabiale. Ce travail élargirait la portée de nos résultats, pour nous permettre de nous positionner dans un débat théorique général sur la nature purement sensorielle, ou sensori-motrice, de la perception ([14], [16]).

Comparaison des modèles

Dans ce travail, nous avons fixé chaque paramètre des modèles puis comparé les modèles. Une piste intéressante de recherche serait donc de laisser libre les paramètres et, lors de leur comparaison, d'intégrer sur l'espace des paramètres. Alors nous ne comparerions plus des modèles particuliers mais des classes de modèles. La comparaison sélectionnerait ainsi la classe de modèles la plus adéquate. Ceci permettrait de conclure plus aisément sur les hypothèses les plus pertinentes faites par les différentes classes de modèles.

Nous souhaitons également réaliser une étude sur l'influence du nombre de paramètres de chaque modèle sur son adéquation aux données. En effet, d'un point de vue théorique, notre modèle M_1 se situe exactement à la limite du sur-apprentissage, car il a autant de paramètres qu'il y a de cas expérimentaux. Nous souhaitons replacer cette question théorique comme une considération centrale. Par exemple, nous pourrions commencer cette

piste de recherche par la question suivante : pourquoi M_2, qui dispose de plus de paramètres que M_1, généralise mieux que ce dernier ?

6.2.2 Proposition de protocoles expérimentaux

Notre travail de modélisation, outre les avancées sur les enjeux théoriques, nous permet également de proposer des améliorations ou des compléments expérimentaux.

Une perspective de recherche est ainsi de complémenter expérimentalement les données sur les cas de conflits. En effet, les données de Lisker et Rossi n'ont qu'une petite intersection avec celles de Robert-Ribes, ce qui nous a limité dans notre étude des cas de conflits. Le Tableau 4.2 ne comporte que cinq cas de conflits. D'autres cas de conflits apparaissant dans le travail de Lisker et Rossi concernent des voyelles qui ne sont pas dans les données de l'expérience de Robert-Ribes. Par exemple, en plaçant la voyelle /œ/ comme intermédiaire, en terme d'arrondissement, entre /e/ et /ø/, nous pourrions simuler, dans nos modèles, le conflit œ/ϵ. De même, à l'inverse, certains des conflits que nous pouvons simuler à partir de l'expérience de Robert-Ribes n'ont pas été testés dans l'expérience de Lisker et Rossi.

Nous pourrions en particulier simuler des cas de conflits mettant en jeu la voyelle /a/, afin de pouvoir déterminer s'il a des coordonnées fixes dans les dimensions u et t, ou si, au contraire, sa position y est indéterminée. Nous possédons en effet un outil de formalisation pouvant traduire naturellement ces deux possibilités théoriques, par l'emploi de distributions uniformes ou de lois Diracs. L'expérience pourrait ainsi venir trancher entre ces deux possibilités théoriques.

Nos modèles permettent de prédire la voyelle perçue par les sujets dans les cas de conflits. Cela est plus précis que ce qui est contenu dans les données issues de l'expérience de Lisker et Rossi, qui ne concernent que le degré d'arrondissement perçu. Nous pourrions ainsi proposer une expérience dont le mode de collecte des données nous permettrait de les comparer à nos prédictions expérimentales sur les voyelles perçues.

Enfin, nous notons que nos modèles de fusion capteur robotique sont symétriques sur les capteurs employés. En d'autres termes, l'opération de fusion prend en compte les incertitudes sur les deux canaux pour produire une estimation finale. Dans l'expérience de Robert-Ribes, seule l'information auditive était bruitée expérimentalement. Nous proposons un complément d'expérience visant à bruiter le canal visuel, et comparer les résultats expérimentaux à notre modèle de fusion capteur.

Annexe A

Formes paramétriques

Loi uniforme C'est une forme paramétrique à un paramètre, qui est le nombre de cas de la variable sur laquelle porte cette distribution. Cette loi indique l'équi-probabilité de tous les cas ; elle sera choisie en général pour modéliser notre ignorance sur un phénomène. Dans notre formalisme, la définition des variables est statique, donc cette loi n'évoluera pas en fonction des données expérimentales. Pour une variable $V : D_V$, k_V, nous notons :

$$\begin{aligned} P(\ V \mid \delta\ \pi) &= P(\ V \mid \pi) = \mathbf{U}_{\mathbf{k_V}}(V) \\ P([V\ =\ v_i] \mid \pi) &= \frac{1}{k_V}. \end{aligned}$$

Loi Dirac Cette forme paramétrique classique possède un paramètre, qui est une valeur particulière n de la variable sur laquelle elle porte : pour cette valeur donnée, la probabilité est 1, et 0 pour toutes les autres valeurs. Nous utilisons la notation standard :

$$\begin{aligned} P(\ V \mid \delta\ \pi) &= \boldsymbol{\delta}_n(V) \\ P([V\ =\ v_i] \mid \pi) &= \begin{cases} 1 \text{ si } v_i = n_i \\ 0 \text{ sinon.} \end{cases} \end{aligned}$$

Loi normale ou gaussienne Cette forme paramétrique classique correspond au choix de ne mémoriser des données expérimentales que la moyenne des valeurs rencontrées et leur écart type. Nous notons :

$$\begin{aligned} P(\ V \mid \delta\ \pi) &= \mathbf{G}_{\mu,\sigma}(V) \\ P([V\ =\ v_i] \mid \pi) &= \int_{v_i-\epsilon}^{v_i+\epsilon} \frac{1}{\sqrt{2\Pi\sigma}} e^{\frac{(v-\mu)^2}{2\sigma^2}}. \end{aligned}$$

Où μ est la moyenne de V dans les données expérimentales δ, et σ est l'écart type de V dans δ.

