Topologies en perception multimodale neuro-inspirée : apprentissage, fusion, prise de décision

Simon Forest

THESE de DOCTORAT DE
L'UNIVERSITE CLAUDE BERNARD LYON 1

Ecole Doctorale N° 512
InfoMaths

Discipline : Informatique

Soutenue publiquement le 16/09/2022, par :
Simon Forest

Topologies en perception multimodale neuro-inspirée : apprentissage, fusion, prise de décision

Devant le jury composé de :

Mme BICHO, Estela	PR	Univ. de Minho	Rapporteure
M. SHEYNIKHOVICH, Denis	HDR	Univ. Sorbonne	Rapporteur
Mme JEAN-DAUBIAS, Stéphanie	PR	UCBL	Présidente
Mme ROHDE, Marieke	PhD	VDI/VDE-IT	Examinatrice
M. TRIESCH, Jochen	PR	Univ. Goethe Frankfurt	Examinateur
Mme HASSAS, Salima	PR	UCBL	Directrice de thèse
M. LEFORT, Mathieu	MCF	UCBL	Co-encadrant
M. QUINTON, Jean-Charles	MCF	Univ. Grenoble Alpes	Co-encadrant
M. LAFLAQUIÈRE, Alban	PhD	Sony AI	Invité

Topologies in neuro-inspired multimodal perception: learning, merging, decision-making

Simon Forest

Laboratoire d'InfoRmatique en Image et Systèmes d'information,
Univ. Claude Bernard Lyon 1

Laboratoire Jean Kuntzmann,
Univ. Grenoble Alpes

Under the supervision of:
Salima Hassas, LIRIS, UCBL
Mathieu Lefort, LIRIS, UCBL
Jean-Charles Quinton, LJK, UGA

To my father.

Acknowledgments

First, I would like to thank the members of jury for their thorough review of my work. They provided me with detailed and valuable feedback, and contributed to a very interesting discussion time during my defense.

I would also like to thank all the colleagues who shared their experience during team meetings, PhD seminars, and project meetings; either from Liris, LJK or other labs. I would like to thank in particular the other members of the Amplifier project, for listening to me as I presented work outside their field, and in return helping me discover the vast disciplines of psychophysics and neuroscience.

Even outside work, I have a special thought for the "non-perma" of LJK and SycoSMA with the midday games and afterworks; Guillaume, Yassine and the rest of office 143; Léo, Juba and the rest of office 12.041; and even the non-PhD-students who still managed (or at least tried) to comprehend what I was doing: Florent, Seb, Sylvain, and many others from ADR, DDT, and SEL.

Finally, I would like to wholeheartedly thank my supervisors, Mathieu Lefort and Jean-Charles Quinton, for their precious advice and support. They kept giving constructive feedback almost weekly for more than four years, showing patience in the hardest steps of the PhD, and offering a fruitful collaboration, as much in writing articles as in solving escape rooms.

Funding

The first three years of this thesis were funded by the "Pack Ambition Recherche" of region Auvergne-Rhône-Alpes as part of the project Amplifier. The fourth year was funded by Polytech Lyon with an ATER.

The internship that preceded the thesis was funded by LabEx Persyval-Lab (ANR-11-LABX-0025-01) in the French program "Investissement d'avenir", the ANR GaG, and CNRS MITI project APF^2.

The internship of Jose Villamar, who contributed to the work presented in chapter 4, was funded by Pôle Grenoble Cognition (FR CNRS 3381 – SFR UGA / Grenoble INP).

Abstract

This thesis focuses on multimodal merging on simulated topologies for use in an active perception context. As an example, humans receive dense information from multiple sensors and use various mechanisms to select and attend to only the relevant signals, for example by moving their eyes towards a target to see it better. Because of irregularities in sensory topologies (cf. fovea), actions can enhance perception, while extracting and merging data also helps choosing the best course of action. Similar needs are faced by artificial systems, e.g. social robots, albeit with their own set of physical constraints. This thesis proposes computational models for use in AI, taking inspiration from neuroscience case studies involving the superior colliculus, a subcortical structure involved in generating saccadic eye movements towards visual, auditory or multisensory stimuli.

When selecting information from multiple signals in a dynamic and multimodal setting, one needs a way to compute robust and reliable decisions. Decision-making in general has been tackled in either psychophysics or robotics using many different algorithms. One contribution of this thesis is to review and compare these algorithms, underlining their spatio-temporal properties, including feature merging, selective attention, etc. Among these models, dynamic neural fields (DNF) display some very interesting characteristics, including selective attention and data fusion depending on stimulus distance and precision. In another contribution, this thesis then makes use of DNF as a signal filtering and merging tool applied to multimodal fusion. This thesis shows how it can apply to model realistically occurences of the ventriloquist effect, a psychophysical effect of audio/visual stimulus localization capture. Then, in order to further study the role of topologies on these cognitive tasks, a final contribution shows that DNF retain their properties in irregular learned topological maps. In this experience, topologies are learnt via growing neural gas in order to extract intrinsic dimensions of the sensory space, but new perspectives, with deeper models, are suggested for application in active perception and embodied cognition.

Résumé

Cette thèse porte sur la fusion multimodale sur des topologies apprises dans un contexte de perception active. À titre d'exemple, les humains reçoivent des informations denses provenant de multiples capteurs et utilisent divers mécanismes pour sélectionner et se concentrer sur les signaux pertinents uniquement, par exemple en déplaçant le regard vers un objet pour mieux le voir. En raison des irrégularités dans les topologies sensorielles (cf. fovéa), les actions peuvent améliorer la perception, tandis que l'extraction et la fusion de données aident également à choisir le meilleur plan d'action. Les systèmes artificiels, par exemple les robots sociaux, font face à des besoins similaires, malgré un ensemble de contraintes physiques qui leur est propre. Cette thèse propose des modèles computationnels pour l'IA, en s'inspirant d'études de cas en neurosciences impliquant le colliculus supérieur, une structure sous-corticale impliquée dans la génération de saccades vers des stimuli visuels, auditifs ou multisensoriels.

Pour sélectionner des informations à partir de signaux multiples dans un contexte dynamique et multimodal, il faut trouver un moyen de générer des décisions fiables et robustes. La prise de décision en général a été abordée à la fois en psychophysique et en robotique, via de nombreux algorithmes différents. Une des contributions de cette thèse est de passer en revue et comparer ces algorithmes, en soulignant leurs propriétés spatio-temporelles, y compris la fusion, l'attention sélective, la mémoire, etc. Parmi ces modèles, les champs neuronaux dynamiques (DNF) présentent des caractéristiques très intéressantes, notamment l'attention et la fusion de données en fonction de la distance et de la précision des stimuli. Dans une autre contribution, cette thèse utilise ensuite les DNF comme un outil de filtrage et de fusion de signaux appliqué à la fusion multimodale. Cette thèse montre comment il peut s'appliquer pour modéliser de manière réaliste des manifestations de l'effet ventriloque, un effet psychophysique de capture de localisation de stimuli audio/visuels. Puis, afin d'étudier plus en détail le rôle des topologies sur ces tâches cognitives, une dernière contribution montre que les DNF conservent leurs propriétés dans des cartes topologiques irrégulières apprises. Dans cette expérience, les topologies sont apprises via un gaz neuronal croissant afin d'extraire les dimensions intrinsèques de l'espace sensoriel, mais de nouvelles perspectives, avec des modèles plus profonds, sont suggérées pour une application dans le cadre de la perception active et la cognition incarnée.

Contents

List of Figures

List of Tables

Glossary

DNF	dynamic neural field
FFI	feed-forward inhibition
FL	fuzzy logic
GNG	growing neural gas
KF	Kalman filter
MLE	maximum-likelihood estimation
NLCA	nonlinear leaky competing accumulator
SC	superior colliculus
SOM	self-organizing map
WS	weighted sum
WTA	winner-takes-all

Preamble

This thesis was part of the Auvergne-Rhône-Alpes project Amplifier (2017–2022). The main goal of the project is to propose a new view on multimodal fusion under the light of active perception. That includes, on one side, a collaboration from labs of neuroscience and psychology in order to highlight the impact of saccades in the ventriloquist effect, a psychophysical effect of audio or visual capture. On another side, this thesis was funded with the initial objective of proposing new computational models of multimodal fusion compatible with saccades, with the option of fitting them to the new psychophysical data. Another aim, not pursued since, was to apply the models in social robotics.

Started on October 1st, 2018, and preceded by a 6-month master internship, this thesis thus first focused on building a neuro-compatible model of multimodal merging with new considerations for dynamic behaviors, something that classical models used in psychophysics (most notably maximum-likelihood estimation — MLE) could not do. Two design choices were quickly taken. First, to draw inspiration from the superior colliculus (SC), a subcortical region studied for both its multimodal merging and saccade mechanisms. Second, to build on the paradigm of dynamic neural fields (DNF), a population-based model of decision-making in a mesoscopic-scale neural map.

These two design choices were later tied to two additional research focuses. The initial comparison between DNF and MLE was put in perspective in a review encompassing decision-making algorithms across different fields, from neuroscience to robotics. And meanwhile, topological constraints found in SC simulation were challenged using a more versatile manifold learning algorithm. Topologies, and their theoretical impact on multimodal fusion and active perception, then started to impose as the guiding thread of this thesis.

This manuscript is structured as follows: after a brief overview of active perception and multimodal fusion (chapter 2 and 3 respectively), we present our three contributions: our review and unifying framework on decision-making algorithms (chapter 4), a DNF model of multimodal merging (chapter 5), and the additional manifold learning for fusion (chapter 6). We follow with a re-contextualization of the contributions (chapter 7) and perspectives for future work (chapter 8).

All this is preceded by a general introduction on general considerations about multimodal merging (chapter 1). This chapter is not meant to substitute to the state of the art and positioning that follows, but rather to provide potential future non-expert readers with an affordable entrance to the topics of this thesis. It is deliberately written with a lighter tone, a tone that will resurface as a transition between main contributions. Expert readers should not worry over the simplified take of these sections, and may skip them if they deem so. They are typeset in a distinct style to make them more recognizable.

Chapitre 1

Introduction à la multimodalité

Warning: English version page 20!
I translated the first chapter in French to introduce my PhD topic to spectators not fluent in English. I make it available here for future wanderers, but the original version below is recommended.

Il n'y a pas si longtemps, je jouais à un jeu de société avec des amis. Dans ce jeu, chaque joueur devait inventer une histoire dans laquelle il devait placer certains mots-clés imposés. Une personne jouant au détective devait alors deviner les mots-clés et trouver quel joueur avait des mots-clés différents des autres. Une stratégie courante dans ce jeu consiste à ajouter des mots « bizarres » non demandés afin d'embrouiller le détective. Pendant mon tour, j'ai inséré le mot « multimodal » dans mon histoire. On m'a immédiatement démasqué : « Multimodal » est trop farfelu pour venir du jeu. Cela m'a surpris. En quoi « multimodal » est-il farfelu ? La multimodalité est partout. Ou, comme nous allons le voir, partout est multimodal.

En guise de démonstration, laissez-moi vous proposer une expérience multisensorielle fictive. Imaginez la scène suivante (figure 1.1) : Un groupe d'amis joue à un jeu de rôle sur table, qui se déroule à une époque médiévale. La tension est à son comble, car il s'agit de la dernière session de la campagne et les joueurs vont bientôt affronter le grand méchant boss final. La nuit avance, animée de dés lancés frénétiquement, d'estomacs souffrant d'une accumulation déraisonnable de malbouffe, et d'yeux plissés à force d'essayer de discerner quoi que ce soit dans la faible lumière de l'unique bougie qui éclaire la pièce — « c'est pour l'immersion », dit Alice, la meneuse de jeu. Il est maintenant temps pour Bob, l'un des joueurs, de lancer un dé à 20 faces pour un test de perception. Voici ce qui se passe :

1. Le dé rebondit plusieurs fois sur la table puis en tombe.
2. On entend le dé rouler au sol.
3. Bob passe sa tête sous la table. Le sol est mal éclairé, il peut vaguement distinguer deux petits objets à l'endroit approximatif où il a entendu le dé pour la dernière fois.
4. Bob tend la main vers l'objet le plus proche. C'est rond et rugueux au toucher.
5. Bob approche l'objet de sa tête. Il ne peut toujours pas le voir clairement mais ça sent la fraise.
6. « Tiens Alice, j'ai trouvé le bonbon que t'as fait tomber tout à l'heure », dit Bob.

7. Bob mange le bonbon sans vergogne. Un goût de sucre et de fraise emplit sa bouche.
8. Bob tend la main vers le deuxième objet. C'est lisse, avec des faces plates et des bords arrondis. Il reconnaît le dé au toucher mais ne peut pas lire sa valeur dans l'obscurité.
9. Bob amène soigneusement le dé au-dessus de la table sans le retourner, et l'approche de la bougie. C'est un 1 ! Échec critique.
10. Alice glousse de façon inquiétante.

FIG. 1.1: Un aperçu de la scène fictive

Dans cette histoire, du point de vue de Bob, ses cinq sens sont stimulés : la vision (1, 3, 5, 9), l'audition (1–2, 6, 10), le toucher (4–5, 8–9), l'odorat (5+) et le goût (7+). Voyons comment ces sens sont stimulés tout au long de l'histoire (figure 1.2).

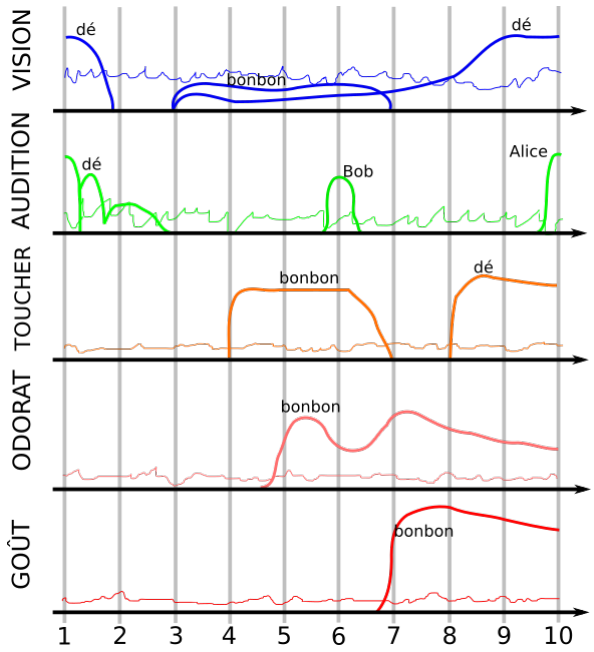


FIG. 1.2: Entrées regroupées par sens. Dans chaque ligne, des traits superposés représentent différents objets. Un bruit de fond est inclus. Les lignes verticales correspondent à des séquences temporelles arbitraires associées au début des événements décrits dans le texte.

Cela fait déjà beaucoup d'informations, nous pourrions tenter de les compresser en regroupant les stimuli correspondants. Chaque sens peut être stimulé par différents objets d'attention : le dé, le bonbon, Alice, et un grand nombre d'autres stimuli qui ne sont pas pertinents pour les décisions et que nous traiterons comme un bruit de fond (objets visibles sur la table, mouvements des autres joueurs, etc.) Il est à noter que certaines stimulations simultanées peuvent être attribuées à la même source (le même dé est à la fois vu et entendu à l'étape 1), et d'autres non (l'objet qui est goûté à l'étape 8 n'est pas le même que celui qui est touché au même moment). Évidemment, la congruence temporelle n'est pas le seul facteur permettant de déterminer si certains stimuli peuvent être combinés[1]. Regroupons maintenant les stimuli en fonction de la source à partir de laquelle ils sont projetés.

1. Dans ce scénario, Bob peut séparer l'objet dans sa main du bonbon dans sa bouche parce que sa main est éloignée de sa bouche. Cette information lui est fournie par les contingences apprises (c'est-à-dire les corrélations entre la perception qu'il a de sa main et son état moteur) et sa proprioception (la perception de la position et des mouvements de son propre corps), parfois considérée comme le sixième sens de l'homme (bien que nous ne l'utilisions pas plus dans ce scénario).

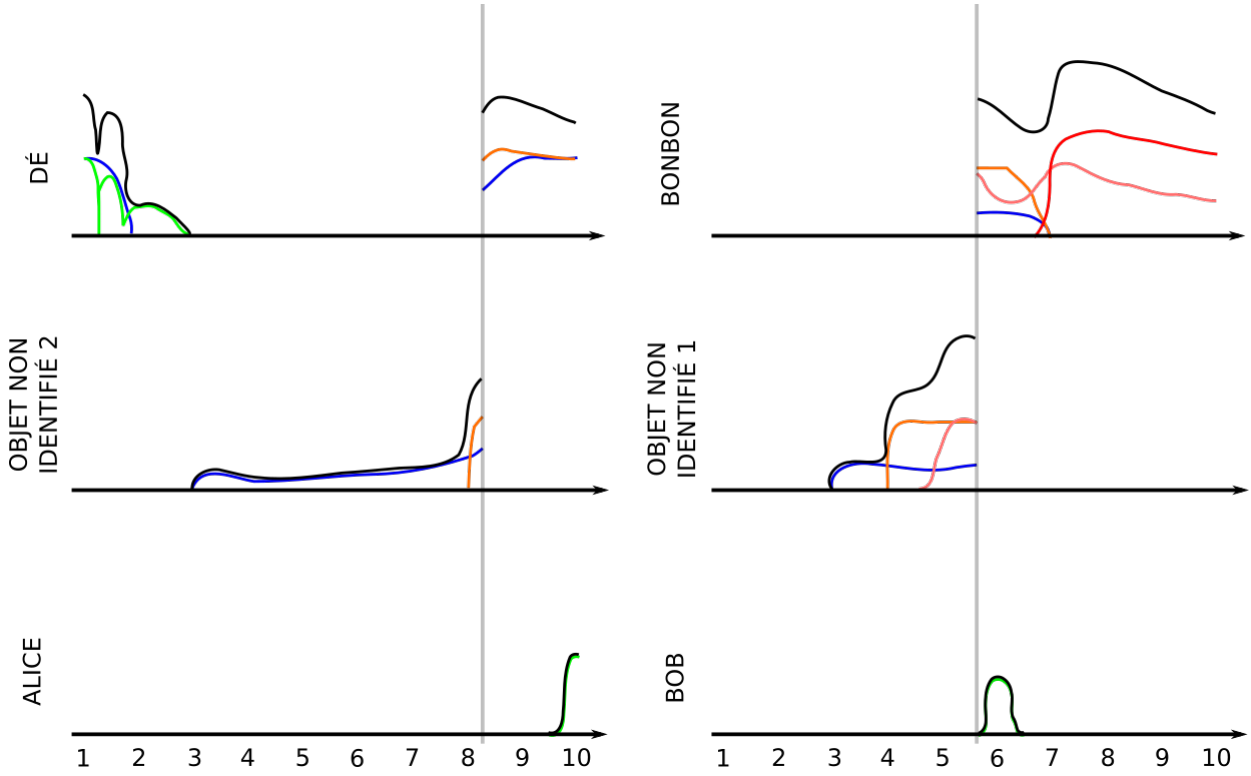


FIG. 1.3: Entrées regroupées par objet. Les sens sont colorés comme dans la figure 1.2. Dans chaque ligne, un trait noir montre la somme de tous les autres.

Avec la figure 1.3, l'enchaînement des événements est plus compréhensible. Après avoir perdu le dé, Bob a le choix entre deux objets non identifiés, avec des informations égales et insuffisantes sur eux (rangée du milieu, étape 3). Il choisit l'un d'entre eux pour y porter son attention, et fait ce qu'il peut pour obtenir plus d'informations à son sujet : il le touche, le rapproche de son visage pour essayer de le regarder de plus près, et le sent. Après avoir identifié l'objet comme étant sans rapport avec son objectif, puis l'avoir éliminé de la scène, Bob poursuit ses recherches et passe au deuxième objet non identifié.

Il faut remarquer que pendant que Bob est en train d'identifier le bonbon, il ne porte pas son attention au hasard sur l'autre objet ou sur un autre stimulus en bruit de fond. Et plus tard, il se concentre sur l'identification, puis la lecture du dé, alors que son sens du goût est complètement accaparé par le bonbon ! C'est un mécanisme qui joue un rôle crucial dans l'intégration multisensorielle : l'attention. C'est comme si certains stimuli pouvaient être renforcés à volonté, et d'autres inhibés s'ils ne sont pas pertinents pour la tâche à accomplir. Et cela tombe bien, compte tenu de la quantité d'informations auxquelles nous sommes exposés à chaque seconde. Pour simplifier[2], nous représenterons cette attention comme un niveau d'intérêt fixe, défini à l'avance pour chaque type d'objet (figure 1.4). Cela constituera une nouvelle couche d'information à ajouter à l'intensité du stimulus.

2. L'attention peut en fait prendre de nombreuses formes et jouer un rôle dans de nombreux mécanismes différents : se concentrer sur une modalité et ignorer les autres, se concentrer sur un objet et ignorer le reste, faire attention aux stimuli les plus saillants… Il ne s'agit ici que d'un cas particulier d'attention descendante guidée par la tâche en cours.

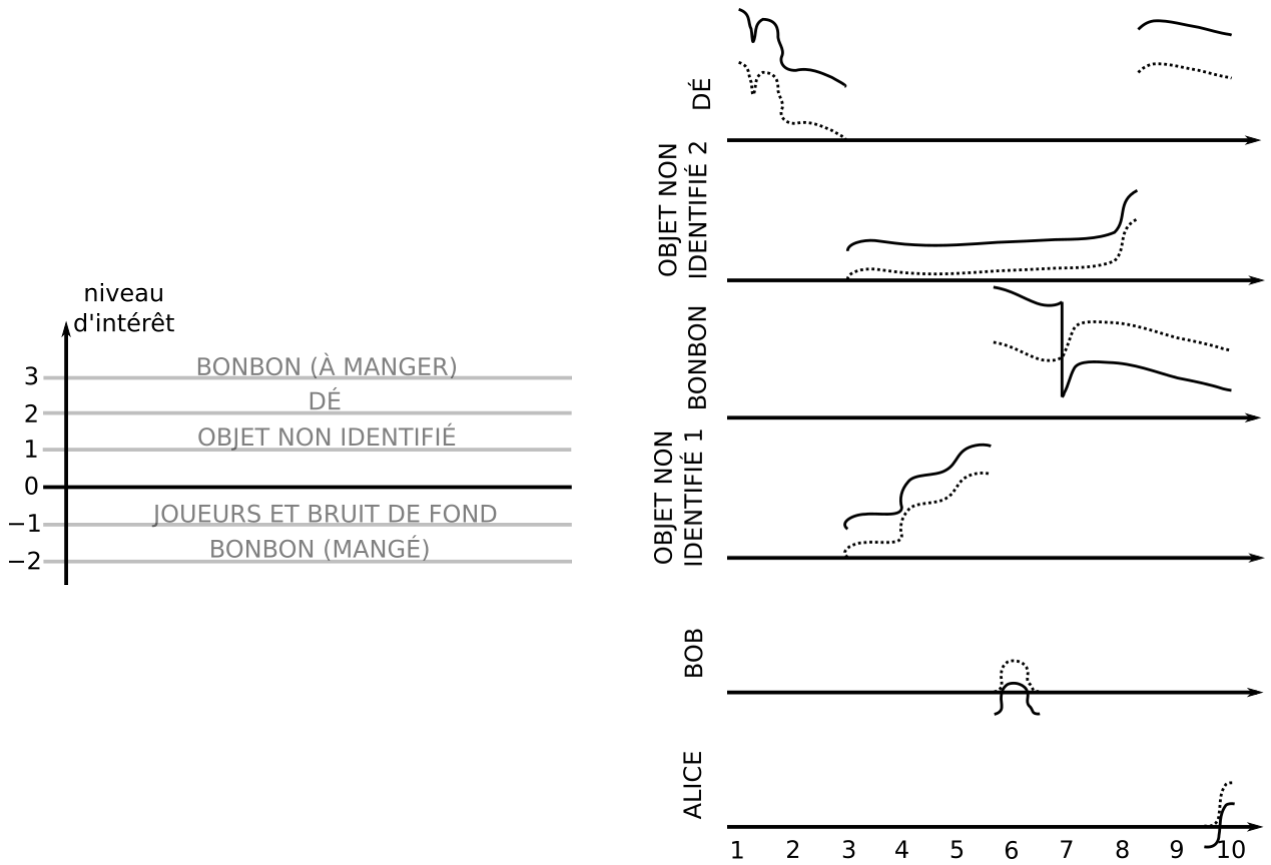


FIG. 1.4: À gauche : niveau d'intérêt accordé à différents objets. À droite : stimulations modulées en ajoutant leur niveau d'intérêt correspondant. Valeur originale en pointillés, mise à jour en plein.

Ce modèle simplifié montre comment les stimuli pertinents peuvent être privilégiés tandis que les stimuli non pertinents sont ignorés. Il s'agit d'un cas d'attention descendante (c'est-à-dire que des informations issues de processus cognitifs de haut niveau, comme notre connaissance de l'importance de chaque type d'objet, guident la perception sensorielle à bas niveau). Notez que même un niveau d'intérêt négatif n'empêche pas un autre stimulus de s'accaparer l'attention. Par exemple, le fait qu'Alice, meneuse de jeu, éclate d'un rire sardonique, devrait être suffisamment intense pour que Bob détourne son attention du dé qui roule.

Multisensoriel et multimodal

Jusqu'à présent, nous avons discuté des cinq sens et de la manière d'accumuler des informations multimodales. Prenons maintenant un peu de recul et expliquons les modalités. Une définition de la « modalité » donnée par le *Cambridge Dictionary* est « une manière particulière de faire ou de ressentir quelque chose ». En première approximation, on peut supposer que chaque sens porte une modalité différente (figure 1.5), par exemple, la vision, l'audition, etc. donnent toutes une information telle que la position de l'objet.

Fig. 1.5: Alors que Bob tend la main pour saisir un objet, deux modalités le guident : une localisation approximative qu'il a estimée à partir du son du roulement du dé (entourée en vert), et la vision qu'il a de deux objets sur le sol (entourés en bleu).

C'est bien le sens de la multimodalité que nous privilégierons pour la majeure partie de cette thèse : devoir combiner une information donnée par un sens et une information donnée par un autre (dans notre cas, la localisation visuelle et la localisation auditive). Mais avant de nous en tenir à ce cadre, il faut noter que la multimodalité peut représenter bien plus, car un seul sens peut fournir plusieurs types d'informations. Tout objet que vous voyez a une position, une orientation, une taille, une forme et une couleur. Tout son que vous entendez a une position, une hauteur et une tonalité. Chacune de ces sources de données peut être traitée comme une modalité différente, même au sein d'un même sens (où un sens est tout ce qui peut être échantillonné par un type de capteurs : vision, audition, toucher…).

Fig. 1.6: Ces objets ont différentes couleurs, luminances et formes. Avez-vous trouvé l'intrus ?

Dans la figure 1.6, même si vous n'avez jamais vu de dé à 20 ou 4 faces auparavant, vous pouvez facilement les classer dans la même catégorie que le dé à 6 faces, et dans une catégorie différente de celle du bonbon. Pour différencier le dé du bonbon, les modalités pertinentes seraient la forme (le dé a des bords droits, pas le bonbon), le contraste (les chiffres sur le dé) ou la texture (le dé est lisse, le bonbon rugueux). Notez que la texture,

par exemple, peut être perçue à travers deux modalités différentes : tactile et visuelle. En général, la manipulation dynamique par nos doigts nous permet de mieux évaluer la texture d'un objet que nos yeux. Mais cela signifie-t-il que nous devrions toujours faire confiance à nos doigts plutôt qu'à nos yeux pour identifier une texture ?

En fait, la multimodalité n'est pas aussi simple que de choisir la modalité la plus adaptée à une tâche donnée. En regardant la figure 1.6, vous pouvez instantanément différencier un bonbon d'un dé. Mais en regardant la figure 1.5, c'est beaucoup plus difficile. Et pourtant, il s'agit exactement de la même tâche. Dans la vie réelle, la perception est rarement unimodale. Consciemment ou non, nous nous résolvons souvent à combiner des modalités, même si certaines d'entre elles ont un impact négligeable. De cette observation émergent de nombreux points de questionnement : Quand une modalité n'est-elle pas suffisante pour prendre une décision ? Quand faut-il combiner des modalités ? Comment savoir si elles sont complémentaires (par exemple, la forme et la texture) ? Redondantes (par exemple, la texture perçue à la fois par la vision et le toucher) ? Est-ce qu'elles parlent du même objet ? Quel importance faut-il alors accorder à chaque modalité ? Et que faire si elles se contredisent ?

Prise de décision

Dans l'histoire qui nous a servi de fil conducteur, Bob a combiné diverses informations pour arriver à plusieurs décisions. Illustrons deux de ces processus :

1. « C'est un bonbon qui doit être mangé. »
2. « C'est le dé dont je dois lire la valeur. »

La figure 1.7 montre certaines des informations qui peuvent être prises en compte afin de prendre une de ces décisions perceptives. Il s'agit d'une représentation symbolique, il est improbable que le cerveau humain encode réellement des éléments d'information unimodaux d'une manière si distinctive.

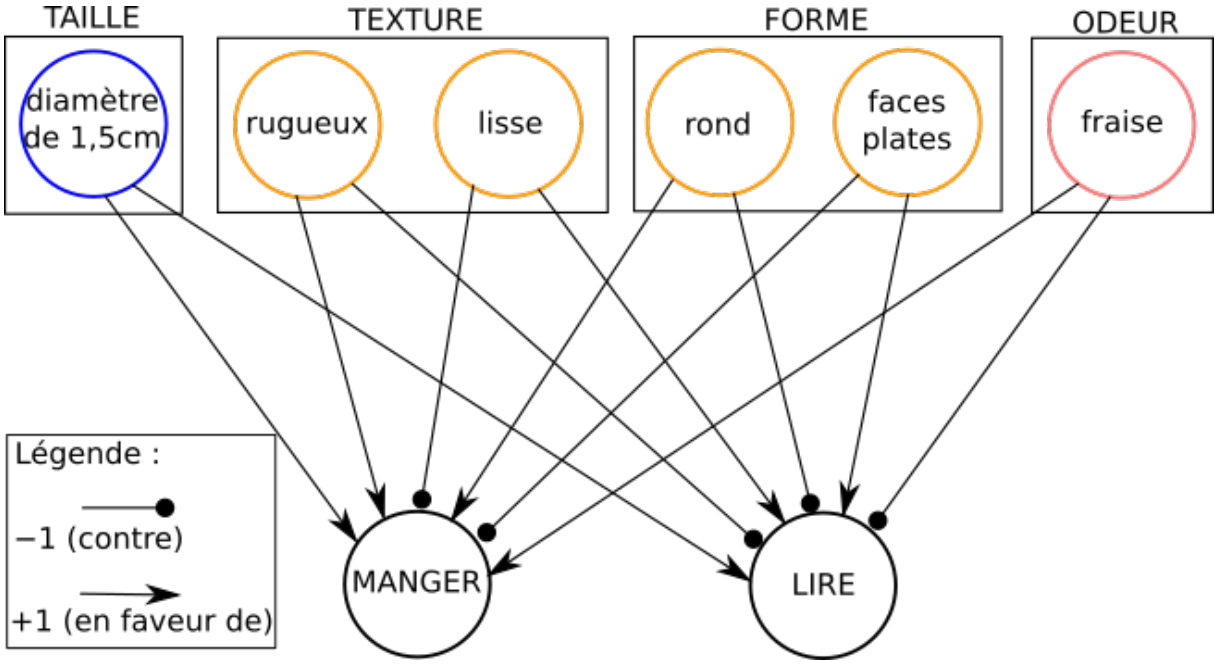


Fig. 1.7: Le flux de décision de Bob. Par exemple, une odeur de fraise ajoutera un indice en faveur de la consommation de l'objet et contre une tentative de le lire.

Voici comment lire le graphique : Sur la ligne du haut sont représentées plusieurs conditions. Lorsque l'une d'entre elles est remplie, elle ajoute un élément de preuve pour

ou contre la décision sur la ligne du bas (fixée arbitrairement à ±1). Dans notre scénario, à l'étape 5, il y a +4 de preuves en faveur de la décision « manger » et −2 en faveur de « lire » (figure 1.8, à gauche). À l'étape 8, c'est −1 et +3 respectivement (figure 1.8, à droite). Gardez à l'esprit qu'il ne s'agit que d'un modèle statique d'accumulation. En réalité, les éléments de preuve devraient s'accumuler au fil du temps et les décisions devraient être mises à jour à chaque étape.

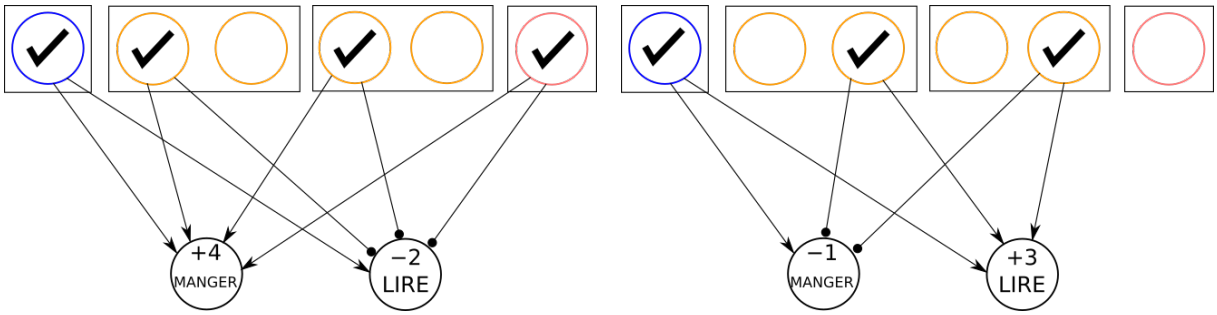


FIG. 1.8: Résultat de la décision pour le premier objet saisi (à gauche) et le second (à droite)

La décision est prise en prenant le choix « le plus positif », mais ce n'est pas tout. En effet, Bob ne mange pas tous les objets de 1,5 cm de texture rugueuse et de forme ronde. Il existe un seuil au-delà duquel les éléments de preuve sont suffisants pour activer la décision de manger l'objet. Pour simplifier, nous pouvons définir un seuil binaire (figure 1.9), bien que les vrais modèles computationnels préfèrent utiliser des fonctions d'activation continues comme des sigmoïdes.

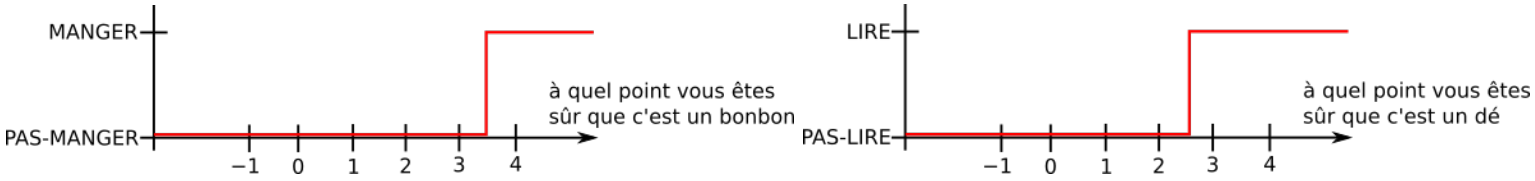


FIG. 1.9: Fonction d'activation d'une décision

Finalement, un processus de décision peut être résumé ainsi :

1. Évaluer la preuve provenant de toutes les modalités. Les éléments de preuve peuvent être modulés par la tâche (par exemple, vous pouvez ignorer le goût dans votre bouche quand vous cherchez un dé), l'action (le toucher que vous ressentez lorsque vous saisissez un objet est plus pertinent que, disons, le toucher de la table sur votre tête en vous contorsionnant dessous), la temporalité (les sons qui se produisent après que le dé a cessé de rouler ne sont pas si pertinents), etc.

2. Les multiplier par un poids : positif s'il favorise une décision, négatif s'il va à son encontre.

3. Faire la somme de toutes les preuves pour chaque décision possible.

4. Tester leur « activation » via un seuil.

5. S'il reste plusieurs choix, prendre la décision avec le niveau de preuve le plus élevé.

Maintenant que nous disposons modèle minimaliste de prise de décision, examinons comment la multimodalité mélange tout cela. Jusqu'à présent, les pondérations données

à chaque condition étaient soit -1 soit $+1$. Cela signifierait que toutes les modalités ont le même effet sur les décisions. Mais en réalité, toutes les modalités n'ont pas la même fiabilité. Peut-être que Bob n'en était pas à son premier bonbon et qu'il en avait des résidus sur les doigts, donc le fait que quelque chose dans ses mains sente la fraise n'était pas si important que ça, et n'aurait dû avoir qu'une pondération de $0,5$. Bob voit peut-être un petit peu la couleur des objets, mais étant donnée l'obscurité de l'environnement, cela ne vaut que $0,01$. Le poids accordé aux modalités peut être modifié de nombreuses façons, en fonction du contexte, de la précision intrinsèque, de l'activité menée, etc. La fonction d'activation peut également être modifiée en une fonction non linéaire quelconque au lieu d'un simple seuil 0/1. Mais comment pondérer toutes les modalités reste une question ouverte. De nombreux facteurs peuvent entrer en jeu : la fiabilité, la congruence (les informations confirmées par d'autres modalités méritent plus de poids), l'attention (les informations provenant de modalités considérées comme non pertinentes méritent moins de poids)… La prise de décision peut s'appuyer sur des mécanismes continus et fluides, et à cet égard, il faut noter que les décisions, et les informations sur lesquelles elles s'appuient, ne prennent pas toujours des valeurs discrètes (dans l'espace ou dans le temps), et peuvent parfois s'échantillonner sur un spectre spatial et temporel continu.

Décisions spatiales

Notre histoire fournit un exemple de décision prise dans un espace continu, dans son étape 4 (figure 1.10). Alors qu'il regarde sous la table, notre protagoniste doit choisir dans quelle direction il va tendre la main.

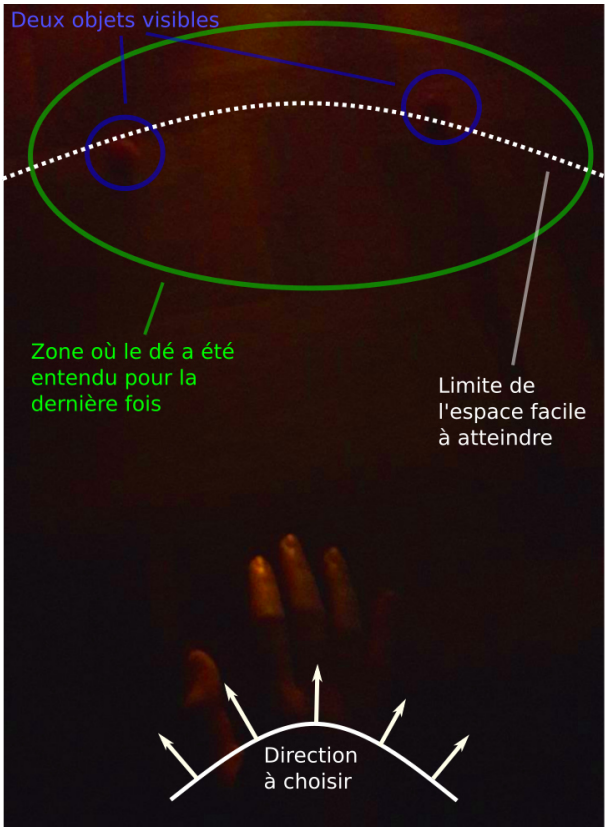


Fig. 1.10: Toutes les informations concernant la position des objets peuvent être regroupées dans un espace topologique commun.

Supposons que nous regardions la scène depuis la main de Bob au niveau du sol. Là encore, il s'agit d'une simplification très arrangeante, car la vision, l'audition et les mouvements de la main se produisent dans des cadres de référence très différents (respectivement

centré sur l'œil, centré sur la tête, et lié au corps via la proprioception). Nous partons du principe qu'à un moment donné, le cerveau projette toutes ces informations sur une topographie commune[3]. La figure 1.11 montre comment nous pourrions représenter la preuve en faveur de la bonne position de la cible, en fonction de la direction de la main.

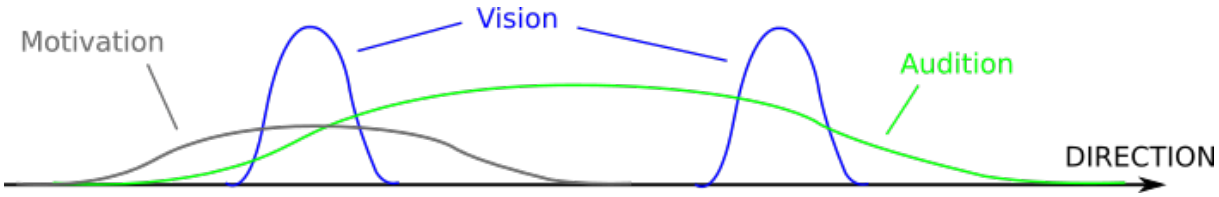


FIG. 1.11: La direction d'intérêt peut être influencée par de multiples modalités, ainsi que par des facteurs cognitifs tels que l'effort (l'objet situé à gauche demande moins d'effort pour être atteint) et, encore une fois, l'attention.

Cette notion d'espace, une forme d'encodage topologique, n'est pas limitée à la position physique. Par exemple, les textures ne sont pas limitées à un état « rugueux » et un état « lisse », on peut trouver toute une gamme de textures entre ces extrêmes. Et toutes les couleurs peuvent être placées sur un espace de représentation inspiré du spectre des couleurs (figure 1.12).

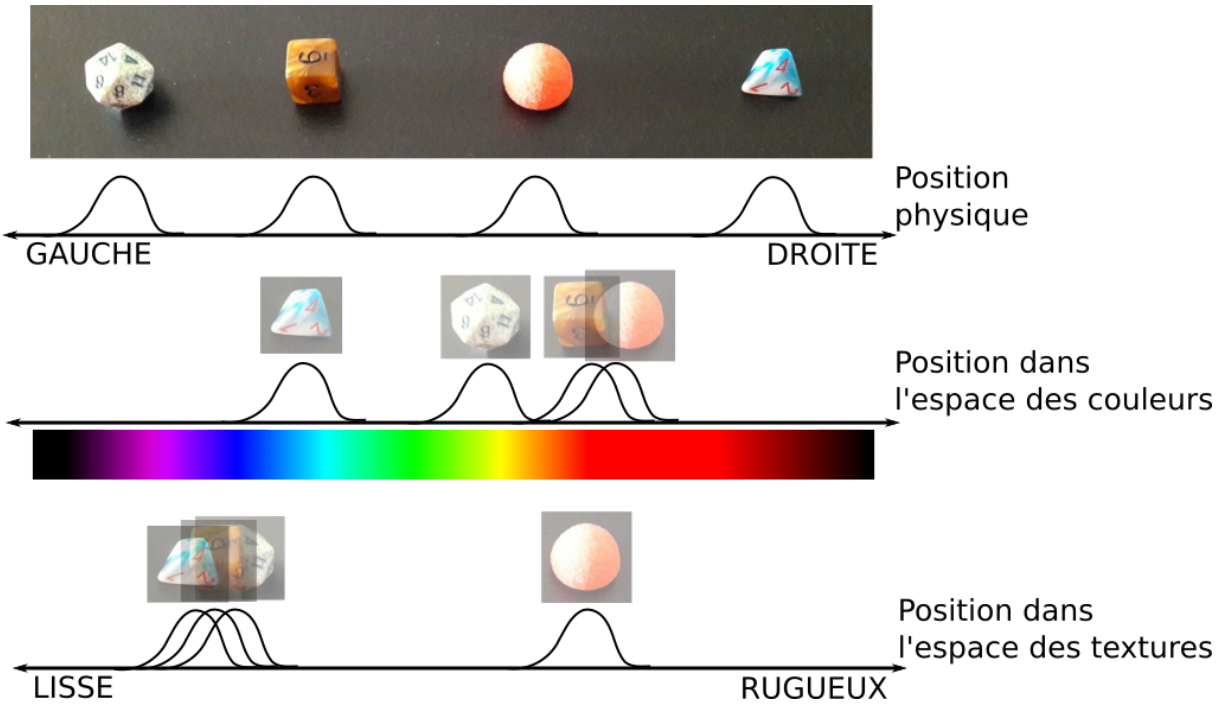


FIG. 1.12: Trois manières possibles de discriminer les objets

Ce n'est pas tout ! Toutes ces dimensions peuvent être combinées[4] pour construire un nouvel espace avec toutes les modalités pertinentes pour une décision (figure 1.13).

3. De telles projections se produisent réellement dans des cartes neuronales, mais probablement pas aussi nettement.

4. Nous nous arrêterons ici, avant d'ouvrir une nouvelle boîte de Pandore connue sous le nom de *« binding problem »*, c'est-à-dire comment des percepts différents sont fusionnés en une seule expérience (ici, comment la localisation et la couleur/texture sont regroupés en points).

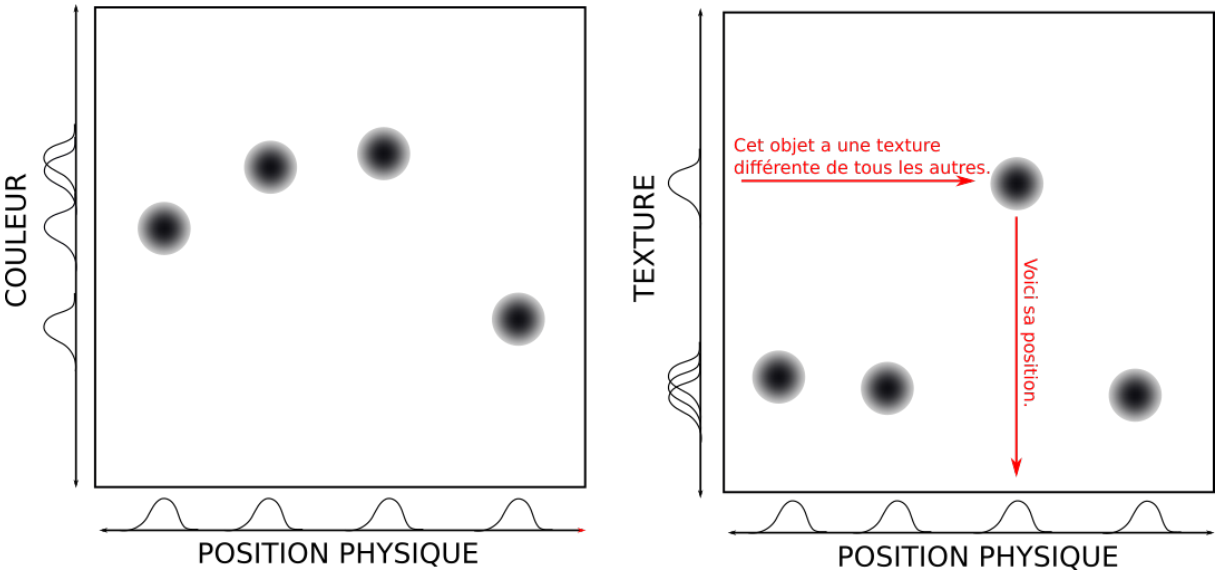


FIG. 1.13: Détection d'objets à l'aide de deux modalités. Chaque point représente un objet dans un espace bidimensionnel couleur-position (à gauche) ou texture-position (à droite).

Maintenant, on peut se poser d'autres questions : comment définir les espaces de la décision ? Quelles sont ses métriques ? La figure 1.13 donne l'impression que la distance entre le rouge et le bleu est la même qu'une distance de 10 cm : est-ce que cela a un sens ? Je suis enclin à répondre que oui, il peut exister une topologie locale dans laquelle cela a parfois un sens. Le problème est que pour chaque configuration, pour chaque environnement, observateur et tâche… la décision prend place dans une topologie différente.

La couleur l'illustre bien. L'espace des couleurs visibles est parfois représenté comme un spectre à deux dimensions (figure 1.14). Mon espace des couleurs observables est probablement différent du vôtre, car je suis daltonien. Le vôtre serait également différent si vous regardiez sous une table dans une pièce éclairée par une bougie. L'environnement, le contexte, et peut-être même vos actions peuvent influencer ce que vous percevez.