Loi de succession de Laplace Cette forme a autant de paramètres que de cas de la variable sur laquelle elle porte ; ces paramètres se calculent facilement au vu des données expérimentales. La distribution résultat ressemble fortement à une normalisation du nombre d'occurrences n_i de chaque cas dans les données expérimentales. Nous notons :

$$\begin{aligned} P(\ V \mid \delta\ \pi) &= \mathbf{L}_{\mathbf{n_1},\ldots,\mathbf{n_{k_V}}}(V) \\ P([V\ =\ v_i] \mid \pi) &= \frac{n_i + 1}{N + k_V}. \end{aligned}$$

Où n_i est le nombre de fois où $[V = v_i]$ dans les données δ, et N est le nombre total de données dans δ.

Question à une autre description Cette forme paramétriques sans paramètres libres, consiste à poser une question probabiliste à une autre description π' pour obtenir un terme. Nous notons :

$$\begin{aligned} P(\ V \mid \delta\ \pi) &= P(\ V \mid \pi) = P(\ V \mid \pi') \\ P([V\ =\ v_i] \mid \pi) &= P([V\ =\ v_i] \mid \pi'). \end{aligned}$$

Annexe B

Matrices de confusions

Ces matrices de confusions sont tirées de [13].

B.1 Perception des stimuli auditifs

Les matrices de confusions auditives pour chaque niveau de bruit sont présentées dans les tables B.1, B.2 et B.3.

	/a/	/e/	/i/	/ø/	/y/	/o/	/u/
/a/	208						
/e/		210		3			
/i/			210				
/ø/	1			207			1
/y/					210		1
/o/						209	1
/u/	1					1	207

	/a/	/e/	/i/	/ø/	/y/	/o/	/u/
/a/	210						
/e/		210		15			
/i/			204		1		
/ø/				195	3	5	2
/y/			6		206		2
/o/						202	4
/u/						3	202

TAB. B.1 – Matrices de confusions auditives pour RSB = 12 dB et 6 dB.

B.2 Perception des stimuli visuels

Pour le test de perception visuelle, la matrice de confusions est présentée table B.4.

B.3 Perception des stimuli audiovisuels

Les matrices de confusions audiovisuelles pour chaque niveau de bruit sont présentées dans les tables B.5, B.6 et B.7.

	/a/	/e/	/i/	/ø/	/y/	/o/	/u/
/a/	209						
/e/		196	1	55	4	3	
/i/		1	155		33		5
/ø/		13		132		22	10
/y/			51		163		10
/o/				23		184	9
/u/	1		3		10	1	176

	/a/	/e/	/i/	/ø/	/y/	/o/	/u/
/a/	210						1
/e/		99	1	74	4	53	10
/i/		3	109	1	86	1	41
/ø/		83	3	105	6	74	15
/y/		6	75		91		36
/o/		17		26		80	12
/u/		2	20	4	23	2	95

TAB. B.2 – Matrices de confusions auditives pour RSB = 0 dB et -6 dB.

	/a/	/e/	/i/	/ø/	/y/	/o/	/u/
/a/	126	18	23	13	26	10	29
/e/	12	47	18	51	21	59	31
/i/	22	21	80	14	65	13	48
/ø/	31	67	30	72	31	61	24
/y/	7	12	36	19	43	12	40
/o/	4	31	5	33	6	47	15
/u/	8	14	18	8	18	8	23

	/a/	/e/	/i/	/ø/	/y/	/o/	/u/
/a/	30	15	16	22	24	16	16
/e/	48	55	55	51	49	57	54
/i/	38	45	47	43	45	43	42
/ø/	45	45	41	46	47	48	47
/y/	25	21	33	18	22	20	25
/o/	11	15	12	16	8	19	12
/u/	13	14	6	14	15	7	14

TAB. B.3 – Matrices de confusions auditives pour RSB = -12 dB et -18 dB.

	/a/	/e/	/i/	/ø/	/y/	/o/	/u/
/a/	209						
/e/		150	140				
/i/		58	68	2	4	4	2
/ø/	1	2	2	107	82	76	95
/y/				64	100	23	73
/o/				12	3	91	10
/u/				25	21	16	30

TAB. B.4 – Matrice de confusions visuelle.