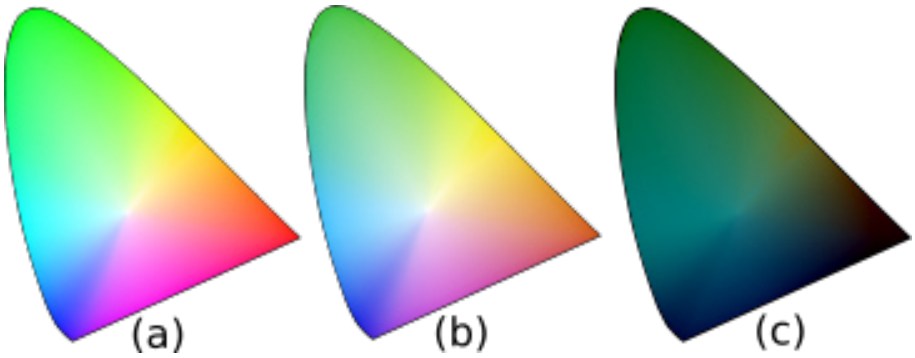


FIG. 1.14: Diagrammes de chromaticité. (a) Couleurs visibles par une personne saine dans de bonnes conditions d'éclairage. (b) Couleurs visibles par l'auteur (daltonien) dans de bonnes conditions d'éclairage. (c) Couleurs visibles par une personne saine sous une table dans une pièce éclairée par une bougie.

Dans de bonnes conditions d'éclairage, Bob n'aurait eu aucun mal à distinguer le dé du bonbon. Mais dans cette configuration, son espace visuel était un grand mélange de taches floues et grises. Et c'est la principale raison pour laquelle il a commencé à toucher les objets dans un ordre arbitraire (le plus proche d'abord, pour minimiser l'effort).

En fait, avec cette relation supplémentaire de cause à effet, nous sommes sur le point de finir la boucle que nous avons commencée au début de cette introduction. Une propriété

fondamentale de la multimodalité est que la fiabilité des modalités varie, et nous pouvons y faire quelque chose, et c'est pourquoi nous agissons. Pendant que vous lisez ces lignes, vos yeux bougent et transportent votre regard sur toute la page. Car tout comme Bob décide de toucher les objets qu'il ne peut pas voir correctement, vous décidez instinctivement de placer les mots plus ou moins là où vous pouvez les lire le mieux, généralement alignés avec le centre de votre rétine. Cette action vous permet d'avoir une meilleure perception, et vous aidera pour la prochaine décision (figure 1.15).

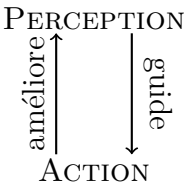


Fig. 1.15: La boucle de la perception active

On serait tenté de dire, à tort, que la fusion multimodale, la prise de décision et la perception active sont trois problèmes différents. Ma thèse est qu'il y a une chose qui les relie tous et qui maintient la boucle en mouvement : l'espace. Un espace sans fin, en perpétuel changement, et aux multiples facettes.

Chapter 1

An introduction to multimodality

Not so long ago, I was playing a board game with some friends. The principle of the game was that each player had to make up a story in which they had to place certain keywords imposed by the game. Someone playing detective then had to guess the keywords and find out which player had different keywords than the others. A common strategy in this game is to add unrequired "strange" words in order to confuse the detective. When it was my turn to play, I inserted the word "multimodal" in my story. I was immediately called out: "'Multimodal' is too far-fetched for it to come from the game." That caught me by surprise. How is "multimodal" far-fetched? Multimodality is everywhere. Or, as we will come to see, everywhere is multimodal.

As a demonstration, let me suggest an imaginary multisensory experience. Picture the following scene (figure 1.1): A group of friends are playing a table-top roleplay game set in a medieval era. Tension is high, as it is the last session in the campaign and the players will soon have to fight the big bad evil final boss. The night is far advanced, dice are thrown frantically, stomachs hurt from an unreasonable accumulation of junk food, and eyes are squinted from trying to discern anything in the dim light of the sole candle that illuminates the room — "it's for immersion", says Alice, the game master. It is now time for Bob, one of the players, to throw a 20-sided die for a perception check. Here is what happens:

1. The die bounces on the table multiple time then falls off it.
2. The die is heard rolling on the floor.
3. Bob puts his head under the table. The floor is badly lit, he can vaguely distinguish two small objects in the approximate area where he last heard the die.
4. Bob reaches to the closest object. It is round and rough to the touch.
5. Bob approaches the object to his head. He still can't see it clearly but it smells like strawberry.
6. "Hey Alice, I found the candy you dropped earlier", Bob says.
7. Bob eats the candy shamelessly. A taste of sugar and strawberry fills his mouth.
8. Bob reaches out to the second object. It is smooth with flat faces and rounded edges. He recognizes the die by touch but can not read its value in the darkness.
9. Bob carefully brings the die above the table without flipping it, and approaches it to the candle. The value is 1! That is a critical failure.
10. Alice chuckles creepily.

Figure 1.1: A look at the fictional scene

In this story, from Bob's point of view, all five senses are stimulated: vision (1, 3, 5, 9), audition (1–2, 6, 10), haptics (4–5, 8–9), smell (5+) and taste (7+). Let us picture how these senses are stimulated along the story (figure 1.2).

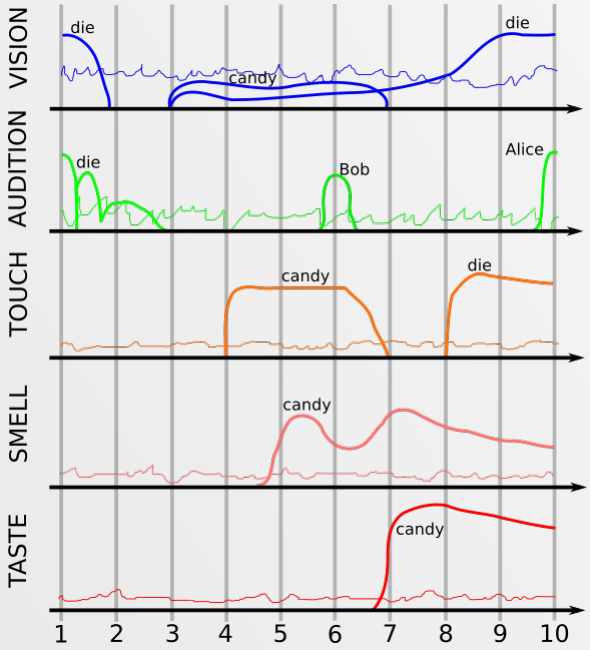


Figure 1.2: Inputs grouped by sense. In a row, superposed lines represent different objects. Some background noise is included. Vertical lines correspond to arbitrary time frames associated with beginning of events described in main text.

That is already a lot of information, so we might be inclined to compress it by regrouping matching stimuli. Each sense can be stimulated by different objects of interest: the die, the candy, Alice, and a big set of other stimuli that are not relevant to decisions and that we will treat as background noise (visible objects on the table, movements of other players, etc.). It is interesting to note that some simultaneous stimulations can be attributed to the same source (the same die is both seen and heard at step 1), and some cannot (the object that is tasted during step 8 is not the same as the one that is touched at the same time). Obviously, time congruence is not the only factor to determine whether some stimuli can be combined[1]. So, what if we regrouped the stimuli according to the source from which they are projected?

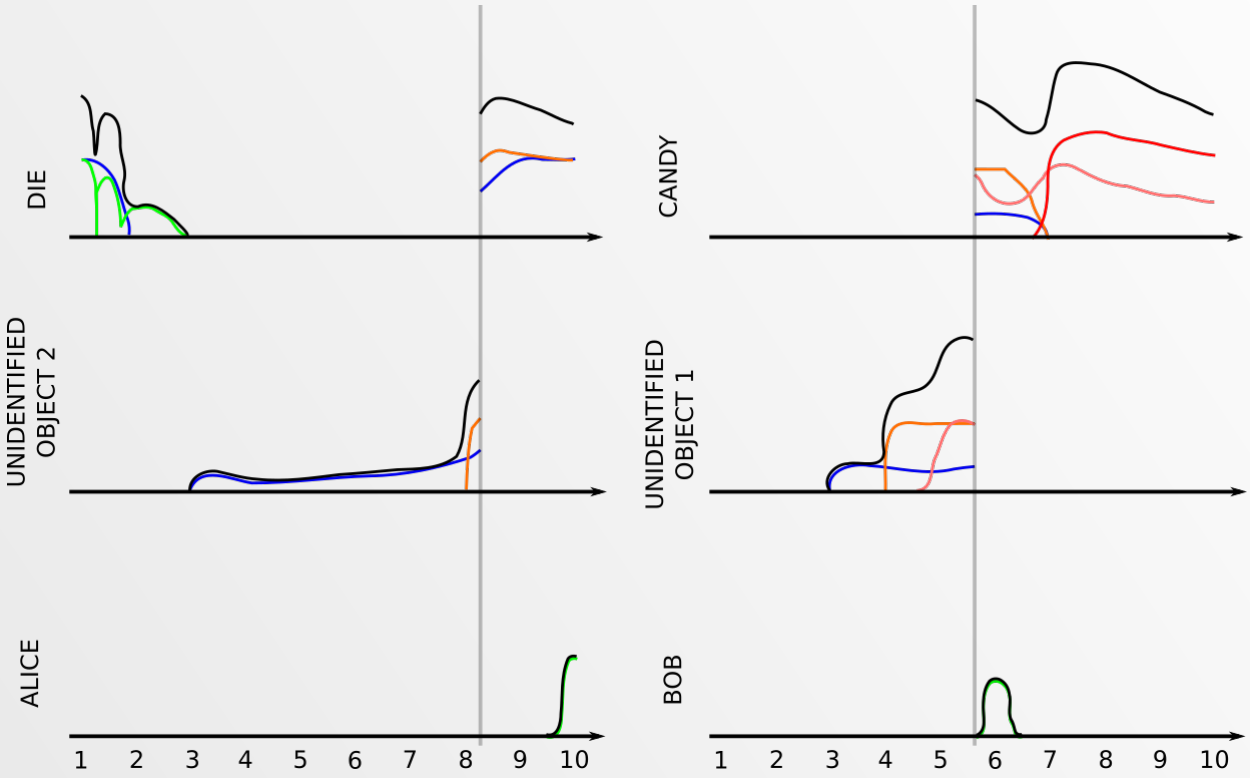


Figure 1.3: Inputs grouped by object. Senses are colored as in figure 1.2. For each row, a black line shows the sum of all others.

Using figure 1.3, the sequence of events is clearer. After the die is lost, Bob has access to two unidentified objects with equal and unsufficient information (middle row, step 3). He picks one of them so that he can focus on it, and does what he can to gain more information about it: he touches it, then brings it closer to his face, so that he can try to watch it more closely, and he smells it. Once the object is identified as something irrelavant to his task, and dealt with, Bob pursues his objective and moves on to the second unidentified object.

[1] In this scenario, Bob can separate the object in his hand from the candy in his mouth because his hand is away from his mouth. That information is given to him by learnt contingencies (i.e. correlations between his perception of his hand and its motor state) and proprioception (the perception of one's own body position and movements), which is occasionally treated as man's sixth sense (although we will not use it further in this scenario).

It is striking that while Bob is in the process of identifying the candy, he does not randomly put his focus towards the other object or some other stimulus in the background. And after that, he focuses on identifying, then reading, the die, as one of his senses is completely overwhelmed by the candy! This is a mechanism that plays a crucial part in multisensory integration: attention. It is as if some stimuli could be enhanced at will, and others inhibited if they were not relevant to the task at hand. And that is convenient, considering the amount of information one is exposed to, every single second. For simplicity[2], we will represent this attention as a static level of interest defined in advance for each type of object (figure 1.4). That will constitute a new level of information to be added to the stimulus intensity.

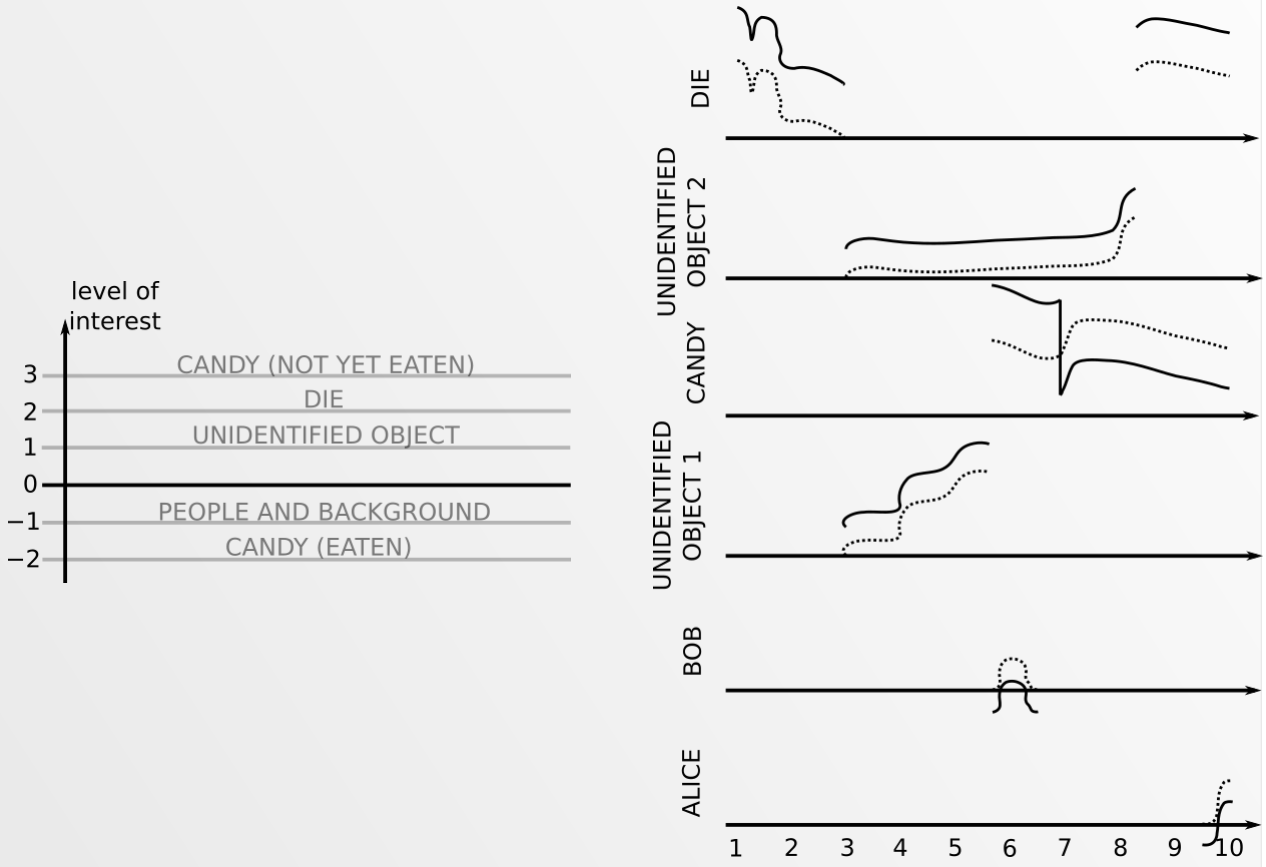


Figure 1.4: Left: level of interest given to different objects. Right: stimulations modulated by adding their corresponding level of interest. Original value in dashed line, updated in full.

This simplistic model shows how relevant stimuli can be preferred while irrelevant stimuli are ignored. It is a case of top-down attention (i.e. information from higher cognitive processes, like the knowledge of the importance of each type of object, guides the sensory perception). Note that even negative interest levels do not prevent another stimulus from taking over the focus. For example, Alice, the game master, bursting out in sardonic laughter should be intense enough to take Bob's attention away from the rolling die.

[2] Attention can actually take many forms and plays a part in many different mechanisms: focusing on one modality and ignoring the rest, focusing on one object and ignoring the rest, attending to salient stimuli... This here is only a particular case of task-driven top-down attention.

Multisensory and multimodal

Until now, we have been discussing about the five senses and how to accumulate multimodal information. Now let us take a step back and explain modalities. A definition of "modality" given by Cambridge Dictionary is "a particular way of doing or experiencing something". In a first approximation, it can be assumed that each sense carries a different modality (figure 1.5), e.g. vision, audition, etc. all give one information such as object position.

Figure 1.5: As Bob extends his hand to grab an object, two modalities guide him: an approximate location he estimated from the sound of the die rolling (circled in green), and the vision he has of two objects on the floor (circled in blue).

That is indeed the meaning of multimodality that we will priviledge for most of this thesis: having to combine one information given by a sense and one information given by another (in our case, visual localization and auditory localization). But before we stick to that frame, we should note that multimodality can represent so much more, for a single sense can provide multiple types of information. Any object you see has a position, an orientation, a size, shape, and color. Any sound you hear has a position, pitch and tone. Each of these datasources can be treated as a different modality, even within the same sense (where a sense is everything that can be sampled through one type of sensors: vision, audition, touch…).

Figure 1.6: These objects all have different colors, luminance and shapes. Can you find the odd one out?

In figure 1.6, even if you have never seen 20-sided or 4-sided dice before, you can easily put them in the same category as the 6-sided die, and a different category than the candy. To differentiate the dice from the candy, relevant modalities would be sharpness (the dice have straight edges, the candy does not), contrast (digits on the dice) or texture (the dice are smooth, the candy has a rough texture). Note that texture for example can be perceived through two different modalities: one tactile, and one visual. In general, dynamic manipulation through our fingers gives us a better appreciation of an object's texture than our eyes. But does that mean that we should always trust our fingers over our eyes when identifying a texture?

In fact, multimodality is not as simple as picking the one modality that is best fit for a given task. When you look at figure 1.6, you can instantly differentiate candy from die. But when you look at figure 1.5, that is much more difficult. And yet it is exactly the same task. In real life, perception is rarely unimodal. Consciously or not, we often resort to combining modalities, even if some of them might have a negligible weight. Many insights emerge from this observation: When is a modality not enough to make a decision? When should modalities be combined? Are they complementary (e.g. shape and texture)? Redundant (e.g. texture perceived by both vision and touch)? Do they even provide information about the same object? Then what weight should be given to each modality? And what if they do contradict one another?

Decision-making

In the story we used as a guiding thread, Bob combined diverse pieces of information to come up with several decisions. Let us illustrate two of this processes:

1. "This is a candy that should be eaten."

2. "This is the die I need to read the value of."

Figure 1.7 shows some of the information that can be taken into account in order to take one of these perceptual decisions. This is a symbolic representation, it is unlikely that the human brain actually encodes unimodal pieces of information that distinctively.

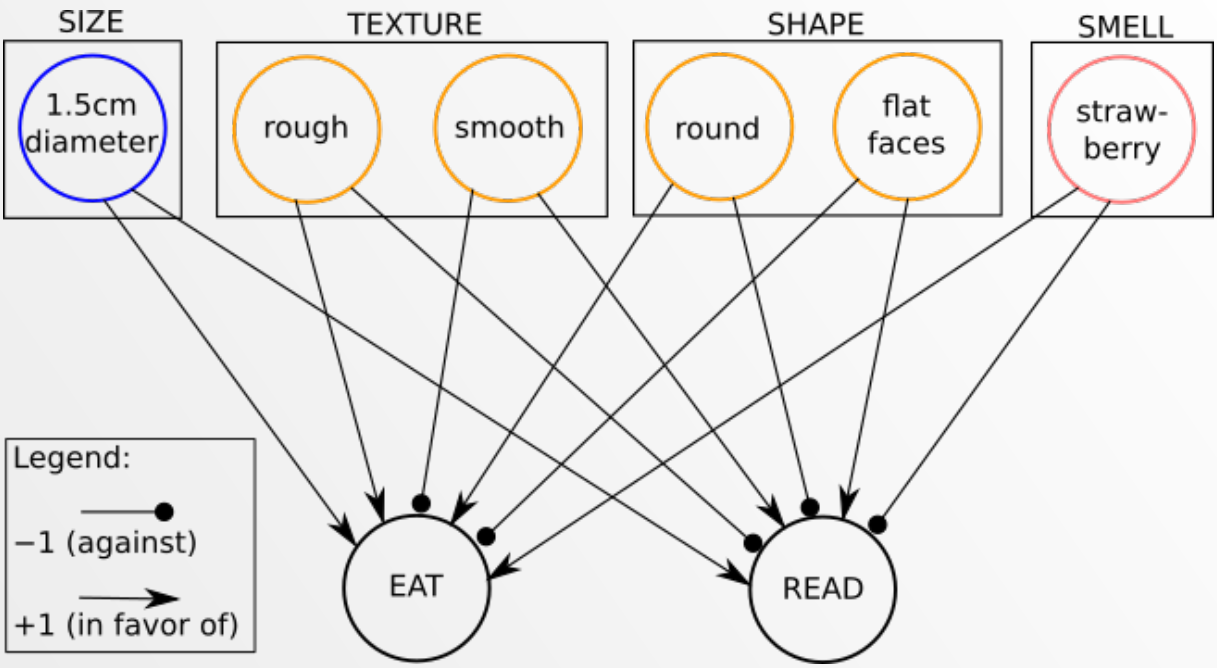


Figure 1.7: Bob's decision flow. For example, a strawberry smell will add evidence in favor of eating the object and against attempting to reading it.

Here is how to read the graph: On the top row are depicted several conditions. When one of them is met, it adds evidence for or against the decision on the bottom row (arbitrarily set to ± 1). In our scenario, at step 5, there is a $+4$ evidence towards the "eat" decision and -2 towards "read" (figure 1.8, left). At step 8, it is -1 and $+3$ respectively (figure 1.8, right). Keep in mind that this is only a static model of accumulation. Realistically, evidence should accumulate over time and decisions should be updated step by step.

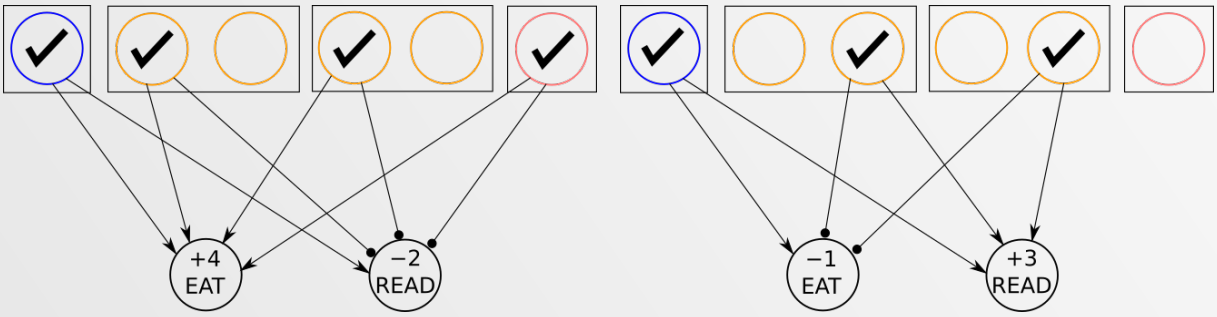


Figure 1.8: Outcome of the decision for the first grasped object (left) and the second (right)

The decision is made by picking the "most positive" choice, but that is not all. Indeed, Bob does not eat every 1.5 cm object with rough texture and round shape. There is a threshold over which evidence is sufficient to activate the decision to eat the object. To simplify, we can set a binary threshold (figure 1.9), though real computational models would rather use continous activation functions such as sigmoids.

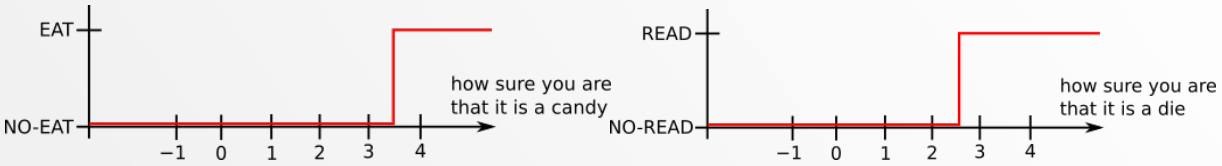


Figure 1.9: Activation function of a decision

Finally, a decision process can be summarized as follows:

1. Evaluate the evidence coming from all modalities. The evidence might be modulated by task (e.g. you can ignore the taste in your mouth when looking for a die), action (the touch you feel when grasping at an object is more relevant than, say, the touch of the table on your head as you crouch below), temporality (sounds that occur after the die has stopped rolling are not that relevant), etc.
2. Multiply them by a weight: positive if it favors a decision, negative if it goes against.
3. Sum all evidence for each possible decision.
4. Test their "activation" with a threshold.
5. If there are multiple choices, pick the decision with the highest evidence.

Now that we have a minimal illustrative process for decision-making, we can consider how multimodality mixes things. Until now, all weights were either -1 or $+1$. This would mean that all modalities have the same effect on decisions. But realistically, not all modalities have the same reliability. Maybe Bob had been eating candy for a while already, and had candy residues on his hands, so the fact something in his hands smells like strawberry was not that important and should only have been ± 0.5. Maybe Bob sees a little of the color of the objects, but given the darkness of the environment it is worth only ± 0.01. The weight given to modalities can be tweaked in many ways, depending on context, intrinsic precision, self activity, etc. The activation function can also be changed into any non-linear function instead of a threshold. But how to weigh all modalities remains an open question. Many factors can come into play: reliability, congruence (information confirmed by other modalities should have more weight), attention (information from modalities considered irrelevant should have less weight), etc. Decision-making can rely on continuous, smooth mechanisms, and to that regard, it is important to note that decisions, and the information they rely on, do not always take discrete values (in space or in time), and can sometimes be sampled from a continuous spatial and temporal spectrum instead.