	/a/	/e/	/i/	/ø/	/y/	/o/	/u/
/a/	210						
/e/		210		8			
/i/			210				
/ø/				202			
/y/					210		2
/o/						210	5
/u/				2	9	1	189

	/a/	/e/	/i/	/ø/	/y/	/o/	/u/
/a/	210						
/e/		210		4			
/i/			210				
/ø/				206			1
/y/					210		5
/o/						210	7
/u/							197

TAB. B.5 – Matrices de confusions audiovisuelles pour RSB =12 dB et 6 dB.

	/a/	/e/	/i/	/ø/	/y/	/o/	/u/
/a/	210						
/e/		207	5	1			
/i/			202		2		1
/ø/		3		193	2	8	4
/y/			3		197		8
/o/				14		201	8
/u/				2	9	1	189

	/a/	/e/	/i/	/ø/	/y/	/o/	/u/
/a/	209						
/e/		203	12	3	1	1	
/i/		6	196		4		
/ø/		1		169	12	89	26
/y/			2	4	165		53
/o/	1			34	1	118	12
/u/					27	2	119

TAB. B.6 – Matrices de confusions audiovisuelles pour RSB =0 dB et -6 dB.

	/a/	/e/	/i/	/ø/	/y/	/o/	/u/
/a/	209		1				
/e/		182	62	1	1	1	
/i/		26	146	1	5	1	3
/ø/	1	1		157	27	116	58
/y/			1	21	144	7	96
/o/		1		20	5	81	10
/u/				10	28	4	43

	/a/	/e/	/i/	/ø/	/y/	/o/	/u/
/a/	207						
/e/	1	130	116	2	1	1	
/i/		79	92				1
/ø/	1	1		86	67	96	75
/y/			1	84	113	26	90
/o/				17	5	77	18
/u/	1		1	21	24	10	26

TAB. B.7 – Matrices de confusions audiovisuelles pour RSB =-12 dB et -18 dB.

Annexe C

Méthode de calcul

Score d'identification correcte corrigé Le score de reconnaissance correcte corrigé par rapport au seuil aléatoire est le premier outil que nous allons utiliser pour analyser les résultats de nos expériences. Ce score nous permet de calculer le pourcentage de réponse correcte pour chaque niveau de bruit. Nous utilisons la formule suivante pour effectuer ce calcul :

$$\text{Score corrigé} = 100\ \frac{|\frac{\text{Items Corrects}}{\text{Total Items}} - \frac{1}{\text{Nb Classes}}|}{1 - \frac{1}{\text{Nb Classes}}}. \tag{C.1}$$

La Figure C.1 présente les scores de reconnaissance correcte corrigés par rapport au seuil aléatoire selon le niveau de bruit.

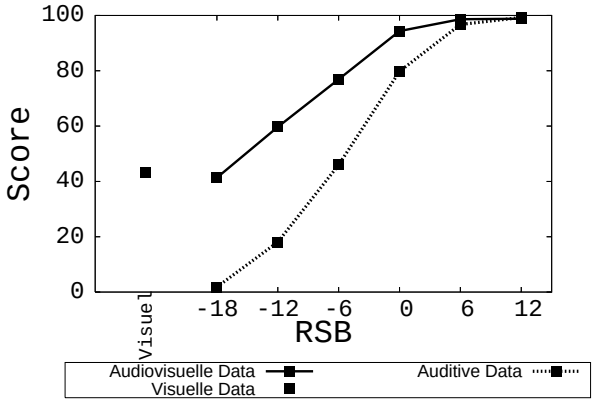

FIG. C.1: Scores d'identification correcte corrigés.

Nous remarquons que la perception audiovisuelle est toujours supérieur ou égale à la perception visuelle seule ou à la perception auditive seule. Nous observons notamment que le score minimum pour les stimuli audiovisuels (-18 dB) est égal au score obtenu avec les stimuli visuels.

Pourcentage d'information transmise sur les traits de hauteur, arrondissement et avant-arrière Pour chaque niveau de bruit, nous avons calculé l'information transmise sur trois traits : arrondissement, avant-arrière et hauteur. Ces traits sont définis dans le Tableau C.1.

$Voyelle$	$avant-arriere$	$arrondissement$	$hauteur$
/i/	1	1	0
/y/	1	0	0
/u/	0	0	0
/e/	1	1	1
/ø/	1	0	1
/o/	0	0	1
/a/	?	?	2

TAB. C.1 – Définition des traits par voyelle (Tableau tiré de [13]).

Il faut noter que pour la voyelle /a/, seul le trait de hauteur est connu. Nous n'avons donc pas tenu compte des stimuli /a/ et des réponses /a/ en ce qui concerne les calculs relatifs aux traits d'arrondissement et avant-arrière.

Ce paragraphe est extrait de [13]. Pour calculer le pourcentage de transmission d'information, nous avons utilisé la méthode de Breeuwer et Plomp. Ce pourcentage est défini par :

$$T = 100\ \frac{H(s,r)}{H(s)}. \tag{C.2}$$

où $H(s,r)$ est l'information transmise des stimuli aux réponses, et $H(s)$ est l'information contenue dans les stimuli. Ces deux quantités sont définies par :

$$H(s,r) = -\sum_i \sum_j p(s_i,r_j) log_2(\frac{p(s_i) * p(r_j)}{p(s_i,r_j)}) \tag{C.3}$$

$$H(s) = -\sum_i p(s_i) log_2(p(s_i)). \tag{C.4}$$

avec
- $p(s_i)$ la probabilité d'apparition du trait s_i dans les stimuli,
- $p(r_j)$ la probabilité d'apparition du trait r_j dans les réponses,
- $p(s_i, r_j)$ la probabilité d'apparition conjointe du trait s_i dans les stimuli et du trait r_j dans les réponses.

La probabilité $p(s_i)$ est connue et est égale à :

$$p(s_i) = \frac{n_i}{n}. \tag{C.5}$$

Et les probabilités $p(r_j)$ et $p(s_i, r_j)$ ne sont pas connues mais peuvent être estimées par :

$$p(r_j) = \frac{n_j}{n} \tag{C.6}$$

et

$$p(s_i, r_j) = \frac{n_{ij}}{n}. \tag{C.7}$$

où n_i est le nombre d'occurrences du trait s_i dans les stimuli, n_j est le nombre d'occurrences du trait r_j dans les réponses, n_{ij} est le nombre d'occurrences du trait s_i dans les stimuli avec les trait r_j dans la réponse, et n est le nombre total des stimuli.

Par exemple, si nous calculons le pourcentage d'information transmise sur le trait d'arrondissement à 0 dB de RSB nous extrayons une matrice 2x2 pour ce trait (Tableau C.2).

	[Stimuli = non arrondi]	[Stimuli = arrondi]
[Reponses = non arrondi]	146	24
[Reponses = arrondi]	4	276

TAB. C.2 – Nombre de stimuli et réponses selon le trait d'arrondissement à 0 dB.

Cette matrice et les sommes de ses lignes et colonnes nous donnent les valeurs de n_i, n_j et n_ij. Nous obtenons alors les probabilités suivantes :

$$p(s_1) = 0.33;\ p(s_2) = 0.66;\ p(r_1) = 0.38;\ p(r_2) = 0.62;\ p(s_1, r_1) = 0.32;$$
$$p(s_1, r_2) = 0.01;\ p(s_2, r_1) = 0.05;\ p(s_1, r_1) = 0.61.$$

Ce qui nous donne un $H(s)$ de 0.92 bit/stimulus et un $H(s, r)$ de 0.63 bit/stimulus et donc un T de 68%.

La Figure C.2 (resp. Figure C.3 et Figure C.4) présente les pourcentages d'information transmise selon le niveau de bruit pour le trait hauteur (resp. avant-arrière et arrondissement).

En ce qui concerne la perception auditive, les résultats montrent que la perception du trait de hauteur est meilleure que celle du trait avant-arrière quand le signal est bruité. Du côté de la perception visuelle, le trait d'arrondissement est clairement identifié. La perception des traits pour les voyelles audiovisuelles est parfaite sans l'ajout de bruit et se dégrade progressivement avec les différents niveaux de bruit. Cependant cette dégradation est limitée : les pourcentages d'information transmise par trait restent toujours égaux ou supérieurs aux scores correspondants en perception visuelle pure. Nous remarquons que la perception audiovisuelle est soumise à un effet de plancher par la perception audiovisuelle.

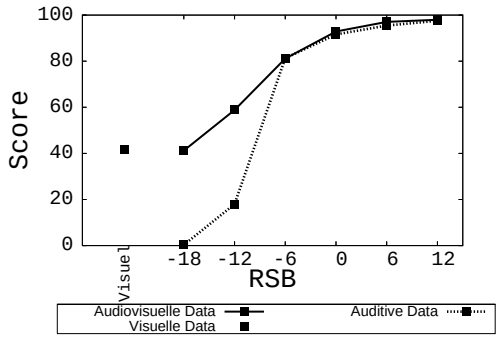

FIG. C.2: Pourcentage d'information transmise selon le trait hauteur.

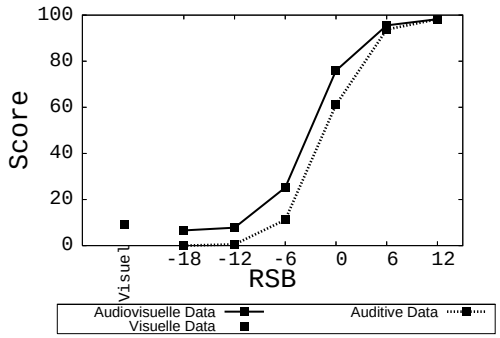

FIG. C.3: Pourcentage d'information transmise selon le trait avant-arrière.

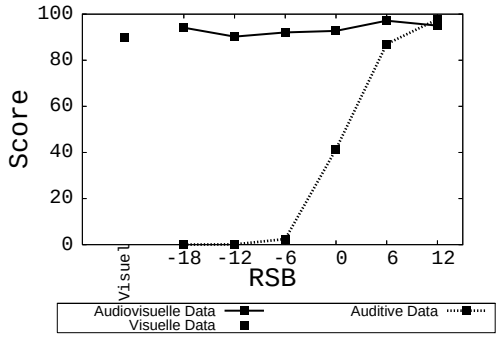

FIG. C.4: Pourcentage d'information transmise selon le trait arrondissement.