Spatial decisions

Our story provides an example of a decision made in a continuous space, in its step 4 (figure 1.10). As he looks under the table, our protagonist has to choose in which direction he should extend his hand.

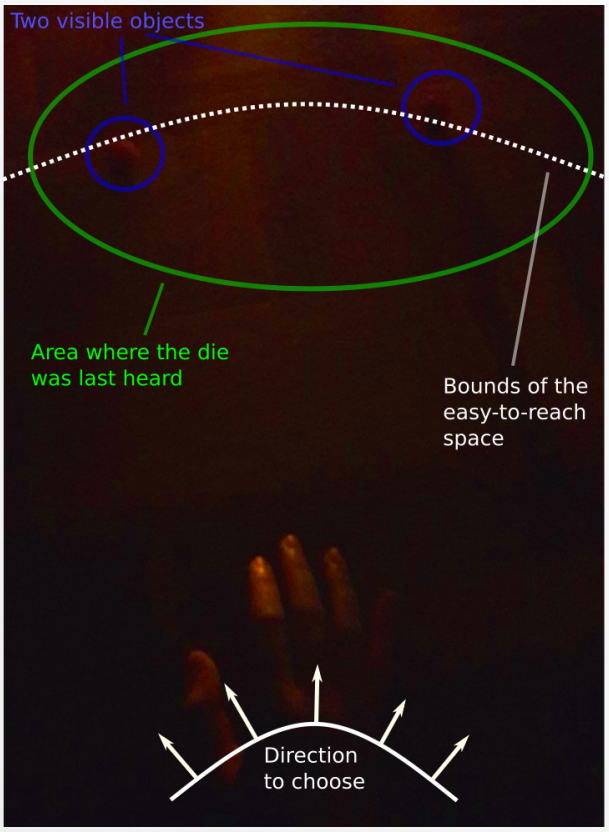


Figure 1.10: All information regarding the position of the objects can be regrouped in a common topological space.

Suppose we are looking at the scene from Bob's hand at ground level. Again, this is a very convenient simplification, as vision, audition and hand movements happen on very different reference frames (respectively eye-centered, head-centered and body-related via proprioception). We make the assumption that at some point, the brain projects all this information on a common topography[3]. Figure 1.11 shows how we could depict the evidence of the right target position depending on the hand direction.

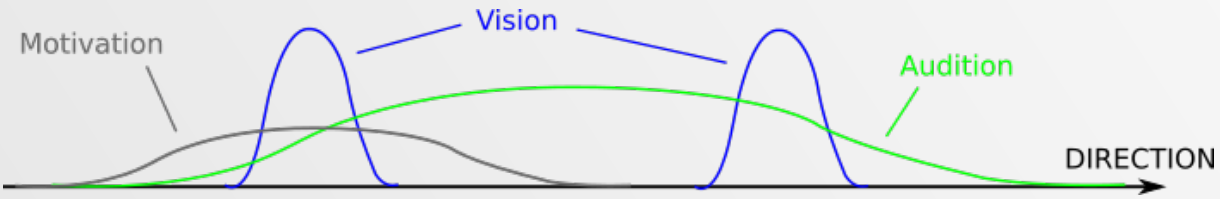


Figure 1.11: The direction of interest can be influenced by multiple modalities, as well as cognitive factors like effort (the object on the left requires less effort to reach), and, again, attention.

This spatial notion, a form of topological encoding, is not restricted to physical position. For example, textures are not limited to one "rough" and one "smooth" qualities, one can find an entire range of textures between these extremes. And all colors can be placed on a representational space inspired from the color spectrum (figure 1.12).

[3] Such projections do happen in neural maps, though probably not to that extent.

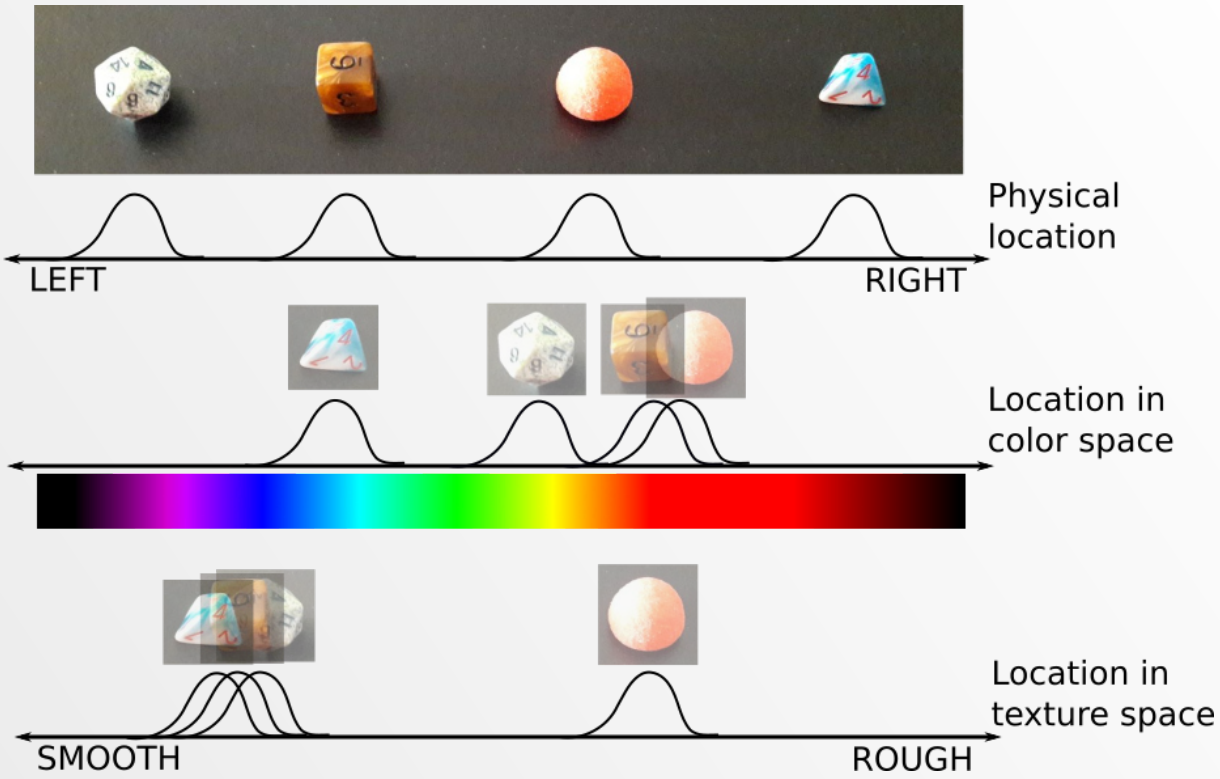


Figure 1.12: Three possible ways of discriminating objects

This is only the beginning! All these dimensions can be combined[4] to construct a new space with all the modalities relevant for a decision (figure 1.13).

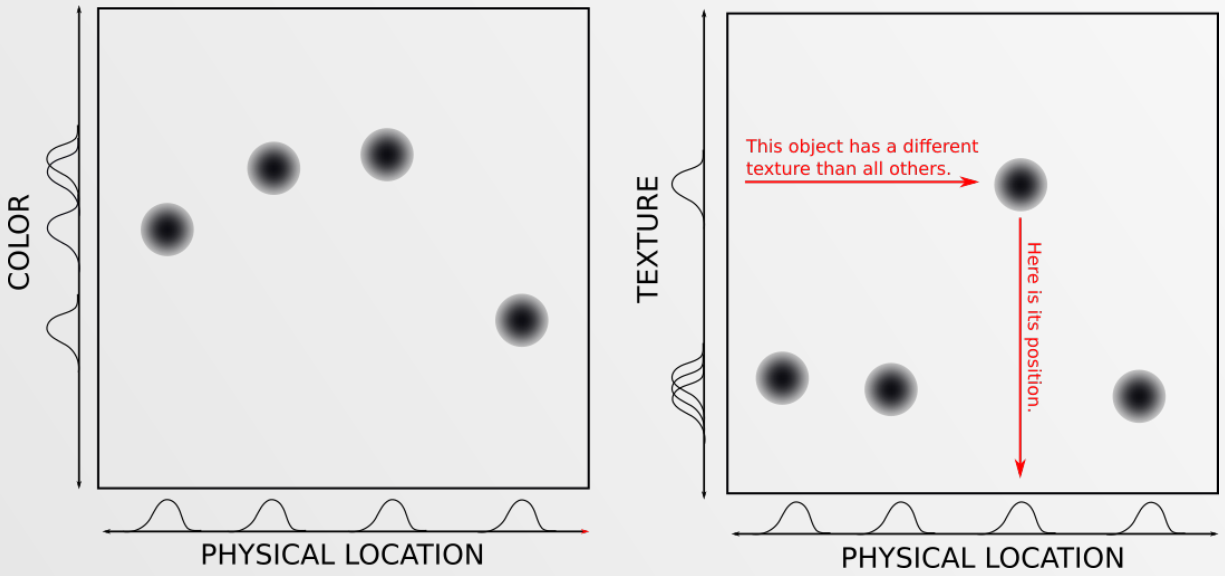


Figure 1.13: Object detection using two modalities. Each spot represents an object in a two-dimensional color-position (left) or texture-position (right) space.

[4] We will stop here, before opening an additional can of worms known as the binding problem, i.e. how different percepts are merged into a single experience (here, how location and color/texture are regrouped in points).

Now, new questions arise: How do we define the spaces of decision? What are its metrics? Figure 1.13 makes it seem like the distance between red and blue is the same as a 10 cm distance: does it make sense? I am inclined to answer that yes, there might exist a local topology in which this occasionally makes sense. The trick is that for every configuration, for every environment and observer and task… the decision lies on a different topology.

Color illustrates this well. The visible color space is sometimes represented as a two-dimensional spectrum (figure 1.14). My observable color space is probably different than yours, because I am colorblind. Yours would become different as well if you were looking below a table in a room lit by a candle. The environment, the context, and maybe even your actions may influence what you perceive.

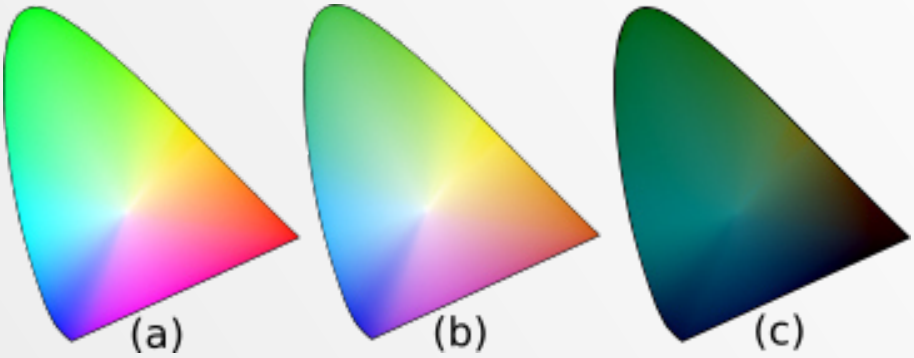


Figure 1.14: Chromaticity diagrams. (a) Colors visible by a healthy person in good lighting conditions. (b) Colors visible by the (colorblind) author in good lighting conditions. (c) Colors visible by a healthy person below a table in a room lit by a candle.

In good lighting conditions, Bob would have had no trouble separating die from candy. But in this setup, his visual space was a big mash of grayscale blurs. And that is the main reason why he started to touch objects in an arbitrary order (the closest one first, to minimize effort).

Actually, with this additional relation of cause and effect, we are about to close the loop we started at the beginning of this introduction. A fundamental property of multimodality is that the reliability of modalities varies, and we can do something about it, and because of this we take action. As you read these lines, your eyes move and carry your gaze all over the page. Because just like Bob decides to touch objects he cannot see properly, you instinctively decide to put the words more or less where you can read them best, usually aligned with the center of your retina. This action allows you to have a better perception, and helps you with the next decision process (figure 1.15).

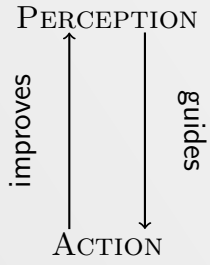


Figure 1.15: The loop of active perception

It is not safe to say that multimodal fusion, decision-making and active perception are three different problems. My thesis is that there is one thing that connects them all and keeps the loop in motion: space. Endless, ever-changing, multi-faceted, space.

As to the story, what happens next? Well, following his failure, Bob's character inadvertently activates a trapdoor and falls in a deep and long slide that brings him to a dark, unexplored basement. He calls his comrades for help, but they somehow seem reluctant to follow him to the depth of the boss's lair.

To be continued at the beginning of chapter 4...

Part I

State of the art and positioning

Chapter 2

Perception translated

Contents

2.1 Introduction

This chapter presents a broad view on perception. This is a mandatory step before we discuss multimodal merging (chapter 3): in order to merge data, you need to collect it first. The theoretical views addressed in this chapter are meant to touch both humans (our source of inspiration) and robots (a potential subject of application). In any case, perception is about an agent receiving information from and intracting with its environment. We will always place ourselves in a context of active perception, i.e. perception results from action and leads to more action. This will motivate our choice of a dynamic computational model in chapters 5 and 6, and although we do not literally implement action in this thesis, our models will fall within a global scope of active perception, as discussed in part III.

2.2 Projections of the world

The world holds a cornucopia of information. Any object within the observable space carries more information than we can count. For example, take a random object on your desk. I will use a coffee cup. Right now, this cup is being hit by waves of photons of different wavelengths sent by an artificial light source. Photons of certain wavelengths will be reflected. Which ones are reflected, depends on some physical properties of the surface coating of the cup. As these light waves are reflected, they reach other objects, including, possibly, my eyes. Specific photoreceptors in my retina are stimulated by these waves, and they send electrical stimulations to my brain, that eventually translates the signal as an intelligible information: I see something pink. Yet from the outside world,

the cup is not pink. It merely possesses physical properties that make it emit specific ranges in the spectrum of light, that my brain identifies as a certain color that I have been taught was pink. If everyone was born blind, the cup would not be pink. Just like we do not say the cup is of infrared color, although, like most warm objects, it certainly emits some infrared light that we do not see. How is it warm? The object has a high internal molecular agitation. I could measure it with a thermometer. Warmth is but a sensation I would feel if I touched the cup with my hand. And it is only one among the many descriptions that could apply to this object.

When I close my eyes, the properties of my coffee cup do not change (quantum considerations aside). When I reopen my eyes, this cup occupies a part of the visible space that is available to my brain. What is visible is already a subset of what is, and what I see is only a projection of what is visible. A topographical projection of everything that lies in front of me. So I sense a projection of the world. And finally, what I perceive is a translation made by my brain of what I sense. Pink is a representation given to the kind of stimulations received by the cones that respond to the light emitted by the cup. A representation that may very well fade away if I turn my attention away from the cup and experience some sort of tunnel vision on my computer screen. So what I perceive is yet another projection of the world. But the topic of attention will come in due time (section 2.3), let us precise our view on representations first.

2.2.1 Internal representations

An important distinction has to be made between the projections that are received by sensors such as the retina, and the projections that are computed by the brain. It is a common intuition that what the retina perceives is projected into the brain to form a conscious representation of what is seen, but that theory faces multiple challenges. To pick one, it is well known that the human eye has a blind spot where the optic nerve passes through the retina. Anyone can find it in a few seconds using very accessible tests. But why do we not see it consciously? One theory is that the brain does have an internal representation of the visible sensory space, and somehow fills in the blanks. However, the necessity of this internal representation is contested.

"The world as an outside memory" Starting from the concern that most studies on visual perception do not account for the blind spot caused by the optic nerve, or the blur due to persistant eye and body movements, [O'Regan, 1992] calls into question the assumption that our seemingly robust and precise vision reflects an internal representation encoded in the brain. He argues instead that what we consciously see is made of high-level knowledge that is repeatedly updated through action. For example, as you read this manuscript, you are not seeing the words you are reading because somewhere in your brain these words are imprinted on a "screen" which is later analysed by the word processing modules in your visual cortex. Instead, the visual cortex processes on-the-fly the pieces of the world that you are currently gazing at. Under this paradigm, the internal representation of conscious vision is an illusion caused by the perpetual availability of the information. Suppose that as you read this text, there is a pile of books in the background of your field of view. Is there a part of your brain that encodes the text of this manuscript and the titles of all the books in the pile? After all, they are all in your field of view. And yet, if you closed your eyes suddenly, you would probably not be able to retranscribe all the titles you were "seeing". You might not even know how many books there are. Unless you gazed at the pile: then you would know how many books there are, and their titles. But you would not be able to tell what you just read in this manuscript. From here comes O'Regan's denomination of an "outside memory": we fetch from the world the pieces of

information we need, when we need it. A similar argument is made by [Brooks, 1991] for artificial intelligence specifically, proposing to reject explicit representations and to use "the world as its own model".

This argument goes strongly in favor of the view that perception is active. [O'Regan and Noë, 2001] follow up on this view with the theory of sensorimotor contingencies. They argue that perception is defined by the rules governing the changes caused by motor actions onto sensory inputs. For example, seeing the distance from one's body to an object is equivalent to estimating the movement required to reach it. Seeing a straight line is recognizing the invariance of sensory data when eyes move in the direction of the line. Under this theory, there is no internal representation of sensory space from which percepts are deduced, because perception is inferred from action (leading to the notion of "active perception", see section 2.3).

Action has to be considered in a broad sense here. When you close your eyes, you are still aware of the presence of the pile of books in the background, even if you have not gazed at it. You can also remember some visual information about it: "a brown cover", "laying on the shelf", "a little to my left". According to the theory of sensorimotor contingencies, you can picture these details in your mind because you are actively focusing on it. You can even point at the pile of books with your eyes closed; not because there is an internal representation of the outside world that your brain would "look at" in lieu of your eyesight, but because you can match some knowledge you have about the object ("a little to my left") to an awareness of one object, among others, lying at a "foveatable by turning my eyes 30 degrees to the left"-position — a position that you have learnt from experience to correspond to a "30 degrees to the left"-direction. And that you have also learnt to be pointable with a specific action of your arm. This awareness of object positions in your field of view can be encoded in only a topographical map of affordable eye gazes.

Neural maps So, the theory of sensorimotor contingencies does not exclude the existence of topographical encoding of information in the brain. Indeed, without going as far as an internal, explicit representation reflecting all that is sensed, the brain certainly contains cortical and subcortical maps that can be linked to organized pieces of information [Rizzolatti et al., 1994]. In the cortex, one possible way of describing cortical maps is under the form of cortical microcolumns, groups of neurons that are interconnected vertically through cortical layers, share some activation and act as a single processing unit [Mountcastle, 1997]. Some of these maps are known to possess topographical connections, i.e., receptive fields of cortical microcolumns retain ordering of the sensory stimuli that activate them [Buonomano and Merzenich, 1998]. While not organized in microcolumns, similar topographical properties have been observed in subcortical maps [Cynader and Berman, 1972]. Note that neural maps are not limited to stimulus positioning, and can encode various observable properties, such as object orientation [Bosking et al., 1997] or color [Li et al., 2014].

We do not mean to solve how information is encoded in the brain. It is sufficient for us to know that there exist topological maps reflecting some pieces of knowledgeable information. Some sort of implicit representations — a misnomer according to [Brooks, 1991], which we may indulge in —, as opposed to explicit internal representations of the world, that we shall agree to discard.

2.2.2 From biological to artificial systems

The distinction between a hypothetical explicit representation, and implicit information coded in cortical maps, is useful to clarify our positioning. As we plan on moving towards

artificial models of perception, it is important that we explain what exactly we model, as that distinction can easily be missed in computer science.

Parallels between human and artificial perception have to be made with caution. The first reason is hardware. Sensors differ by their disposition (two eyes opposed to an arbitrary number of cameras), regularity (fovea on one side, mostly regular resolution on the other), or sensitivity. Meanwhile, processing power differs by orders of magnitude: human brains achieve very complex tasks with minuscule energy compared to computers, which on the contrary scale very easily to very demanding calculation tasks. A computer can calculate long multiplications in an instant, contrarily to humans. Humans can learn to recognize a new object in seconds, while computers may need hours of training. Yes, a lot of inspiration can be drawn from biological systems when designing computer models. But do not expect that tasks, including perception, will be carried out exactly the same way by living and artificial agents.

The second reason is representations. Artificial agents have explicit representations of everything they sense. Pixel-perfect scans of sensory inputs can be copied and stored indefinitely, meaning an artificial agents not only has a permanent inner representation of the present scene as sampled through its sensors, but also the past, only being limited by memory storage and processing power. Any potential implicit representation (through image processing, scene segmentation, object identification…) is produced out of this known explicit representation of the world. This constitutes a slight change from the human paradigm, where computational sets are fed by multiple pathways, which include (and not exclusively) part of (and not all) the sensory inputs. The internal explicit representation of the perceived scene, if it exists, would be reconstructed after the fact.

2.2.2.1 Feed-forward dataflow

We can try to reconcile these two kinds of systems by underlining a dataflow made of what they have in common. We will consider object space from three points of view: physical space (containing all the properties of an object, which can be sampled but never fully known), sensory space (the raw description of what sensors sample from the physical space), and feature space. The latter contains all the (mostly intelligible) dimensions that are relevant for a computational task, object positions, color space, etc. We acknowledge that there are complex pathways from sensory to feature space, which we will sometimes replace with simple computational models (cf. logpolar transformation in chapter 5 or manifold learning in chapter 6). The dataflow we propose is summarized in figure 2.1. Note that only one side of perception is represented: feedback from decision to physical, sensory and feature space exists but is not depicted here.

In both systems, sensory data feeds into feature space through undisclosed pathways. Feature space is constituted of maps holding different dimensions and playing a part in decision-making. Some of these maps can be multimodal. The possibility of an explicit internal representation, containing all pieces of information that are consciously observed, is not particularly relevant.

In artificial intelligence, the shape of feature maps can be chosen arbitrarily, and is often set to be regular, because the data is. That differs from biological maps, that can take various forms depending on physiological factors. One such example can be found in the superior colliculus (SC). The SC will come up multiple times in this thesis. Not only does it hold measurable topographical projections of sensory inputs, it is known to contain multisensory neurons and is involved in attentional and active mechanisms. So it is a bit of an all-in-one structure with regards to topics studied here.