Annexe D

Probabilité d'identification des traits d'arrondissement, de hauteur et avant-arrière

D.1 Modèle M_0

Les Figures D.1, D.2, D.3 présentent le pourcentage d'information transmise par trait (arrondissement, hauteur, avant-arrière) comparé à celui des réponses des sujets.

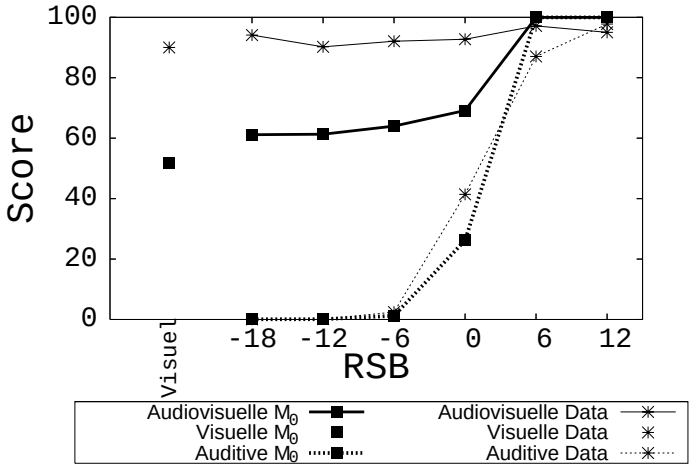

FIG. D.1: Probabilité d'identification du trait d'arrondissement pour le modèle M_0.

Trait d'arrondissement Nous observons que la probabilité d'identification du trait d'arrondissement pour le visuel seul par le modèle M_0 est très inférieur au score réalisé par les sujets. Ce score s'explique par le fait que les voyelles sont ordonnées dans

notre espace : dans l'agencement des voyelles, nous retrouvons bien les deux catégories : les voyelles arrondies (/e/ et /i/) et les non arrondies (/ø/, /o/, /y/ et /u/). Cependant, l'ordre que nous avons choisi rend les voyelles /i/ et /ø/ voisines alors qu'elle n'ont pas le même arrondissement.

De plus, étant donné que lorsque le bruit est très élevé (RSB = -18 dB) la perception audiovisuelle ne dépend que de la modalité visuelle, il est normal de retrouver une faible probabilité d'identification du trait d'arrondissement par le modèle M_0.

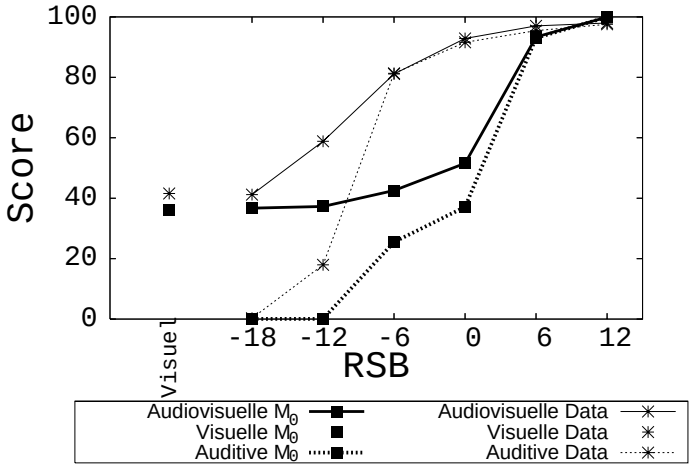

FIG. D.2: Probabilité d'identification du trait de hauteur pour le modèle M_0.

Trait de hauteur La probabilité d'identification du trait de hauteur en visuel est très similaire à celle des sujets. Cependant, pour l'auditif les probabilités sont très inférieures. L'ordre des voyelles que nous utilisons ne correspond pas à un «classement» selon le trait de hauteur : des voyelles telles que /o/ et /y/ n'ont pas le même trait de hauteur alors qu'elles sont voisines. Etant donné que la probabilité d'identification auditive est faible, il est normal de retrouver une probabilité d'identification faible pour la perception audiovisuelle.

Trait avant-arrière Pour le trait avant-arrière, nous observons que les résultats sont similaires à ceux des sujets lorsque le bruit est très faible ou très élevé. Dans le cas de bruits très élevés ($RSB = -18\ dB$ ou $-12\ dB$), la probabilité d'identification de ce trait est nulle que ce soit pour le modèle M_0 ou pour les sujets. Dans le cas de bruits très faibles ($RSB = 6\ dB$ ou $12\ dB$), la probabilité d'identification est quasiment de 100% du fait que le score d'identification correcte est aussi de 100%. Pour un niveau de bruit moyen, nous observons que les probabilités d'identification sont inférieures à celles des sujets ce qui est probablement dû aussi à l'agencement des voyelles.

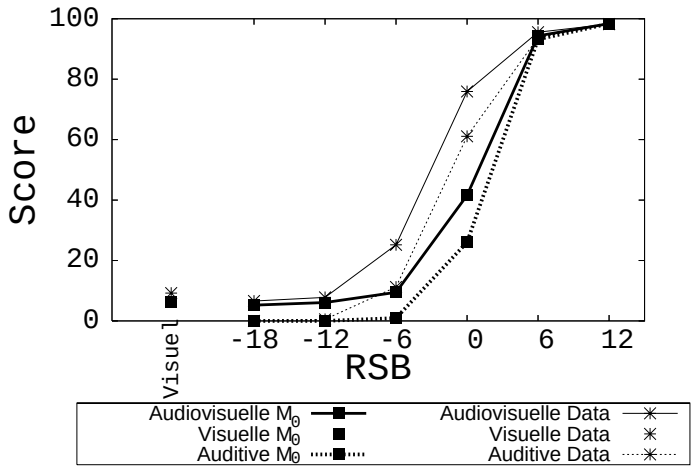

FIG. D.3: Probabilité d'identification du trait avant-arrière pour le modèle M_0.

D.2 Modèle M_1

Les Figures D.4, D.5 et D.6 présentent le pourcentage d'information transmise par trait (arrondissement, hauteur, avant-arrière) comparé à celui des données.

Trait d'arrondissement Nous observons que les probabilités d'identification du trait arrondissement du modèle M_1 sont très bonnes comparées à celles des sujets. Ceci peut s'expliquer par le fait que l'espace des voyelles est discret et sans relation entre les voyelles. La forme paramétrique utilisée est la loi de succession de Laplace (voir Paragraphe 3.2) qui ne fait que reproduire les données. Par conséquent, c'est normal d'obtenir les mêmes scores.

Trait de hauteur Comme précédemment, la probabilité d'identification du trait de hauteur est similaire à celle des sujets.

Trait avant-arrière Nous observons que la probabilité d'identification du trait avant-arrière est aussi similaire à celle des sujets.

D.3 Modèle M_2

Les Figures D.7, D.8 et D.9 présentent le pourcentage d'information transmise par trait (arrondissement, hauteur, avant-arrière) comparé à celui des sujets.

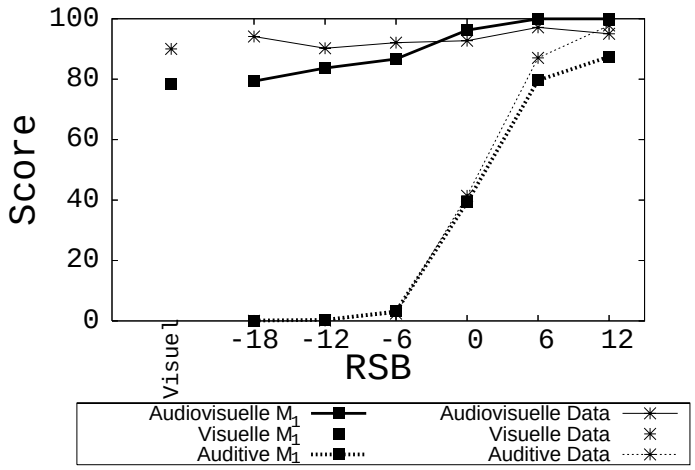

FIG. D.4: Probabilité d'identification du trait d'arrondissement pour le modèle M_1.

Trait d'arrondissement Nous observons que la probabilité d'identification du trait arrondissement est similaire à celle des sujets. Par conséquent, nous pouvons en déduire que l'agencement des voyelles dans notre espace respecte totalement ce trait.

Trait de hauteur De manière générale, la probabilité d'identification du trait de hauteur est similaire à celle des sujets. Seule exception pour RSB = -6 dB où la probabilité est cette fois supérieure aux données des sujets.

Trait avant-arrière Nous observons que la probabilité d'identification du trait avant-arrière est supérieur à celle des sujets.

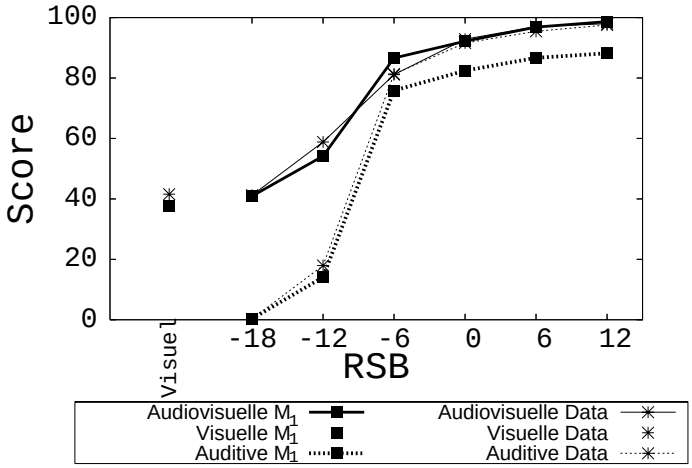

FIG. D.5: Probabilité d'identification du trait de hauteur pour le modèle M_1.