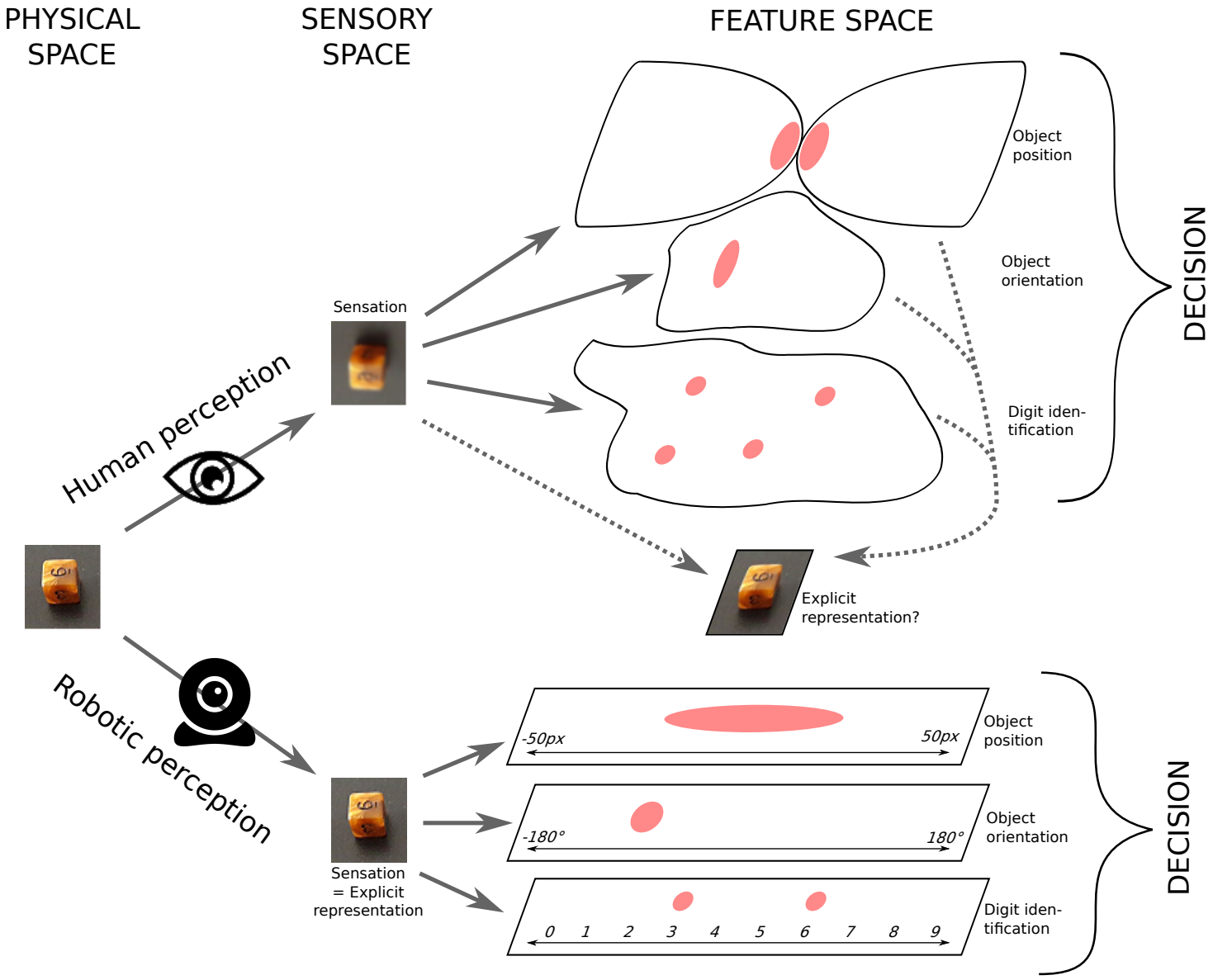


Figure 2.1: Schematic dataflow for feedforward data processing, for biological systems on top and artificial systems on bottom. The existence of an internal representation of the visual scene in the brain is disputed and can be left out of our dataflow. This representation does not show the existing feedbacks from decision to physical, sensory and feature space.

2.2.2.2 Neural feature map example: the superior colliculus

The SC, or its non-mammalian equivalent the optic tectum, is a subcortical structure that has been extensively studied for its performance of multisensory integration [Meredith and Stein, 1986] and its role in generating eye saccades and other movements [Gandhi and Katnani, 2011]. It is made of multiple layers of neurons, with the superficial ones dedicated mostly to visual processing, and the deep ones to sensorimotor processing [King, 2004]. Topographical neural maps have been found in the SC, receiving not only visual stimuli, but also auditory and somatosensory [Knudsen, 1982, Wallace and Stein, 1996]. Meanwhile, bursts of activity in the deep layers of the SC have been shown to correlate to motor commands for gaze shifts [Kustov and Lee Robinson, 1996] and fixation [Gandhi and Katnani, 2011] using retinal coordinates [Klier et al., 2001]. While collicular maps are not retinotopical, there exists a mapping from retinal to collicular coordinates, which can be approximated by a logpolar transformation [Ottes et al., 1986]. Concretely, the retina is split into left and right hemifields, that are each projected in a heterogeneous way, increasing the size of stimuli reaching the center (called fovea), and decreasing their size in the periphery. That is consistent with the heterogeneous distribution of sensors in the retina, which is very dense around the fovea and gradually decreases with distance. As a result, the SC can be depicted as two connected hemifields forming a hourglass shape (figure 2.2).

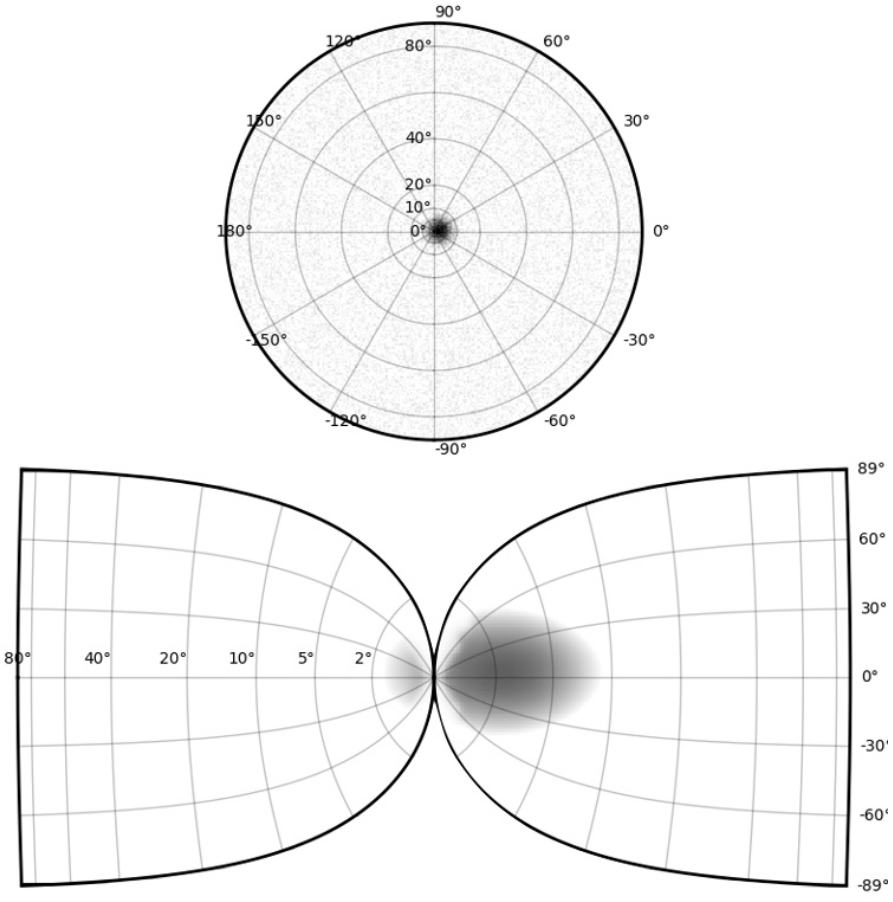


Figure 2.2: Projection of a visual stimulus from retina (top) to SC (bottom). Pictures produced with code adapted from the sources accompanying [Taouali et al., 2015].

The SC displays a rare example of a neural map where the relationships between input, shape and output are (in part) explainable. Its influence on eye movements justifies an encoding of decisions in retinal coordinates, so the feature space lays out the properties of the sensory space. Most importantly, one can argue that the decision-making justifies the

disposition of the collicular map. The latter has been designed, either through evolution, or cerebral plasticity and development, so that it would output commands in the right coordinates, giving an adequate weighting to either saccades or fixation decisions.

This is an aspect of decision feedback that is often overlooked, especially with regards to artificial models. Figure 2.1 only shows the feedforward flow of information, that is not all. Indeed, feedback from decision to physical space is evident (for example, you push an object, it moves away). Feedback to sensory space too (you move your eyes, you perceive the object at a different relative position). But other kinds of feedback can be taken into account, most notably when a developmental point of view is adopted. This would fit under the global paradigm of embodied cognition, which hypothesizes that decisions are made by the body as a whole, instead of the body mindlessly reacting to decisions made by independent cognitive processes. Whether structures like the SC fit this paradigm is still an ongoing research question. Recent work, observing that SC inactivation had an impact not only on saccade generation but on decision-making itself, suggests that it does [Jun et al., 2021].

2.2.2.3 Artificial models of feature space learning

As we mentioned, feature spaces in artificial systems face little constraint. Their construction is often left to the choice of the developer. However, some methods do allow to learn some kinds of feature space automatically. The main recourse is to extract implicit representations from sensory space and find a low-dimensional manifold on which they can be projected.

Nowadays, the first methods that come to mind in manifold learning are deep neural networks [Bengio et al., 2013]. Considered the de facto standard in computer vision, they consist in learning the function from sets of (high-dimensional) inputs to their expected response (of lower dimension), in the form of dozens of layers made of nonlinear combinations of hundreds of parameters. A feedforward pass in the model is a form of dimensionality reduction, and a class of neural networks named variational autoencoders is built on this property [Kingma and Welling, 2019], although an intrinsic dimensionality can also be found earlier in the intermediate layers [Ansuini et al., 2019]. Representation learning has also seen recent progress with contrastive learning algorithms, either supervised [Khosla et al., 2020] or self-supervised [Chen et al., 2020]. In these, the general idea is to learn common points between objects that represent the same concept, and differences between objects that do not. This is the current standard in computer vision. Anyway, deep learning is a wide and very active domain, and this thesis is focused more on biological inspiration and multimodal fusion than on pure learning, so we will not expand on this further in this part. Our work is not incompatible with deep learning, but we will stick to more parsimonious methods of manifold learning when we need it.

There are other, lighter methods designed for manifold learning: self-organizing maps (SOM) [Kohonen, 1982], neural gas (NG) [Martinetz and Schulten, 1991], and growing neural gas (GNG) [Fritzke, 1995b] in particular. These ones will be on focus in chapter 6. We pick those because they are easy to set up and require no supervision, contrarily to deep neural networks. To summarize, a network of neurons is made to expand in a sensory space by repeatedly drawing an input, picking the neurons that match it better and bring them closer to it. The goal is for all possible inputs in sensory space to be represented fairly by the prototypical inputs of the neurons. Then, the connections between neurons, either fixed beforehand (in SOM), or learned via Hebbian-like rules (NG/GNG), form a new topology of intrinsic dimensionality. Note that without additional learning rules, the newly created space is extracted directly from the sensory space regardless of the decision-making task, i.e. learning is unsupervised. So, the prototypes that compose this space are

not necessarily (all) the relevant ones. If all inputs are drawn from a visual scene where the only thing that changes is the position of an object, then a 3D intrinsic dimension can be found (out of as many dimensions as the number of pixels times the number of channels in the camera). If the object changes its position and color, then we can find up to 6 intrinsic dimensions (three positional axes and three color channels), even if the task is only one of localization.

This is one issue that differentiates biological from artificial systems. In the former, years of evolution have lead to complex interwoven processes allowing perception and decisions to improve themselves in an active loop. We have taken the example of the SC, where sensory projections are tied to neural activity in a topographical map, which itself correlates to decision-making and actions that affect perception in return. In artificial systems, it is up to the human operator to make the connections. It is entirely possible to optimize representations to perform tasks in a passive way. We argue that a lot can be gained from closing the loop and allowing decision to feed back into perception and feature processing. This is supported by decades of studies in psychology and neuroscience. While we do not implement active perception ourselves, we feel that it is an important piece of context. In particular, the role of attention will guide some of our choices for most of the contributions.

2.3 Active perception

There are different ways for perception to be active. Macroscopically, perception may require acting on the world, e.g. turning on the light, approaching an object to see it better, turning the page of a book… It may also mean acting on oneself, e.g. turning the eyes to place an object at the center of the field of vision where it is seen better. Or, as put forward in the theory of sensorimotor contingencies, action is how perception occurs in the first place. Let us take a broader look at this topic.

2.3.1 Theoretical frameworks

In this section, we evoke two of the main psychological frameworks of active perception. While they might seem to contradict each other, it might be the case that they are both partially true. Nevertheless, we will use this presentation to issue a general positioning for the rest of the thesis with regards to these theories.

Ecological psychology Ecological psychology, first theorized by [Gibson, 1960], makes the perceiver inseparable from its environment. Under this theory, perception relies not on sensory stimulations but on information present in the environment and experienced through affordances [Lobo et al., 2018]. Affordances are an idea of what the environment may offer that can be acted upon. For example, a ball is not perceived by its diameter in centimeters, or its weight in grams, but by the possibility it affords an organism to lift it, push it, or throw it.

The theory of sensorimotor contingencies is strongly inspired from ecological psychology, in that it sees all perception as active and puts the perceiver in perpetual interaction with its environment.

Cognitivism Cognitivism treats cognition as an information processing system separate from behavior. It is closely followed by computationalism, which assumes that cognition results from neural computations [McCulloch and Pitts, 1943, Dietrich, 1994, Chalmers,

2011]. In particular, [McCulloch and Pitts, 1943] laid the groundwork for most neuro-inspired models of artificial intelligence.

When we ultimately will develop our computational models, our positioning will evidently have close ties to computationalism. But that should not distance us from other paradigms, such as embodied cognition that we mentioned earlier. Embodied cognition is actually classified into postcognitivism, a broad theory that challenges the cognitivist concept of placing cognition and affiliated representation inside the brain. [Villalobos and Dewhurst, 2017] argue, however, that postcognitivism is not antinomic to computationalism, as long as it does not rely on representations. We have already made the choice to leave out some explicit internal representation of the world. The neural maps we plan on exploiting could count as some sort of implicit representation of percepts, so we are only half-way there, but at least the door is left open to present the upcoming computational models in the light of postcognitivist paradigms.

2.3.2 Selective attention

One important key to active perception in computationalist models is found in selective attention [Rizzolatti et al., 1994, Kustov and Lee Robinson, 1996]. That is the mechanism that allows one to concentrate on a given task and ignore irrelevant stimuli. It can be observed in many ways, but to name a few:

- Eye gaze allows to fixate an important target to get the best vision of it while it is of interest. Distractors may appear in the field of view but the observer is sometimes able to ignore them while the task is in progress.
- A noteworthy consequence, the gorilla experiment proposes a famous case of inattentional blindness. The experiment consists in showing people a video of students making passes with a basketball, and asking them to count the number of passes made within one specific team. In the middle of the video, unbeknownst to the viewer, an actor in a gorilla costume traverses the screen and waves at the camera. When the original experiment was led, roughly 50 % of viewers remained completely oblivious to the gorilla for the entirety of the video[1] [Chabris and Simons, 2010].
- The cocktail party effect designates the phenomenon where an observer in a noisy environment (such as a party) is able to isolate a specific set of sounds from the rest (e.g. someone talking in the middle of a crowd) [Arons, 1992].

In general, attention can be divided into two categories [Sternberg, 1996]:

- Top-down attention is guided by cognitive fluxes of information. This includes focusing on someone's voice in the middle of a crowd because the listener is interested in what they have to say; or, for 50 % of unaware viewers, filtering out the vision of a gorilla because it does not participate in the basketball challenge.
- On the contrary, bottom-up attention is triggered by salient features present in raw stimuli. This includes noticing a gorilla in the middle of a basketball game for the remaining 50 % because a gorilla has nothing to do there; or picking up when someone calls your name when you are busy doing and/or listening to something else.

Both phenomenons have been observed and modeled, so it is likely that neural systems contain a combination of the two [Sternberg, 1996].

1. Unfortunately, the experience is no longer reproducible, since lots of people have already seen the video one way or another. I have tested the experiment in a classroom and found only a disappointing 10 % still failed to notice the gorilla. The number was propably low-balled due to the social pressure on students unwilling to admit their obliviousness in front of their classmates.

2.3.2.1 First theoretical models

One of the first models of selective attention was proposed by [Broadbent, 2013]. It makes the assumption that unattended stimuli are filtered out early at the sensory level by top-down processes. For example, if you asked someone to attend to a sentence whispered in their left ear, they would ignore a sentence whispered in their right ear at the same time.

Broadbent's model was contradicted soon after: participants will notice if you whisper their name in the unattended ear [Moray, 1959]. A new model was then proposed by [Treisman, 1960], in which unattended signals were attenuated instead of canceled completely.

An alternative model also suggested that signals be filtered much later, after being processed and interpreted [Deutsch and Deutsch, 1963, Norman, 1968]. Finally, the consensus is that both steps occur [Neisser, 2014].

2.3.2.2 Covert and overt attention

Another way attention can be decomposed is between covert and overt attention. The former happens on a cognitive level and is not manifested psychophysically. The latter is manifested through actions. Evidence suggests that shifts in covert attention are involved in preparing overt shifts [Kustov and Lee Robinson, 1996]. This has been observed most notably in eye saccade generation [Yuval-Greenberg et al., 2014], and in the SC in particular [Krauzlis et al., 2013].

The SC is an interesting example. We mentioned that its deep layers were associated with a topographical map in retinal coordinates, so that bursts of activation would correlate to gaze shifts commands. This means that some amount of covert attention takes place in the SC. This would be mostly bottom-up attention, since signals computed from the superficial layers correlate directly to retinal stimulation [Ottes et al., 1986], although the addition of a top-down modulation is very likely [Fecteau et al., 2004]. Our follow-up question is: how does the SC make this selection? The signals it receives can be very dense, especially if we account for its multisensory input data. Indeed, the SC is also involved in generating saccades towards auditory stimuli [Jay and Sparks, 1987]. Given the abundance of possible targets in a multimodal world, selecting and attending to specific pieces of information becomes primordial. Active perception plays a part in filtering multimodal signals. But at the same time, merging multimodal data may help selecting better targets, as indicated by spatial congruence and temporal synchrony among others.

2.4 Discussion

This chapter focused on perception from an active point of view. Depending on the paradigm, action is either a way to improve perception or the actual mechanism that makes perception. Attention is at the core of this process. Not only does it prepare and stabilize action, it also improves perception by acting as a filter. This is necessary because of the large amount of multimodal information that can be sensed. But at the same time, all this information has to be accumulated and merged in some way, so that the appropriate targets are attended to and the appropriate actions are taken.

The contributions of this thesis are mainly focused on artificial models of multimodal fusion. But there can be fusion in action and fusion without action. Based on biological inspiration, we would be inclined to work on active perception. Our contributions, in part II, are actually quite light in that regard. We use models containing some amount of bottom-up attention, which we believe is necessary for fusion. It might be a drop in the ocean of active perception, but at least that piece of context is present and accounted for.

To expand further on interactions between multimodal merging and attention, a discussion on this is held in [Macaluso et al., 2016]. Here are the main take-aways:

- Attention certainly plays a role on fusion, both bottom-up (e.g., whether or not to ignore low-intensity stimuli in case of conflict) and top-down (e.g. during complex, semantic tasks).
- There is context-dependent interaction between attention and the type of modalities in focus. For example, bottom-up attention will give more importance to vision in spatial tasks, and more to audition in temporal tasks. This effect is diminished by top-down attention.
- Stimulus complexity influences by which nature attention affects multisensory integration, with more complex stimuli favoring top-down attention and simpler favoring bottom-up.

So, attention goes hand-in-hand with multimodal fusion, and the reasons behind the former may actually be found in the latter. This is why we felt the need to discuss active perception before multimodal perception, which brings us to next chapter. In the following, we will focus on multimodal merging in particular.

Chapter 3

Psychophysical and neural accounts of multimodal merging

Contents

3.1 Introduction

There are plenty of reasons to merge modalities. Sometimes one is incomplete (e.g. you hear a sound coming from behind a surface). Sometimes it is ambiguous (e.g. you see two persons in front of you, and need to interact with one). Sometimes it is unreliable (e.g. seeing in a dark room). Merging occurs as a natural solution to lift some of these issues. But it comes with its own challenges. In order to merge modalities, one has to find the right knowledge to combine, i.e. stimulations carrying information about the same property of the same object (e.g. someone's lips moving linked to sound source localization). One has to put them in a common reference frame, and then solve potential incongruences that may arise (e.g. you see someone talking at the center of a television but sound is coming from the sides).

Mechanisms of multimodal merging in humans or animals are studied in both neuroscience and psychophysics, with strong interactions between the fields. Psychophysics are defined as "the scientific study of the relation between stimulus and sensation" in [Gescheider, 1997]. Before even looking at the neural mechanisms underlying multimodal perception, some particular behavioral effects have been discovered early on, from which we cite the most noteworthy ones in this section.

McGurk effect In 1976, McGurk and MacDonald discovered an illusion happening when exposed to an incongruent audiovisual speech signal [McGurk and MacDonald, 1976]. For example, if you are presented with a video of someone pronouncing the syllable "ga" and, simultaneously, a recording of someone pronouncing "ba", then there is a high

chance you will consciously hear "da". An advantage of this illusion is that it can work without any preparation and it remains active even if you are aware of it[1].

Rubber hand illusion First described by [Botvinick and Cohen, 1998], the rubber hand illusion is obtained by hiding someone's arm behind a screen and placing a fake arm on the other side, visible to the subject. After stroking both the real and fake arm in synchronization for a certain amout of time, the subject starts feeling ownership of the fake arm and believing that they feel touch on it.

Ventriloquist effect Named after ventriloquist shows, where spectators are under the impression of hearing a puppet talk while the sound is produced by the puppeteer, the ventriloquist effect describes an audio or visual capture of stimulus localization [Bertelson, 1999]. Most commonly known on the side of visual capture (i.e. the localization of a visual stimulus captures the localization of an auditory stimulus, deemed less reliable), the effect was also reported to function the other way around (audition may capture vision when it is blurred a lot), or to lead to a compromise (an incongruent signal perceived in-between the visual and auditory stimuli) [Alais and Burr, 2004].

The common point between these three effects is that a subject is confronted to (spatially or phonetically) incongruent stimuli and a capture or interpolation effect happens to merge them into a single percept. Similar effects have been observed with other types of incongruences. Visual and tactile perception of object size can capture one another [Ernst and Banks, 2002]; same with textures [Calvert et al., 2004, chap. 7]. In the temporal domain, the perceived times of occurence of a visual flash and an auditory click presented in succession are biased towards one another [Fendrich and Corballis, 2001]. The perceived number of occurences can also be influenced [Shams et al., 2002].

Multimodal merging manifests in many different ways and involves many different mechanisms. We can isolate a few challenges: Under which conditions are modalities combined? How are they combined when they are sensed in different reference frames? What factors modulate the response? We review part of the literature on these questions in the next three sections respectively. We follow with some theoretical models of multisensory integration in section 3.5. Note that this chapter gives a quick overview of the state of the art, but each chapter of contributions will also have a dedicated bibliographical study.

3.2 Factors in modality combination

One must draw a line between stimulations coming from a common source and stimulations from different sources. In the first case, there is reason for combining them into a single percept. In the second case, it means either they should be processed separately, or one should be selected over the other.

Knowing when to combine stimulations is already an arduous task, as many mechanisms can be used by the brain to infer that they come from a common source. Many factors can intervene, we describe here the most important ones.

1. For a demonstration, the following video goes straight to the point: https://www.youtube.com/watch?v=aFPtc8BVdJk. Try watching it with sound off, listening to it without looking, then watching with sound on. You should perceive three different sounds: respectively "ga", "ba" and "da". Note that it does not work on everyone — a minority of people still hear "ba" in the bimodal condition.

Temporal synchrony The most reliable factor would be temporal synchrony, i.e. we mostly combine stimulations that occur simultaneously. A strict synchrony is not necessary however [Munhall et al., 1996], and it might depend on the modalities involved, with audition being favored in the temporal domain [Calvert et al., 2004, chapter 2]. But in any case, some proximity in time is necessary for the combination, for example there is an estimated 200 ms time window for vision to influence sound in the McGurk effect [Van Wassenhove et al., 2007], and up to 100 ms in the ventriloquist effect [Slutsky and Recanzone, 2001].

Spatial congruence Similarly to time, spatial congruence can have an effect on the combination, with some limitations [Jack and Thurlow, 1973, Slutsky and Recanzone, 2001]. The ventriloquist effect is a clear case of multisensory integration of spatially-incongruent stimuli.

Attention Finally, as we mentioned earlier, attention and task relevance may influence how modalities are combined [Talsma et al., 2010, Macaluso et al., 2016]. There is no unanimous theory as to by how much attention plays into modality combination. Does the McGurk effect occur because we are visually attending to the scene? There is evidence in favor of this, or at least, that diverting attention away from the scene greatly reduces the effect [Alsius et al., 2005]. Reverse arguments have been made for the ventriloquist effect [Bertelson et al., 2000, Vroomen et al., 2001]. This is a topic that can only be studied case by case, and that we cannot expect to review or generalize here.

3.3 Reference frames

Another open question on multimodal merging is how the modalities are placed in a common reference frame before being compared or interpolated. After all, the sensory space of visual perception is not the same as the sensory space of auditory perception. On the sensor side, they do not even have the same reference frame — vision is eye-centered, audition is head-centered. But there is a point where they are put in a common space so that they can be merged and a decision can be made.

Some neurons are able to receive stimulations originating from different modalities at the same time. Multimodal maps have been found in some brain regions, that contain a mix of multisensory and monosensory (in multiple modalities) neurons [Allman et al., 2009, Meredith et al., 2020]. We evoked the SC in the previous chapter in the context of saccades, and it happens to contain such multimodal maps.