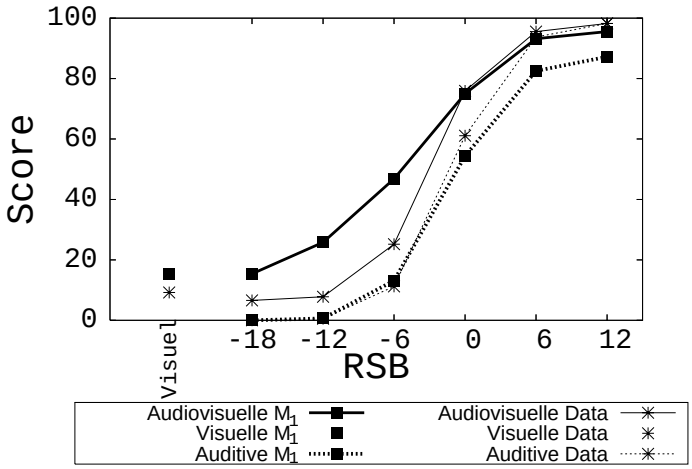

FIG. D.6: Probabilité d'identification du trait avant-arrière pour le modèle M_1.

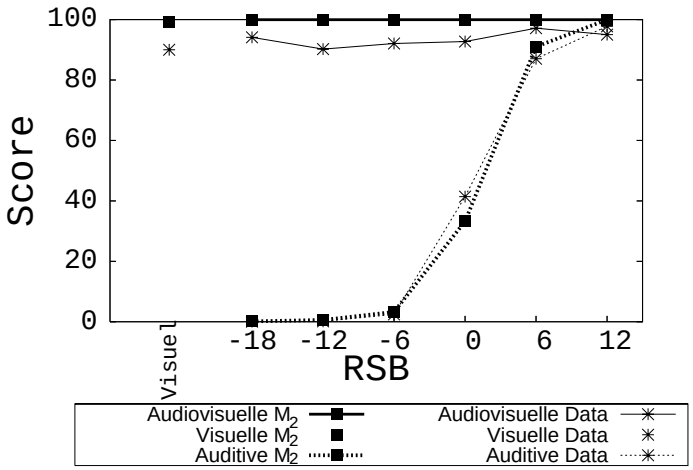

FIG. D.7: Probabilité d'identification du trait d'arrondissement pour le modèle M_2.

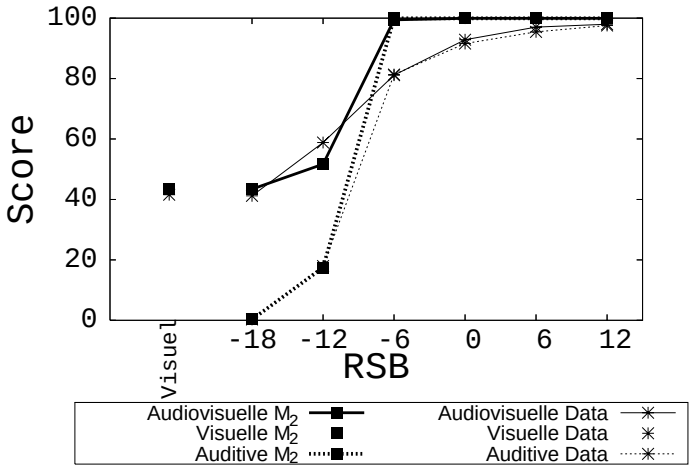

FIG. D.8: Probabilité d'identification du trait de hauteur pour le modèle M_2.

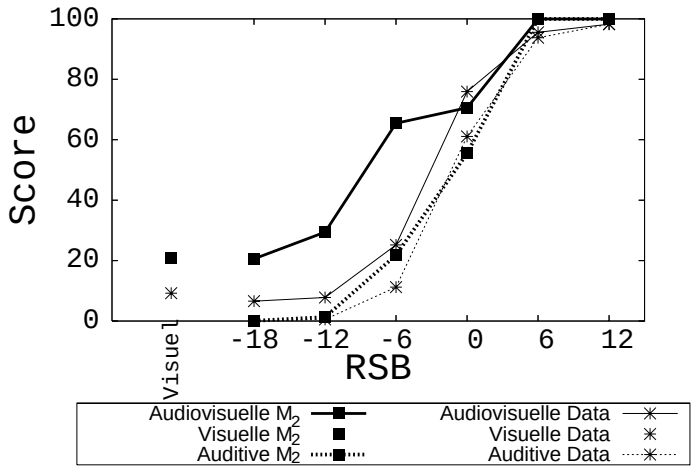

FIG. D.9: Probabilité d'identification du trait avant-arrière pour le modèle M_2.

Annexe E

Structure perceptive

Les Figures E.1, E.2, E.3 et E.4 présentent la structure perceptive pour différents niveaux de bruits.