The SC is known to integrate cues from multiple modalities, including visual, auditory and somatosensory [Wallace and Stein, 1996, Calvert et al., 2004]. Its deep layers have been reported to contain unisensory visual, auditory and somatosensory neurons, and multisensory neurons [King, 2004]. It has been suggested that alignment between sensory modalities is — at least in SC — guided by vision [Knudsen and Brainard, 1991]. Despite the apparent vision dominance, behavioral experiments show that other modalities do have a strong effect on the visuomotor decision-making process, positive if congruent, negative if not [Stein et al., 1989].

Calibration Little is known on how multisensory maps are formed. In the SC for instance, evidence suggests that they are developed after birth [Wallace and Stein, 1997]. It is possible that they are calibrated from sensory experience. Indeed, even in adulthood, some recalibration effects are known to happen.

Recalibrations have been observed under the form of after-effects of fusion of incongruent stimuli, most notably after ventriloquism [Radeau and Bertelson, 1974, Frissen et al., 2012, Mendonça et al., 2015, Bosen et al., 2017], but not exclusively [Xu et al., 2018]. It appears that after being presented spatially-incongruent audiovisual signals, subjects will quickly, but temporarily, recalibrate their auditory localization map to compensate the discrepancy, leading to a bias in future sound localization towards the previous visual stimulus position.

3.4 Modulations

Once stimulations are set on a common ground, there should be a way to form a single percept out of them. There are two main scenarios possible: either the stimuli are congruent, and one can expect an enhanced response, or they are in conflict, that needs to be solved.

Multisensory enhancement Responses to multimodal stimuli in the brain depend on their congruence. On simple tasks, at a neural level, if stimuli are congruent then one observes some multisensory enhancement, i.e. the neural discharge in the multimodal condition is stronger than the sum of discharges in all unimodal conditions [Wallace and Stein, 1996]. On the contrary, conflicting stimuli lower the response.

Interpolation Solving a conflict means either finding an interpolation between the unimodal stimulations, or selecting one (which is a special case of interpolation where one weighs for 100 %). Historically, it has been thought that modalities captured the decision in tasks where they were specialized: mostly visual capture in spatial tasks, and auditory capture in temporal tasks. That has been challenged by psychophysical experiments showing that the modality deemed the most reliable would take precedence [Ernst and Banks, 2002], and even vision, if blurred enough, could give in to audition in spatial tasks [Alais and Burr, 2004]. When modalities have the same reliability, and if conditions for interpolation are met (e.g. the unimodal components are not too spread out), the multimodal stimulation is perceived at a midpoint between its unimodal components.

Modulation by attention Unsurprisingly, the weighting of modalities can also be modulated by attention [Driver and Spence, 2004], a process which is very context-dependent. For example, there is evidence against attention modulating fusion in the ventriloquist effect [Bertelson et al., 2000, Vroomen et al., 2001], but there is evidence in favor of it for the McGurk effect [Andersen et al., 2009] and the rubber hand illusion [Thériault et al., 2022].

3.5 Models for fusion

Numerous theoretical models have been advanced to account for multisensory integration in the brain at behavioral and/or mechanistic levels, with or without neural plausibility. Some models, like the first one below, are purely probabilistic. Others attempt to explain what computations lead to fusion as described in psychophysics. For reference, we propose a quick overview of these paradigms in this section. For some algorithmic implementations, we refer the reader to chapter 5.

Maximum-likelihood estimation In many instances, the merging effect was quantified, and often found to correspond to maximum-likelihood estimation (MLE) [Ernst and Banks, 2002]. The prerequisite for this comparison is for the effect to be measurable along a spatial dimension: stimulus localization for [Alais and Burr, 2004], object shape for [Ernst and Banks, 2002]. Psychometric functions allow to estimate the reliability of each modality (inversely proportional to the variance of answers given by the subjects in unimodal trials). Then, the average answer in a bimodal trial lies at a barycenter of each unimodal component, weighted by their reliability. See [Rohde et al., 2016] for a full description of the protocol.

While the MLE model is easily applicable to spatial multimodal merging as in the ventriloquist effect, it is not so much to other domains as in the McGurk effect — where is "da" between "ga" and "ba"? There is a chance that there exists a cortical map in which these sounds are topologically organized, but psychophysical measures of it are not that straightforward [Jiang and Bernstein, 2011]. The ventriloquist effect, on the contrary, can easily be placed spatially, both in psychophysics and in neuroscience. This justifies why we put some more focus on this specific effect in chapter 5.

The Bayesian brain There is no clear consensus as to how the SC, or the brain in general, performs statistically-optimal integration as described by MLE. Either the brain computes and memorizes a representation of stimulus uncertainty, or the activity in some neurons actually encodes a measure of probability. The latter hypothesis is priviledged, under the form of probabilistic population coding [Deneve et al., 2001, Pouget et al., 2002, Pouget et al., 2003], sometimes referred to as the "Bayesian brain" [Knill and Pouget, 2004]. There is little experimental neuroscientific evidence supporting this theory, but it was shown through simulations to be plausible [Ma et al., 2006].

Predictive coding A related theory supposes that some regions of the brain perform predictive coding [Srinivasan et al., 1982]. In some versions of this theory, Bayesian priors are represented by probabilistic models of the environment, that serve as prediction to following perceptions. Discrepancies between prediction and perception are then used as feedback to update the internal model. Like the rest, there is no proof that the brain implements this type of coding, although plausible comparison to cortical microcircuitry have been made [Bastos et al., 2012]. Folding back to multisensory integration, [Talsma, 2015] argue that it could indeed be explained in light of predictive coding, as long as attentional processes are implicated.

Free energy Predictive coding has later been integrated into the free-energy principle [Friston and Kiebel, 2009, Parr et al., 2022], which states that decisions in the brain are oriented towards minimizing free energy, and thus avoiding surprises [Friston, 2010]. Under this principle, multimodal merging is bound to occur in a statistically-optimal manner, being the best way of minimizing the probability of surprises. One criticism of this theory is that under this principle, humans would look for a dark room and stay there. [Friston et al., 2012] tackle this so-called dark-room problem, adding the precision that surprises depend not only on sensations but on the agent as well, so a dark room would be surprising if the agent expected a stimulating environment.

3.6 Discussion

Multimodal merging is a very intrincate process involving many more mechanisms than we could describe in a single chapter. Our goal was only to give a quick overview, as

we do not intend to implement a full bio-realistic model of multimodal merging. We will however propose a bio-inspired computational model of it. The main challenges to take away are: finding what properties should be combined, placing relevant stimulations in a multimodal reference frame, and weighing them appropriately depending on their reliability (and possibly other factors). In the context of this thesis, we wish to use a model that allows attentional behaviors. To that end, the probabilistic models presented above may not be sufficient.

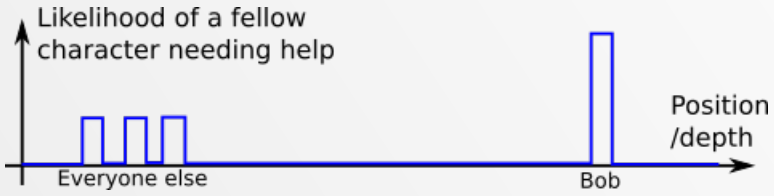
In another category of models, neurons can be simulated as individual processing units, where decisions are encoded by firing rates or membrane potential. At a population level, they may encode perception and decisions at a mesoscopic level, differing from the models presented above that offer more of a macroscopic overview. Neuro-inspired implementations can be fit for multimodal merging tasks, either through learning (see a brief overview on this in the beginning of chapter 6) or by exploiting spatial (and possibly dynamic) properties of stimuli. Models in the latter category, without learning, will be our focus in this thesis. To that end, in chapter 4, we propose a review of decision-making models, ranging from MLE to these neuro-inspired implementations. Here we pose multimodal fusion as a particular case of decision-making. In particular, in chapter 5, we will pick one of the models in this review, one known for its dynamic and spatial integration properties, and study how to benefit from its properties in the context of multimodal merging. The ventriloquist effect will serve as a benchmark there.

Part II

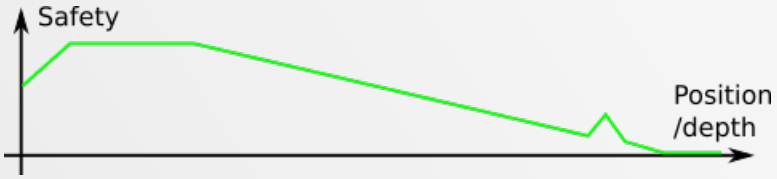
Contributions

Interlude

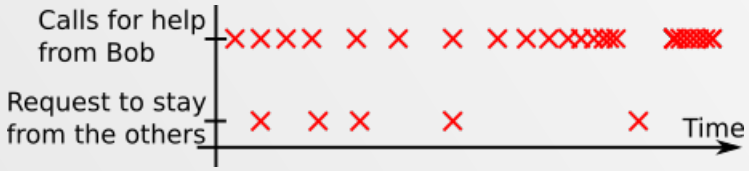
Back to the story. After Bob's character fell from a trapdoor into a dark, deep pit, and despite distant cries for help attesting to his survival, the four other players wholeheartedly agree that it was nice knowing him and that they should keep exploring the dungeon horizontally. Eve is struck by remorse though. Eve plays a healer, meaning her main task is to keep the other characters alive, even Bob's. She has to choose between joining Bob's character and staying with the group. On the one side, Bob is more likely to need help than anyone else. On the other side, the expectancy of a healer being useful is higher in the company of three troublemakers than one.



Both choices are defensible. Meanwhile, Eve must watch out for her own safety, and Bob is not very reliable in that regard.



There is a clear argument in favor of staying with the group. But as Eve ponders, Bob's pleas for help get increasingly insistent. The other players ask her not to go, but not with the same persistance.



So, if Eve were to consider the increasing accumulation of arguments on Bob's side, she would probably favor helping him. Then what choice will she make? Well, it depends on what she sets her main task to be: helping people in highest danger, optimizing the expectancy of being useful, surviving, doing what the other players ask... Or possibly a combination of some of these criteria. There are many different ways of integrating the available information and making decisions. And there are many ways to model such a decision-making process. This is precisely what we are going to formalize in the upcoming chapter.

Chapter 4

Paradigms of decision-making

Contents

The content of this chapter is adapted from a review in progress. Mathieu Lefort and Jean-Charles Quinton started preparations on this work before my thesis, then I picked it up and did the benchmark, part of model formalization and implementation, and most of the writing. This work was made in collaboration with Flora Gautheron, who does a PhD involving accumulator models (DDM, LCA, etc.), and Léo Pio Lopez, who specializes in Bayesian models. Additionally, I co-supervised an internship by Jose Villamar, who did a part of the model formalization and added functionalities to the code. The review is currently in preparation for journal submission.

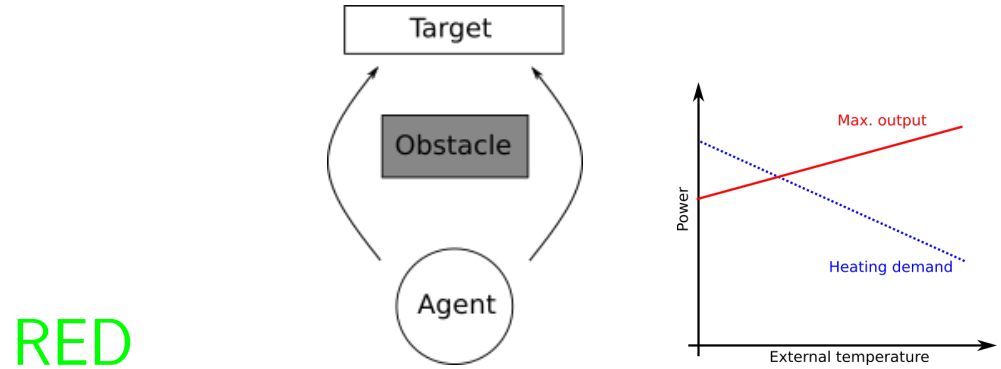
4.1 Introduction

Decision-making is defined as "the process of acting upon the best information available in order to determine the most appropriate course of action" (Oxford dictionary). This definition touches as much human behaviors as artificial systems created by humans. One domain that springs to mind is robotics, which involve a large set of tasks: selecting an object, navigating towards it, reaching and grasping it, manipulating it. Each step requires taking in information from not only the robot and its target, but also its environment, other robots (multi-agent systems), humans, and more. And even one process in robotics can involve other disciplines, most notably computer vision (categorization, tracking…) and machine learning (classification, behavior prediction…).

In humans, decision-making is studied from numerous points of view, including psychophysics, neuroscience, social science, economics. In these domains, studies ask "what decision is made" as much as "how the decision is made" [Gold and Shadlen, 2001, Gold and Shadlen, 2007, Lepora and Gurney, 2012], but in artificial systems, this duality is less blatant. The "what" has the main focus, with models developed to explain or reproduce complex behaviors, but the "how" is often relegated to dynamic integration on the sensor side. The actual selection process is rarely discussed, although studies on this aspect are gaining traction, with the recent development of "explainable" artificial intelligence for example.

As we look into artificial implementations of decision-making, it is striking that as the problems to tackle increase in complexity, the employed solutions become increasingly task-focused and less system-focused. Today, the pinnacle of dynamic task resolution in high-dimensional settings is found in deep reinforcement learning [Mnih et al., 2015, Arulkumaran et al., 2017]. In a word, a model is trained by running simulations and setting rewards depending on the outcome, and behaviors are obtained by tuning a neural network made of thousands of neurons so that the reward expectations are maximized. Learning is being put forward as the go-to bridge from a problem to its solution, but little focus is put on the capabilities of the system itself. On the contrary, devices such as Braitenberg vehicles [Braitenberg, 1986] show that even the simplest systems can exhibit many interesting properties, as long as we put a bit of focus into how the system works.

As to "how" the decisions are made, there are actually many similarities between domains. Be it a human saying how a word is colored (figure 4.1, a), a robot choosing whether to turn left or right (figure 4.1, b), or a heater increasing or decreasing in power to adjust room temperature (figure 4.1, c), decision-making is always about taking in a set of various features extracted from stimulations, and outputting a *condensed* and *exploitable* view of it.



(a) (b) (c)

Figure 4.1: Examples of decision-making tasks. (a) Presentation used to demonstrate a type of color–semantics Stroop effect [Scarpina and Tagini, 2017]. (b) Situation where an agent has to choose a direction to reach a target. (c) Graph giving boiler temperature command depending on outside temperature.

Condensed, because one expects an unambiguous response to a set of stimuli that can be diverse, conflicting, and sometimes extremely dense. In the second example, a single camera may acquire a mountain of evidence: position and nature of objects, their relative proximity, and perhaps all sorts of visual indications such as warning signs or movement detection. All of this has to be integrated into a single decision, namely, what direction the robot should follow. Two issues need to be addressed here.

1. The information may or may not carry topological meaning. In example (a), the decision space is categorical ("green" or "red"). In example (b), one could think

of the 3D environment in which the scene takes place, or the 2D horizontal plane, but for this given task a 1D axis of orientation could suffice. A decision would be any angle in a continuous 360° interval. Example (c) is not about physical positions, but the decision still takes place in a continuous space, with variable temperature measurements calling for different degrees of power. The information is never infinite, even when processed at the level of rawest sensory inputs, because every agent is at some point limited by the resolution of its sensors. But features can be high-dimensional enough that sometimes not all of it can be processed. In this high-dimensional space, a robust decision has to be made. Given multiple temperature measurements advising for different power adjustments (c), a single command has to be made, possibly in a trade-off. When choosing between two paths, left or right (b), going in the middle is possible but ill-advised. And sometimes, no compromise is possible (a), as adding an underlying topology (color spectrum) would lead to unwanted decisions (such as answering "yellow" because it is at a midpoint between red and green in the color spectrum).

2. Information may come from different modalities, which do not necessarily fit well together. Before choosing which modalities should weigh in, and by how much (which is already a decision-making problem in itself), one needs to find a common ground for the decision to take place. Should the topology be unimodal or multimodal? We treat that question in chapter 6. For now, we always assume that there is an underlying (1D) space in which inputs can be projected and a decision can be made.

Exploitable, because the output of a decision-making model is meant to be used by another system: either motors or other decision modules. It will sometimes be of use to the same model in cases of recursion. Indeed, some stimuli can be time-dependant, and some algorithms make use of previous internal states to generate the next output. Internal states can take multiple forms, from priors to membrane potentials, and are not always directly interpretable (cf. activation of hidden layers in neural networks). But in the end, models must produce either an activity (which we define as a set of values for each possible choice) or an output (a singular value usable as an intelligible answer, e.g. a motor command).

Nevertheless, from one domain to another, from one task to another, characteristics expected from a decision-making system can be very diverse. A focus can be put on dynamic and spatiotemporal properties. Systems that are meant to interact with their environment may generate sensorimotor behaviors [Lepora and Pezzulo, 2015] in which decisions influence actions and reciprocically. In particular, decisions might take the form of a sequence of events, as hypothesized by reinforcement learning [Kaelbling et al., 1996] for instance. Especially in dynamic tasks, it may be expected from the system that it guarantees some amount of stability: the ability to focus on targets and also react to sudden stimuli, while showing robustness to noise and to unwanted distractors. Another challenge is in data integration, as decision may require interpolation from incomplete or ambiguous data, generalization from randomly fluctuating inputs, and merging possibly incongruent signals.

Meanwhile, decision-making models have to face a variety of contraints. The first is an issue of scalability. Inputs may bear a high dimensionality, which are increasingly harder to process for models of high complexity. At the same time, algorithms may be faced with computational constraints: limited processing power, memory, time… Part of this can be mitigated with some optimizations: some algorithms may be required to perform numerical approximations (e.g. replacing convolutions with FFT), data reduction (PCA), and changes in structure and coding of data. Another constraint is the possibility (or not) to parallelize computations for speed. Sometimes, it might even be a requirement to

make an algorithm distributed (as in multi-agent systems), or centralized instead.

This chapter presents a review of decision-making algorithms, with an increased focus on neuroscience, psychophysics and robotics. We make the choice to highlight the readily-available properties of a representative sample of existing models, and present them inside an unifying framework. In particular, we restrict ourselves to models with intrinsic behavioral properties, not learned properties. We also refrain from presenting all derivatives of popular architectures in this framework, and mostly stick to the most essential implementations.

4.1.1 State of the art

Many decision-making architectures have been developed in many different fields. We cannot make an exhaustive review of all of it, so we make an overview in the fields of robotics and neuro-inspired cognitive systems. To pick a representative sample:

- Classification algorithms (decision trees, deep neural networks, self-organizing maps) can be used to learn relationships between data and a potential decision (often made legible under the form of a "best-matching unit"). This implies the execution of a training phase, so this is not our focus in this chapter.

- Same goes for regression methods. Among others, we can cite uses of Gaussian processes [Rasmussen, 2004], Gaussian mixture models [Plataniotis and Hatzinakos, 2000] and locally-weighted projection regression [Vijayakumar and Schaal, 2000] in robotics [Khansari-Zadeh and Billard, 2011] for example. See [Sigaud et al., 2011] for a survey and unifying framework on these methods.

- Fuzzy logic (FL) describes operations made on fuzzy sets, where truth values are no longer binary but instead compared to membership functions expressing possibility values (between 0 and 1) [Zadeh, 1965, Bellman and Zadeh, 1970, Dubois et al., 2004, Dubois and Perny, 2016]. By fuzzifying sensory inputs and combining their membership functions, one can create fuzzy commands, that can be exploited in computer vision [Krishnapuram and Keller, 1992, Sobrevilla and Montseny, 2003], data fusion [Russo and Ramponi, 1994], or robotics [Wakileh and Gill, 1988, Bajrami et al., 2015, Qureshi et al., 2018].

- Probabilistic models process data to make inferences about probability distributions or parameters. Many models have been used for data fusion, from maximum-likelihood estimation (MLE) through Bayesian inference [Castanedo, 2013] to Kullback–Leibler divergence minimization [Doki et al., 2015] (used in e.g. variational autoencoders [Kingma and Welling, 2019]). Bayesian models in particular have an implementation in Kalman filters [Kalman, 1960], which find many applications in robotics [Chen, 2011].

- In voting systems, evidence is gradually accumulated over time until a given threshold is reached. The simplest of these is the drift-diffusion model (DDM), on which the race model is built [Vickers, 1970, Bogacz et al., 2006, Bogacz et al., 2007]. Once the threshold is reached, the unit with maximal potential makes the decision. DDM have numerous extensions: feed-forward inhibition (FFI), Ohrstein–Uhlenbeck model (OUM), leaky competing accumulator (LCA) [Usher and McClelland, 2001], and pooled inhibition model (PIM) [Wang, 2002]. These models are mostly used in psychophysics and neuroscience [Gold and Shadlen, 2007, Ratcliff and McKoon, 2008].

- Dynamic neural fields (DNF) are population-coded accumulator models running on a topological map [Amari, 1977, Schöner et al., 2015]. The decision is read from a weighted sum or argmax of the model output. DNF have had numerous applications in neuroscience [Wijeakumar et al., 2017, Buss and Spencer, 2018] and robotics [Sandamirskaya, 2014, Tekülve et al., 2019, Grieben et al., 2020].

- Finally, some of the previous paradigms can be combined into hybrid architectures [Sun et al., 1999]. That is one of the model families explained in [Goertzel, 2014].

In some of these, there are already partially cognitive components. Hybrid models have been advanced as a basis for artificial general intelligence using the concept of "cognitive synergy" [Goertzel et al., 2011], i.e. the coordination of multiple different processes leads to smooth and rich new behaviors. MLE reflects computations observed in psychophysics [Ernst and Banks, 2002], while DNF simulate the interaction of cortical columns in neural maps [Amari, 1977]. Some relations can be drawn between these models, for instance [Bitzer et al., 2014] and [Gepperth and Lefort, 2016] argue that DDM and DNF respectively provide a plausible implementation of Bayesian inference. On the other side, models used in robotics can also retroactively be used to explain cognitive behaviors [Lepora et al., 2012]. Usually, different domains call for different models, and there have been little attempts at unifying these decision-making algorithms in a field-agnostic formal setting. That is one of the objectives of this chapter.

4.1.2 Objectives

As we mentioned, we want to pick a representative sample of existing learning-free decision-making algorithms. We find that they are mostly divided in three families: logic-based models, probabilistic/Bayesian models, and dynamic accumulators. Our selection is made of both simple and advanced models from each category, going from bare aggregators such as winner-takes-all (WTA) and weighted sum (WS), to more complex methods such as FL, DNF and KF. We formalize them using common notations in order to emphasize their different characteristics: topology-based interaction between processing units, output aggregation, recursion…

We set up toy examples to display the qualitative properties of each algorithm. Our focus is mostly on the decision, although we can also measure numerical values of model activity, a quantification of model internal state. Our purpose here is not to tune or train models to fit complex behaviors, we stick to the emergent properties of standard and isolated models.

In next section, we describe the models selected for the formal unification and comparison, as well as the scenarios they are tested on and the way their outputs are read. Section 4.3 gives the results and comparisons of all the models on all the scenarios. We conclude and add perspectives in section 4.4.

4.2 Methods

In this section, we start by describing the experimental setup in section 4.2.1. Then we explain how models are evaluated in subsection 4.2.2. Then we present the core of all models in subsection 4.2.3. In the end, each simulation is made of a combination of up to three parts: scenario generation, model processing, and aggregation (sometimes included in the model). See figure 4.2 for a formalization of this process.

4.2.1 Scenarios

Time scale Our simulations take place in discrete time. Scenarios are defined over a finite series of timesteps, the step Δt being constant throughout the simulations. While the value of Δt can have visible effects on the behaviors of models with temporal integration, we use a sufficiently low step time in order not to hinder the performance of any model. This choice is consistent with real-life applications of decision-making algorithm, with the perception of artificial agents being limited to a certain amount of frames per second, as well as psychological modeling, with neurons having a finite fire rate.

Working space All scenarios will take place in a topological space X. For computational reasons, and even if the original decision space is not necessarily discrete, we discretize X into a regular set $\{x_1, x_2 \dots x_n\}$. A stimulus i is characterized by an amplitude a_i at position x_i. Position is to be taken in a very broad sense, as the x_i could for example designate semantics for simulations that do not rely on a topology.

A scenario is characterized by a set of sparse stimuli $\{a_{i_1}, \dots, a_{i_N}\}$. Equivalently, we can write a scenario as a set of amplitudes for all positions: $\{a_1, \dots, a_n\}$, assuming that most a_i (where there is no stimulus) are equal to 0.

Some models are designed to operate on continuous topological maps, and are hardly able to process sparse inputs. For these models specifically, stimuli are projected in a base of Gaussians of fixed shape. The projection is also discretized on positions $\{x_1, x_2 \dots x_n\}$. See for example the second square of figures 4.7 and 4.14. We do not consider this preprocessing step to be part of the model itself, but instead a byproduct of sensory perception (e.g. receptive fields, that describe the continuous sensory area in which a stimulus can excite a neuron).

Noise treatment The following models have very different relations to noise. Perception and control result of multiple, interwoven processes, and models integrate this bundle of mechanisms (and all its aleas) with different levels of abstraction. Some will consider that noise is part of the decision process, and treat it like a supplementary parameter. Some assert that noise is statistically estimable from the inputs, and that estimation is part of the results. Some do not process noise unless it is added manually to the inputs. To put all models on an equal measure, the scenario we use are all deterministic. The different approaches to integrating noise will be discussed further in the manuscript for one or two of the models, and especially in chapter 5, where it plays a crucial role.

Simulation plan We set up eight non-stochastic scenarios that determine the inputs to give to all models. They are presented in table 4.1. Each stimulus plotted is shown as a thick bar. For non-spatialized models, only the amplitude of the bar is taken into account. For some other models, the bar will be replaced by a Gaussian.

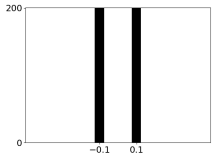
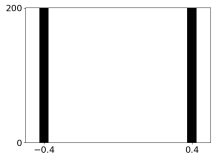
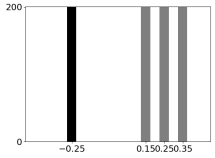
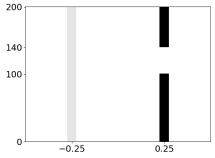
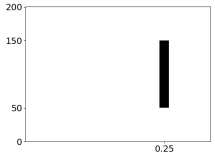
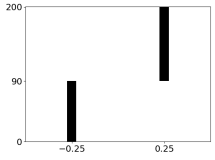
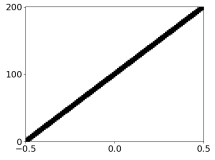
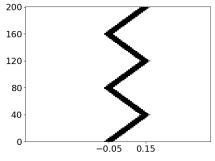
The scenarios were picked to show the various spatiotemporal properties of the models, so they include cases where, depending on tasks, interpolation between signals is likely, and cases where it is not, as well as dynamic settings to evaluate attentional properties and reactivity.

4.2.2 Aggregators

Depending on the model, two kinds of inputs can be read, sometimes both:

1. A positional decision $\bar{x}$, possibly accompanied by an activation value $\bar{y}$. $\bar{y}$ can sometime be related to an estimation of the certainty of the decision.

Table 4.1: List of scenarios. For all inputs, the horizontal axis gives the stimulus positions, and the vertical axis their evolution over time. Stimuli are represented as thick bars for better visibility, with hues proportional to stimulus amplitudes. Each scenario contains between 1 and 4 stimuli.

Inputs	Values		Problematic
	Left	Right	
	1	0.99	Given two *close* stimuli, does the model select one over the other? or merge them into a single percept?
	1	0.99	Given two *distant* stimuli, does the model select one over the other? or merge them into a single percept?
	1	0.5 0.5 0.5	Given a strong stimulus on one side and multiple lower stimulus on the other side (with a summed amplitude higher than the isolated one), which side does the model favor?
	0.1	1/0/1	Is the model robust to temporary obstruction? Or does it lose focus as soon as the stronger target disappears?
		0/1/0	How fast does the model react to a stimulus appearing or disappearing and take/lose focus?
	1/0	0/1	How fast does the model react to a new stimulus appearing instead of another one and switch focus?
		1	Can the model track a target?
		1	Can the model smoothen trajectories?

2. A set of activity values y_{i_k} for all stimulated x_{i_k}.

In order to make a decision, we want to extract a singular value $\bar{x} \in X$ after the model processing in all cases. It does not necessarily have to be one of the x_i. When a model does not include a way of reading the decision directly, we need to add an aggregator to compute the decision localization from the model activity. It takes the following form:

$$\bar{x}(t) = \sum_i w_i(t) x_i \tag{4.1}$$

This is a weighted sum of all evaluated positions. The weights w_i depend on the activities y_i, and can be configured in mainly two ways:

- Plain barycenter:

$$w_i(t) = \frac{y_i(t)}{\sum_j y_j(t)} \tag{4.2}$$

  i.e., all units contribute proportionally to their activity.

- Mean of maxima (or argmax): Let $S(t) = \text{argmax}_i(y_i)$.

$$w_i(t) = \begin{cases} 1/|S(t)| & \text{if } i \in S(t) \\ 0 & \text{if not} \end{cases} \tag{4.3}$$

  where $|S(t)|$ denotes the size of set $S(t)$. In short, this is a barycenter of all units of maximum activity. Very often, there is a unique maximally-activated unit, in which case this aggregator is essentially an argmax.

4.2.3 Models

4.2.3.1 Representation convention

As one of the objective is to propose a unified frame of analysis of the models, they will be depicted using a common formalism, captioned in figure 4.2. The entire evaluation process is split into two or three parts, the model being separated from the scenario generation, and its aggregator if one is necessary. Some varying properties of the models can be seen in the following depictions:

- The topology on which the decision takes place is shown as a black line (e.g. figure 4.4). For models that do not require knowledge of the topology, the line is dotted (e.g. figure 4.3).

- Some models are iterative. We show the intermediate steps from a state at time t to a state at time $t + \Delta t$, but the time loop is not explicitely represented (the state at $t + \Delta t$ replaces the one at t, then links to the one at $t + 2\Delta t$, etc.). Instead, when a previous state is used recursively, it is highlighted in gray in our representation.

- Some models contain some amount of interaction (i.e. the potential at position x_i depends from the potential at position $x_{j \neq i}$). In our depiction, this always results in a vertical step (models with two rows, e.g. figures 4.4, 4.8 and following).

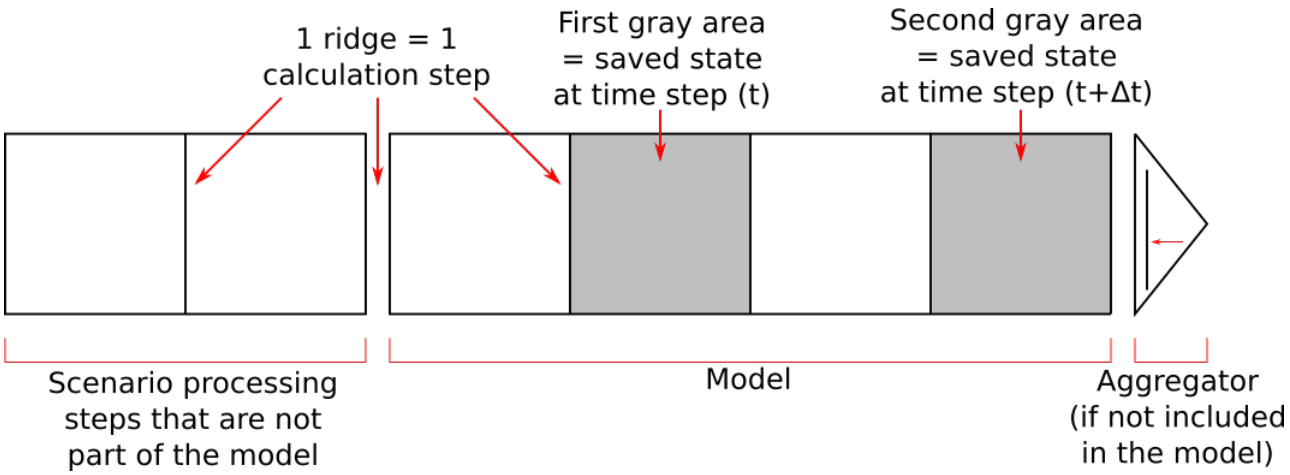


Figure 4.2: Legend for the schematics of the models. Models with recurrent states are shown unfolded, i.e. the process to go from time step t to time step $t + \Delta t$ is visible. The recurrence can be pictured by furling the pattern so that the gray areas touch each other. The aggregator part is shown attached to the model when the latter produces a readable direction directly, and detached if it has been added retrospectively.

4.2.3.2 Logic-based models

Noise integration Models in this family take any information as a truth value. Noise in inputs would be taken as is and not filtered in any way. Most notably, these models would be rendered totally useless in noisy competition tasks: for example, add a bit of white noise to a scenario made of two very similar stimuli, and the output will start switching back-and-forth randomly between the two stimuli.

Winner-takes-all (WTA) This is the simplest model of all. The decision is made at the position of the stimulus of highest intensity. It amounts to applying the mean of maxima aggregator directly on the input (figure 4.3).

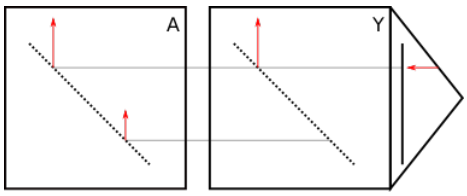


Figure 4.3: Main steps of a WTA model. Explanations in text and figure 4.2.

For WTA to fit in model formalization, we arbitrarily define its activity $\bar{y}(t)$ at position $\bar{x}(t)$ as:

$$\bar{y}(t) = \max_i(a_i(t)) \tag{4.4}$$

Fuzzy logic (FL) As shown in figure 4.4, this model functions in two steps. First, the inputs are fuzzified using a truncated triangular distribution [Dubois et al., 2004], so that they can express a possibility value between 0 and 1, everywhere in the topological space. Then, they are accumulated using a minimax:

$$\begin{cases} y_i(t) = \min\limits_j \Big(\max \Big(1 - a_j(t), P\big(x_i, x_j(t)\big) \Big) \Big) \\ P(x, x') = 1 - \lambda |x - x'| \end{cases} \tag{4.5}$$

where λ specifies the slope around the stimuli.

The decision is then found by using a mean of maxima.

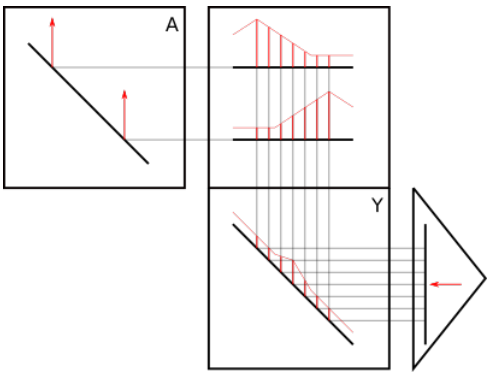


Figure 4.4: Main steps of a FL model

4.2.3.3 Distribution-based models

This class of models operates on interpolations of inputs. Consequently, it is necessary for the inputs to be placed in a topology, as there is no telling that a barycenter of categories $\bar{x}_i$ makes sense.

Noise integration Noise is one side of the estimation. It is not that each presentation contains a certain amount of noise, but instead that each presentation is assumed to vary following a probability distribution that is asserted by the model. In our implementation, inputs are assumed to aggregate into a Gaussian.

Weighted Sum (WS) This model consists of a plain barycenter of the inputs (figure 4.5).

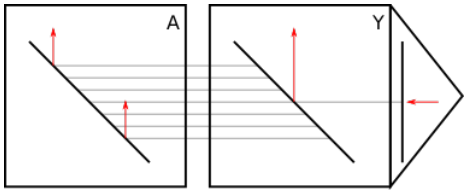


Figure 4.5: Main steps of a WS model

For WS to fit in model formalization, we arbitrarily define its activity $\bar{y}(t)$ at position $\bar{x}(t)$ as:

$$\bar{y}(t) = \sum_i a_i(t) \tag{4.6}$$

Maximum-likelihood estimation (MLE) This is the main paradigm used in multisensory integration [Ernst and Banks, 2002, Rohde et al., 2016]. Given stimuli drawn in Gaussian distributions of estimated position $\bar{x}_i$ and variance σ_i^2, MLE models the decision as a Gaussian distribution of mean m and variance s^2 given by:

$$\begin{cases} m = \sum_i \dfrac{1/\sigma_i^2}{\sum_j 1/\sigma_j^2} \bar{x}_i \\ s^2 = \dfrac{1}{\sum_j 1/\sigma_j^2} \end{cases} \tag{4.7}$$

Our implementation is not directly compatible with this paradigm. Our models are meant to receive individual trials, while MLE operate on a distribution of trials. In particular, we do not use variable input variances σ_i. Oppositely, MLE does not take into

account stimulus intensities a_i. So, for readers interested in what MLE would give in our scenario, we can simulate it using a variable transform $a_i = 1/\sigma_i^2$, making the amplitude a measure of stimulus reliability. Equation (4.7) then becomes:

$$\begin{cases} m = \dfrac{\sum_i a_i \bar{x}_i}{\sum_i a_i} \\ 1/s^2 = \sum\limits_i a_i \end{cases} \tag{4.8}$$

which is exactly the same as our implementation of WS, with $\bar{x}(t) = m$ and $\bar{y}(t) = 1/s^2$ (figure 4.6).

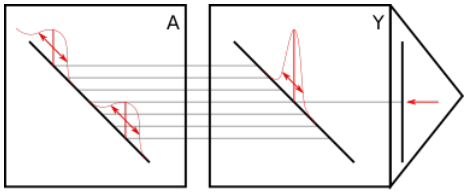


Figure 4.6: Main steps of a MLE model

Kalman filter (KF) This model acts as a time-related MLE: instead of interpolating between two inputs at the same time, it interpolates between a new aggregated input at time $t + \Delta t$ and its older interpolation at time t (figure 4.7). This time, the variance is estimated directly from the inputs in the decision space. Consequently, an unambiguous presentation has less variance (so more weight) than a presentation with two or more stimuli. Also, it is necessary here to assume continuous stimuli, as a sparse input made of a single Dirac would have zero variance, rendering the model quickly useless.

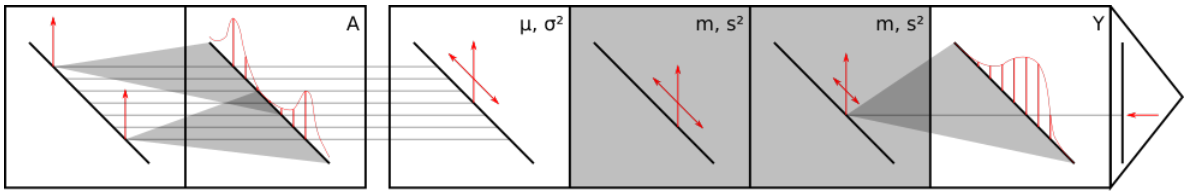


Figure 4.7: Main steps of a KF

The activity takes the form of a Gaussian of mean m and variance s^2. We compute the mean μ and variance σ^2 of the input in order to update the activity:

$$\begin{cases} m(t + \Delta t) = K(t + \Delta t)\, \mu(t + \Delta t) + \big(1 - K(t + \Delta t)\big)\, m(t) \\ s^2(t + \Delta t) = \big(1 - K(t + \Delta t)\big)\, p(t) \end{cases} \tag{4.9}$$

with K the Kalman gain, defined as:

$$K(t + \Delta t) = \frac{p(t)}{p(t) + \sigma^2(t + \Delta t)} \tag{4.10}$$

and p the extrapolated estimate uncertainty:

$$p(t) = s^2(t) + q \tag{4.11}$$

where q is a parameter representing the process noise.

Initial values $m_0 = m(t = 0)$ and $s_0^2 = s^2(t = 0)$ may have an influence on the behavior of this model. In particular, a low s_0^2 will give a strong influence of the prior position m_0

over the incoming inputs. To make this prior knowledge negligible, we set a high initial variance $s_0^2 = 1$ (and $m_0 = 0$ to stay as neutral as possible).

In any case, the model output activity can be represented as a Gaussian of mean $m(t)$ and standard deviation $s(t)$:

$$y_i(t) = \exp\left(-\frac{(x_i(t) - m(t))^2}{2s^2(t)}\right) \tag{4.12}$$

However, the KF does not need an aggregator, as its decision can directly be read as the predicted mean:

$$\bar{x}(t) = m(t) \tag{4.13}$$

4.2.3.4 Accumulators

Inspired from neuroscience, accumulators are a whole family of models consisting of units that accumulate evidence over time [Bogacz et al., 2006, Roxin, 2019]. This section describes the main accumulator models that can be used for two (or more) alternative choice tasks, which do not necessarily take place in a given topology. Each processing unit represents a possible decision, its potential (internal activity) starts as zero and increases gradually as evidence in favor of the decision is brought. The relations between the different models is synthesized in figure 4.13.

Noise integration These models see noise as a part of the decision process. Either added to the sensory inputs as an outcome of background stimulations and sensor imperfections, or embedded as an inevitable side-effect of microscopic neural mechanisms, noise favors bifurcation when a dynamic system is stuck in an unstable equilibrium. For instance, given two distant competitors of similar intensity, a small amount of noise is sufficient to ensure that one is selected over the other. Temporal integration is complementary to the stochasticity, as it permits keeping a random decision stable, contrarily to WTA and FL. For this reason, it is very common to add a supplementary parameter to the implementation of accumulators, which determines the amount of (often white) noise added to all units. This is very different to models such as KF, for which adding white noise to the inputs would cause very little change to the results. Noise integration is at the heart of next chapter, in which this distinction will be discussed further.

Topology As depicted in figure 4.13, DNF is a special kind of accumulator model that relies on and exploits a topology. This makes a big enough difference that it has a separate subsection. On the contrary, models presented in this subsection are meant to process sparse inputs, that may lie in a topology (e.g., left/right) or not (e.g., blue/red). Consequently, the argmax aggregator is the only one that is always suitable for these models. Given the system dynamics, there should be no ambiguity anyway.

Drift-diffusion model (DDM) The DDM is the seminal accumulator model, and a baseline on which other models are based [Bogacz et al., 2006]. Given a stimulus of intensity a_i, the model accumulates an activity y_i over time [Ratcliff and McKoon, 2008]:

$$\tau \frac{\Delta y_i}{\Delta t} = a_i \tag{4.14}$$

This is equivalent to:

$$y_i(t + \Delta t) = y_i(t) + \frac{\Delta t}{\tau} a_i(t + \Delta t) \tag{4.15}$$

although we will keep the first, lighter writing style for all the following models, as it is easier to read.

Our implementation is actually made of several DDM units in parallel. So when multiple (traditionally two) stimuli are put in competition, one way to make a decision is to run one DDM per stimulus and pick the first to have its activity reach a given threshold. This algorithm is called a "Race model" [Bogacz et al., 2006]. In our case, for comparison purposes, we will instead add the argmax aggregator at all times (figure 4.9).

Feed-forward inhibition (FFI) This model (figure 4.8) is designed to put several stimuli in competition. Each accumulator is stimulated by one stimulus and inhibited by all others:

$$\tau \frac{\Delta y_i}{\Delta t} = a_i - w_- \sum_{j \neq i} a_j \tag{4.16}$$

The actual implemented equation is found from (4.16) the same way equation (4.14) is found from (4.15).

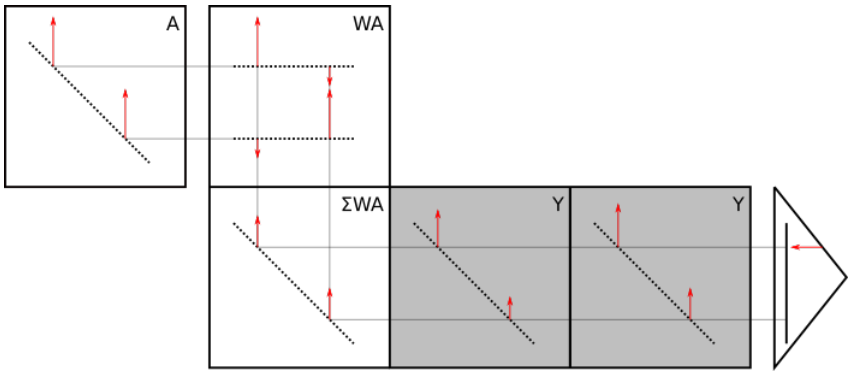


Figure 4.8: Main steps of a FFI

Ohrstein-Uhlenbeck model (OUM) This (figure 4.9) is an upgrade of the DDM with the addition of a leakage term $k > 0$:

$$\tau \frac{\Delta y_i}{\Delta t} = a_i - k y_i \tag{4.17}$$

It allows the accumulator activity to converge when the stimulus amplitude stagnates, contrarily to the previous two models, in which activity may diverge to infinity. All the models that follow include this stabilization term.

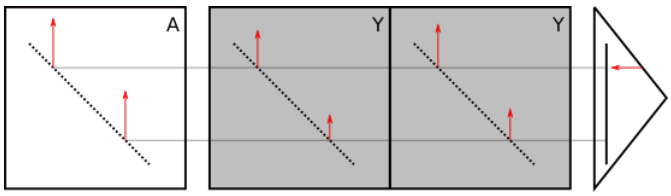


Figure 4.9: Main steps of a DDM or OUM. The difference between the two is that given constant inputs a_i, DDM activity will increase indefinitely, whereas OUM activity should converge to a_i/k due to the leakage term.

Leaky competing accumulator (LCA) The novelty of this model (figure 4.10) is that the activities are put in competition and inhibit each other [Usher and McClelland, 2001]. Also, a term of self-excitation is added:

$$\tau \frac{\Delta y_i}{\Delta t} = a_i - k y_i + w_+ y_i - w_- \sum_{j \neq i} y_j \tag{4.18}$$

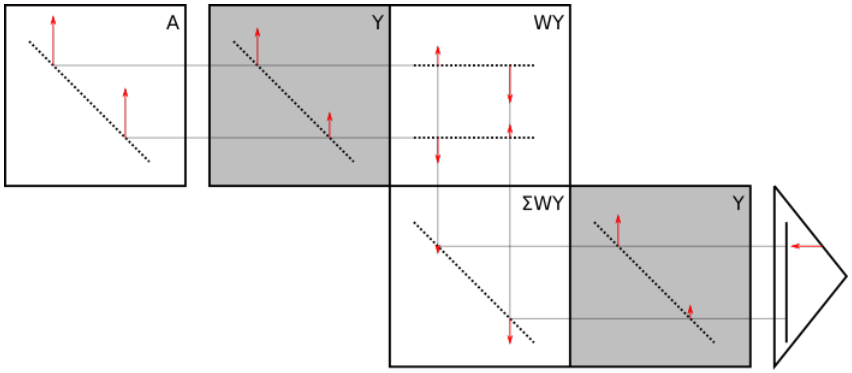


Figure 4.10: Main steps of a LCA

Nonlinear LCA (NLCA) Now (figure 4.11), we differentiate the model activity from its output. The output is obtained by putting the potential through an activation function [Bogacz et al., 2007]:

$$\begin{cases} \tau \frac{\Delta u_i}{\Delta t} = a_i - k u_i + w_+ u_i - w_- \sum_{j \neq i} y_j \\ y_i = f(u_i) \end{cases} \tag{4.19}$$

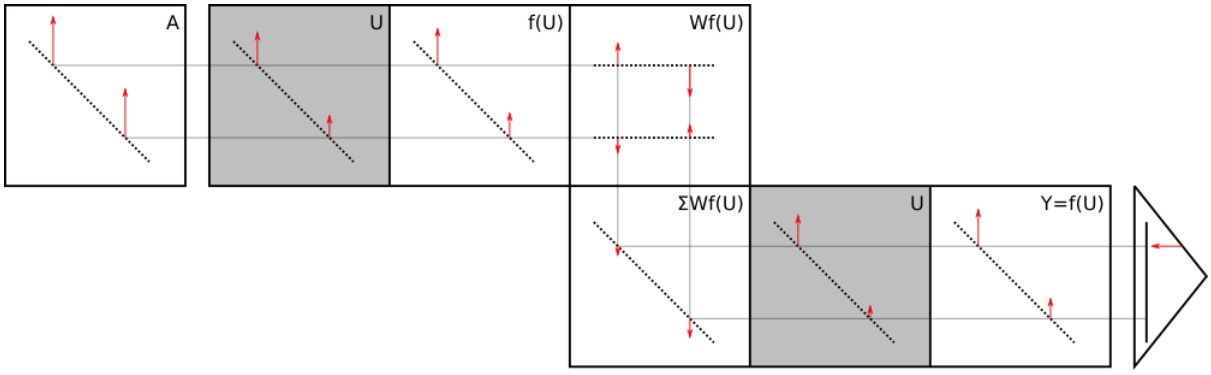


Figure 4.11: Main steps of a NLCA

Pooled inhibition model (PIM) Contrarily, to LCA, in the PIM (figure 4.12), inhibition is shared [Wang, 2002]. A new accumulator is added that gets stimulated by the others and inhibits them all:

$$\begin{cases} \tau \frac{\Delta y_i}{\Delta t} = a_i - k y_i + w_+ y_i - w_- y_I \\ \tau \frac{\Delta y_I}{\Delta t} = -k_I y_I + w_I \sum_j y_j \end{cases} \tag{4.20}$$