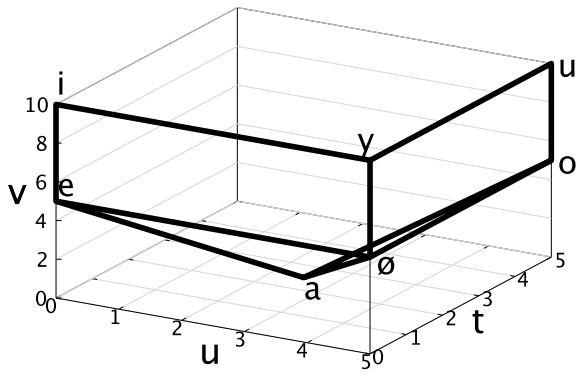

FIG. E.1: Structure perceptive pour RSB = 12 dB et 6 dB.

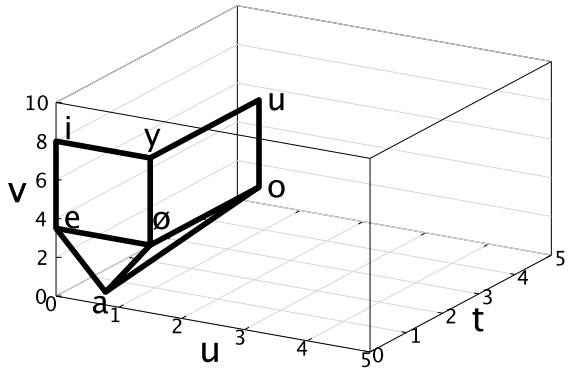

FIG. E.2: Structure perceptive pour RSB = 0 dB.

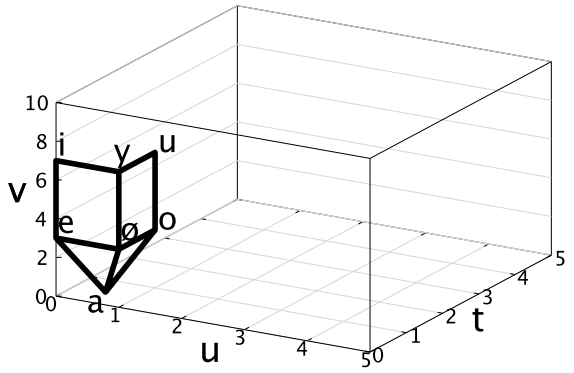

FIG. E.3: Structure perceptive pour RSB = -6 dB.

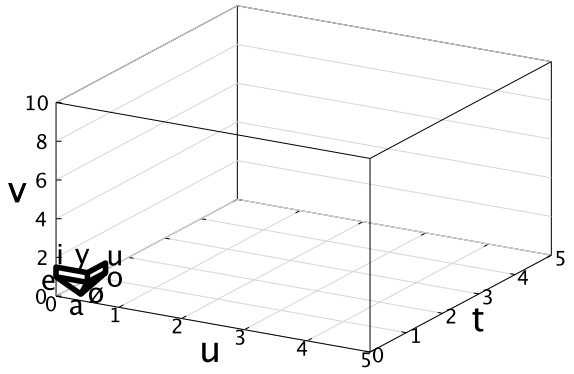

FIG. E.4: Structure perceptive pour RSB = -12 dB et -18 dB.

Bibliographie

[1] D. Alais and D. Burr. The ventriloquist effect results from near-optimal bimodal integration. 2004.

[2] M. Banks. Neuroscience : what you see and hear is what you get. 2004.

[3] P. Battaglia, R. Jacobs, and R. Aslin. Bayesian integration of visual and auditory signals for spatial locallization. 2002.

[4] J. Diard. *La carte bayésienne un modèle probabiliste hiérarchique pour la navigation en robotique mobile*. PhD thesis, Institut National Polytechnique de Grenoble, 2003.

[5] K. Drewing and M. Ernst. Integration of force and position cues for shape perception through active touch. 2005.

[6] M. Ernst and M. Banks. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, 2002.

[7] M. Ernst and H. Bulthoff. Merging the senses into a robust percept. 2004.

[8] S. Gepshtein and M. S. Banks. Viewing geometry determines how vision and haptics combine in size perception. 2003.

[9] D. Kersten, P. Mamassian, and A. Yuille. Object perception as bayesian inference. 2004.

[10] D. Knill, D. Kersten, and P. Mamassian. Implications of a bayesian formulation of visual information for processing for psychphysics.

[11] O. Lebeltel. *Programmation Bayésienne des Robots*. PhD thesis, Institut National Polytechnique de Grenoble, 1999.

[12] L. Lisker and M. Rossi. Auditory and visual cueing of the rounded feature of vowels. *Language and speech*, 1992.

[13] J. Robert-Ribes. *Modèles d'intégration audiovisuelle de signaux linguistiques : de la perception humaine à la reconnaissance automatique des voyelles*. PhD thesis, Institut National Polytechnique de Grenoble, 1995.

[14] J.-L. Schwartz. Une théorie de la perception pour le contrôle de l'action. *Percevoir : Monde et langage*.

[15] E. Simonin. Carte bayésienne et apprentissage. Technical report, Master 2 Mathématiques, Informatique, 2004.

[16] P. Viviani and N. Stucchi. Motor-perceptual interactions. 1990.