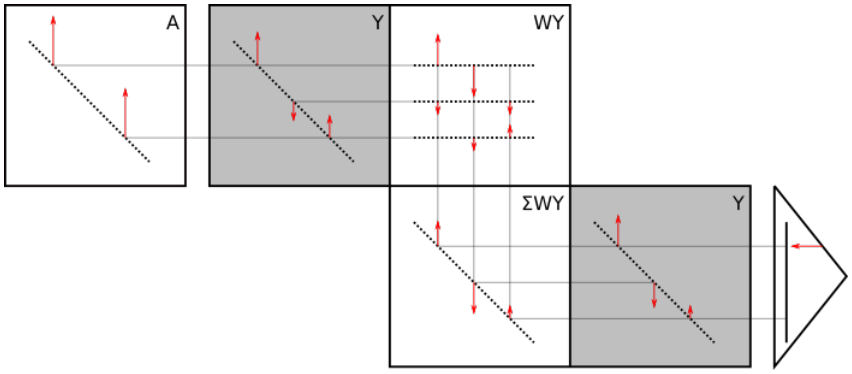


Figure 4.12: Main steps of a PIM

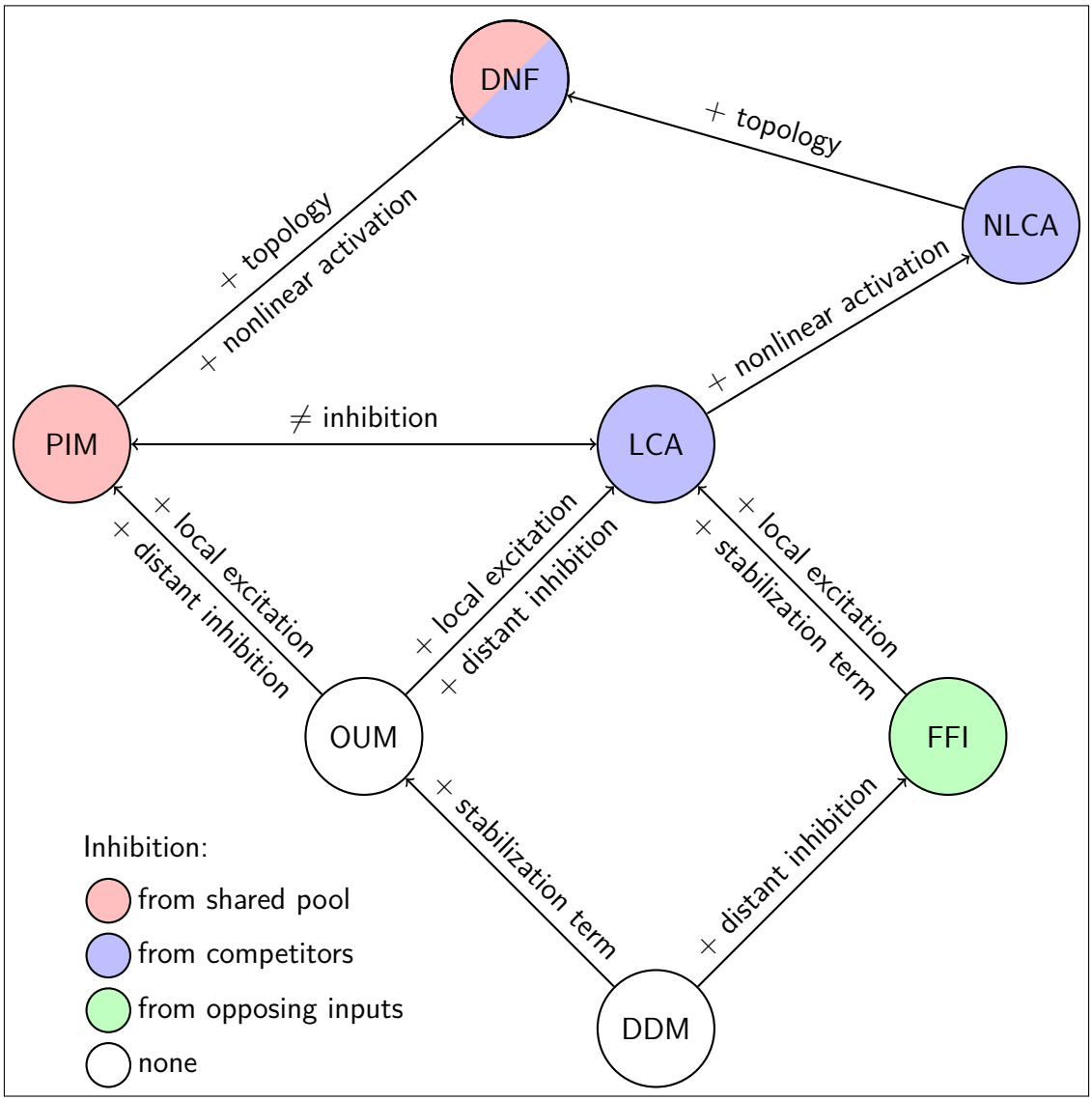


Figure 4.13: Relations between accumulator models. Adapted from [Bogacz et al., 2006] with added DNF.

4.2.3.5 Dynamic neural fields (DNF)

DNF describe the evolution of mean field potential over a continuous domain such as the average membrane potential of neurons on a mesoscopic scale [Trappenberg et al., 2001, Wilimzig et al., 2006]. They can be used to bridge the gap between microscopic-scale neural processes and macroscopic behavioral data [Fix et al., 2011, Taouali et al.,

2015].

DNF originated as a mathematical model of neural dynamics [Wilson and Cowan, 1973, Amari, 1977]. While the first descriptions of membrane potential in neural maps date back to the 1950s, [Wilson and Cowan, 1973] were among the firsts to propose an algorithmic implementation of it. [Amari, 1977] expanded their work by describing in details the behaviors that could emerge from this model of neural dynamics. There are three main categories of behaviors: a monostable field where all excitation eventually dies out; a monostable field where activity increases indefinitely; a bistable field where two different states can be reached depending on the successive presentations received as inputs. To summarize, in a bistable field, once a stimulus is selected, switching focus becomes much harder. This property has made DNF a popular computational model in the study of attention mechanisms [Rougier and Vitay, 2006, Babaie-Janvier and Robinson, 2019]. Extensive analytical studies on the emerging properties of DNF have been made by Gregor Schöner, John Spencer and their teams, and are now condensed in a book [Schöner et al., 2015].

From a computational aspect, DNF (figure 4.14) can be seen as an extension of NLCA to a regularly discretized continuous domain, where each unit acts as an accumulator:

$$\begin{cases} \tau \dfrac{\Delta u_i}{\Delta t} = a_i - k u_i + w_+ \sum_j \exp^{-\frac{\|x_i - x_j\|^2}{\sigma_+^2}} y_j - w_- \sum_j \exp^{-\frac{\|x_i - x_j\|^2}{\sigma_-^2}} y_j + h \\ y_i = f(u_i) \end{cases} \tag{4.21}$$

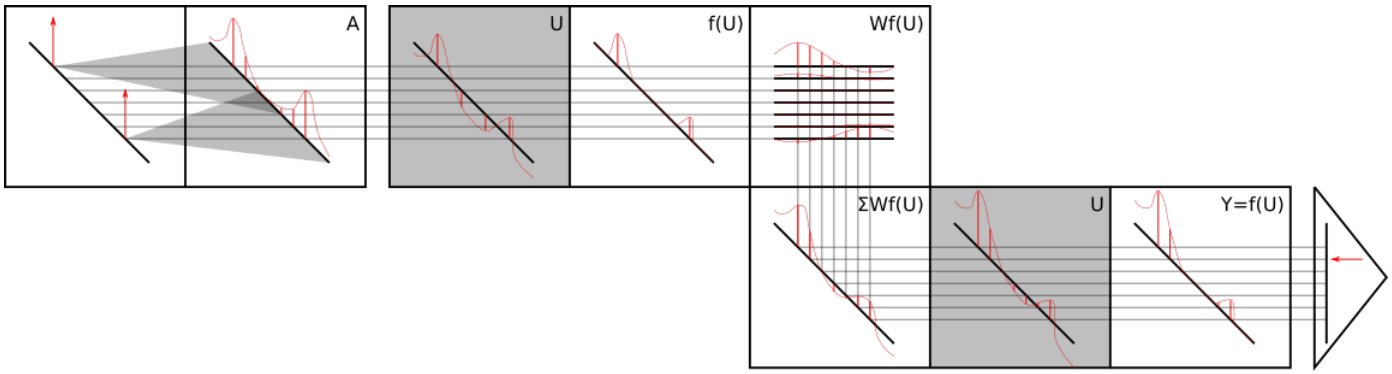


Figure 4.14: Main steps of a DNF

The amount of interaction in the model is determined by parameters w_+, w_-, σ_+ and σ_-. DNF are updated by convoluting their output wit a kernel made of a difference of Gaussians, shaped like a mexican hat: strong close-range excitation and moderate long-range inhibition ($w_+ > w_- > 0$, $\sigma_+ < \sigma_-$) [Amari, 1977]. As a result, close-by units enhance each other while distant ones go in competition, until a stereotypical peak of activity (sometimes called a bubble) emerges. Depending on their parametrization, DNF may achieve various behaviors [Schöner et al., 2015]: selection or interpolation between several conflicting signals [Taouali et al., 2015], robust selective attention in presence of noise and distractors [Fix et al., 2011], working or long term memory of stimuli [Sandamirskaya, 2014]…

The resting level $h \leq 0$ is a parameter that does not appear in all accumulator models. It serves to create an initial resting state with negative potential, so that activity is only produced once strong enough stimuli are received. That parameter can be adjusted easily to filter out low noise or weak stimuli, but it is not always necessary. We maintain it to 0 in this implementation.

In parallel, one common variant of DNF is to integrate all output activity into a global inhibition term (as if $\sigma_- \to +\infty$). This ensures that competition between stimuli

encompasses the entire field (instead of a more or less wide neighbourhood). In that case, DNF can also be seen as a generalization of PIM (see recap figure 4.13).

As the DNF activity converges into (usually one) bubble, a decision can be interpreted from either a WS or WTA of the outputs y_i. The only situation where these aggregators can yield different results is in cases where more than one bubble reach a stable state, and that can be easily prevented with a high enough σ_-.

4.2.3.6 Comparison

A comparison of the main design properties of the models is given in table 4.2. We can already see that depending on the task (stimuli topologically correlated, sparsity of inputs, time relation), some models are more suitable than others. But these models can also be classified according to the level of abstraction at which they compute activity. Figure 4.15 shows how some of these models fit on Marr's hierarchy [Marr, 1982]. For example, models based on Bayesian theory are at the level of computational theory, making assumptions on the distribution of inputs and outputs, and explaining the processing with a theoretical paradigm, with little focus on how the computation is made. Models such as FL, DNF, FFI and NLCA are on the representation–algorithm level, where the operations are explained but the outcome is measured after the fact, and not theorized beforehand. Zooming in on the latter three models, we can look at the units that constitute them and can be likened to sets of DDM or OUM. These can be placed at the hardware level, as they simulate the physical operations that implement decision-making, mimicking actual neurons or cortical columns.

To discern even more the most complex models (FL, KF and DNF), we can differentiate them by the kernel with which inputs are confronted: triangular for FL, Gaussian for KF, a difference of Gaussians for DNF. This competition also does not occur at the same time for all models. For both FL and DNF, inputs are divided, matched to the kernel then put in competition with each other. But for KF, inputs are matched to a Gaussian, all together, then aggregated. For both KF and DNF, the state of the model is gradually updated over time. But for FL, the state is reset at each time step, i.e. there is no memory trace.

Table 4.2: Main characteristics of the models

Model	Use topology	Sparse inputs OK	Time dependency	Interaction	Nonlinearity	Initialization	Main parameters	Aggregation	Noise
WTA	N	Y	N	N	N	–	–	Single output	Ignored
FL	Y	Y	N	Y	Y	–	λ	WTA (MoM)	Ignored
WS/MLE	Y	Y	N	N	N	–	–	Single output	Modeled
KF	Y	N	Y	N	N	m_0, s_0^2	q	Predicted mean	Modeled
DDM	N	Y	Y	N	N	0	τ	WTA	Parametrized, added to input
FFI	N	Y	Y	Y	N	0	τ, w_-	WTA	Parametrized, added to input
OUM	N	Y	Y	N	N	0	τ, k	WTA	Parametrized, added to input
LCA	N	Y	Y	Y	N	0	τ, k, w_-	WTA	Parametrized, added to input
NLCA	N	Y	Y	Y	Y	0	τ, k, w_-	WTA	Parametrized, added to input
PIM	N	Y	Y	Y	N	0	$\tau, k, w_+, w_-, k_I, w_I$	WTA	Parametrized, added to input
DNF	Y	N	Y	Y	Y	h	$\tau, k, w_+, w_-, \sigma_+, \sigma_-$	WTA or WS	Parametrized, added to input

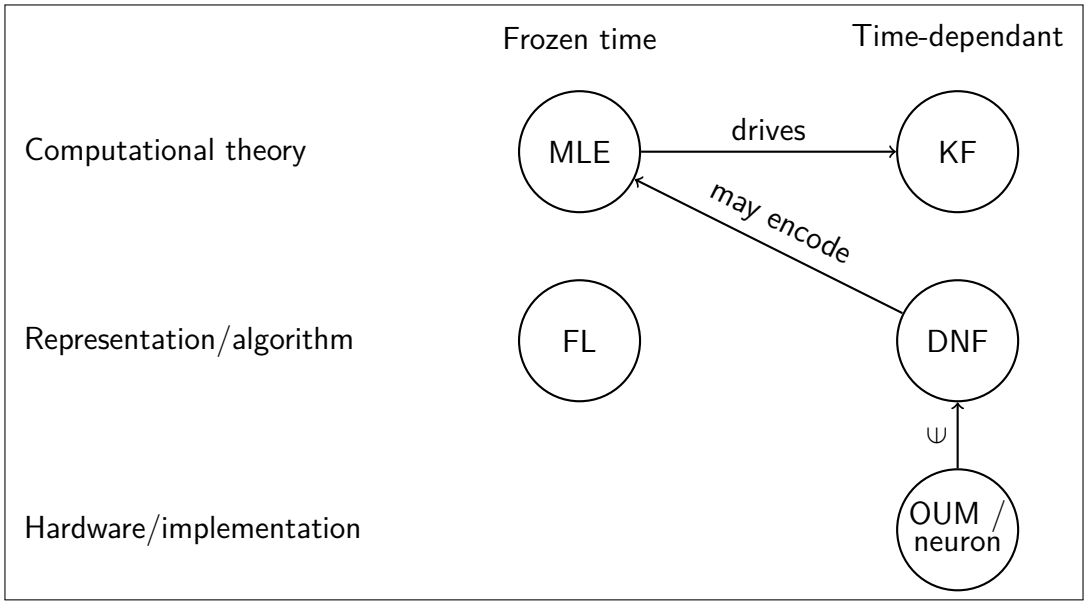


Figure 4.15: Positioning of models in Marr's hierarchy. FFI, LCA, NLCA and PI can be put in the place of DNF, and DDM of OUM, following the relationships described in figure 4.13. Models in the first column operate independently of time: at each time step, an output is given as if time was frozen and inner operations had fully converged. Models in the second column are iterative and may behave differently depending on simulation time step.

All this has repercutions on the way models integrate inputs in space and over time. In order to give a broad overview of the achievable properties, we select a varied sample of all models presented. We pick two model instances from each subsection: WTA and FL, WS and KF, FFI (the simplest model with interaction) and NLCA (the most complete accumulator model outside DNF), and DNF. Since DNF can produce very different behaviors depending on its parametrization, we use two very different set of parameters to give a glimpse of the range of properties available. Initialization is made to zero potential, also setting $h = 0$. We take ReLU as the activation function, and WS as the aggregator. The parameters used for all models are listed in table 4.3. Parameter values were selected through expert knowledge and preliminary tests.

Table 4.3: Parameters of model implementations. Stimuli are placed in an interval $[-2, 2]$ with discretization step $\Delta x = 0.01$. Inputs for KF and DNF are convolved with a Gaussian of amplitude 1 and standard deviation 0.035. Euler integration is applied with time step $\Delta t = 0.01$.

Model	λ	q	τ	k	w_+	w_-	σ_+	σ_-
WTA								
FL	4							
WS/MLE								
KF		0.00005						
FFI			0.01			0.1		
NLCA			0.01	1	0.9	0.25		
DNF1			0.2	1	0.25	0.15	0.05	0.5
DNF2			0.01	1	0.1	0.05	0.15	100

Here is a quick breakdown of the way parameter values are picked. For FL, λ determines how likely stimuli are to interact. With a high λ, they are unlikely to mix and

the model acts closer to a WTA. With a low λ, a midpoint is easily reached and the model acts closer to a WS. We picked a value in-between. For KF, with a high q, internal variance remains high and the model always gives a strong weight to new inputs. q has to be low enough for internal states to matter, but not too low, otherwise new inputs are eventually ignored. For DNF (and FFI/NLCA likewise), we refer the reader to next chapter.

4.3 Results

The evolution over time of activities and decisions of the 8 models in the 8 scenarios (from figure 4.1) is given in figure 4.16. To describe a result, we use the acronym of the model followed by the scenario letter in superscript, e.g. WTA^{E} designates the output of WTA in the fifth scenario. Before we detail the main takeaways, here are some explanations to help understand the figure:

- WTA returns a single position with an activity equal to the maximum intensity. The activity is plotted with a thick line for visualization. When the input is completely empty (beginning of WTA^{E}), the center of the field is returned by default. Same goes for WS^{E}, FFI^{E} and NLCA^{E}.
- As a reminder, FL returns the intersection (minimum) of truncated triangles. In FL^{A}, the triangles overlap slightly in the middle of the field. In FL^{B}, they do not overlap, all that remain are the truncatures: either $1 - 1 = 0$ (fuzzification of the left stimulus) or $1 - 0.99 = 0.01$ (right stimulus). This results in a 0.01 plateau centered on the left stimulus (the only place where it is not truncated to 0). This is why a decision can be made even if no activity is visible.
- KF activity is a Gaussian, where the variance is updated depending on the variance of all inputs projected in a base of Gaussians. The less variance in the input, the thinner the output.
- In $\text{NLCA}^{\text{A/B}}$ and $\text{DNF}^{\text{A/B}}$, we can distinguish two phases. First, peaks appear at the position of each stimulus, of apparent equal activity. After a certain delay, the slight superiority of one stimulus (or group of stimuli) allows one peak to grow stronger, self-excitate more than it is inhibited by the others, and inhibit the others more than they self-excitate. The shift is not visible for NLCA decision (red line) because it is discrete, but since DNF use a barycenter for aggregation, we can see clearly the potential shift from undecided competition to selection ($\text{DNF}_1{}^{\text{A}}$, $\text{DNF}_2{}^{\text{B}}$, $\text{DNF}_1{}^{\text{C}}$, $\text{DNF}_2{}^{\text{C}}$).

Selection and interpolation As we can see in scenarios A and B, some models are specialized in selecting only the strongest stimulus (WTA, FFI, NLCA) and some at making an interpolation (WS, KF). Two can implement both behaviors. FL will either select (FL^{B}) or interpolate (FL^{A}) depending on the proximity between the stimuli. The gap at which it switches behaviors can be controlled through its slope parameter λ. This distinction can also be made with DNF, except it is controlled mostly by the width of lateral excitation σ_+, which determines how close stimuli must be to be able to fit inside a single bubble of activity. But the width of lateral inhibition σ_- is not neglictible. $\text{DNF}_1{}^{\text{B}}$ shows a case where neither selection nor interpolation occurs: the interaction kernel is too thin for the stimulated region to affect each other, so the two stimuli are selected separately. If this particular behavior is unwanted, it is common to use a global inhibition term (i.e. an infinite σ_-).

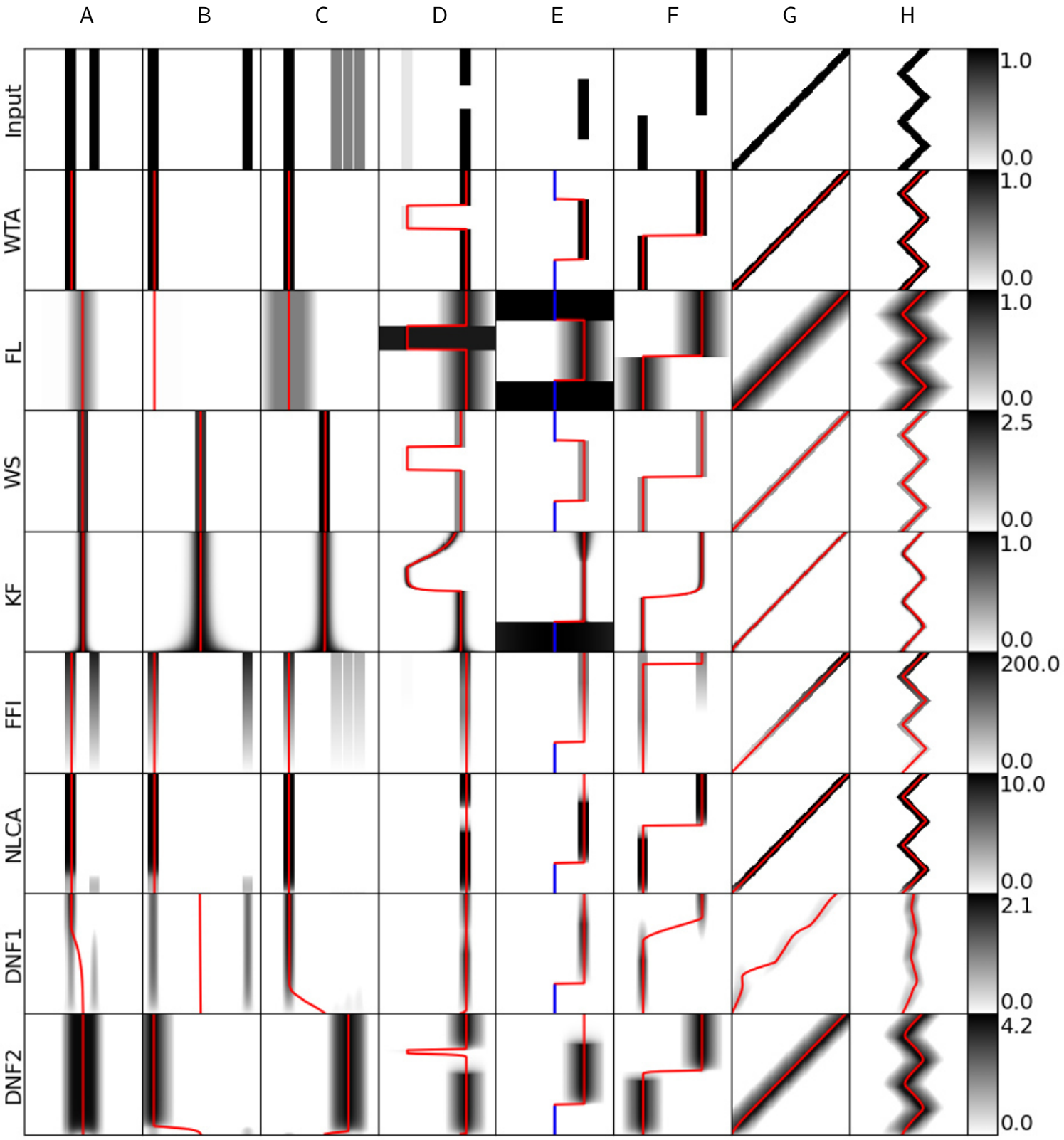


Figure 4.16: Evolution over time (y-axis, starting at bottom) of activities $\{y_i\}$ or $\bar{y}$ (grayscale surfaces) and decisions (red lines) of 8 models (rows 2–9) in 8 scenarios (row 1). Blue segments indicate a default decision (in the middle of the field) when models have an empty or invalid output.

Greediness Scenario C opposes one strong stimulus to a concentrated group of smaller stimuli of higher total intensity. Most models will take the greedy approach and pick the strongest stimulus, as selecting the group requires taking their proximity into account. WS and KF do it to a certain degree, as the bigger weight of the group attracts the barycenter towards it. Regarding DNF, the outcome will again depend on the width of the interaction kernel. DNF favor stimuli that match its kernel. With a thin kernel, the lone stimulus will be picked more easily than the group, in which every component goes in competition with one another. With a large kernel, the group can be merged into a single, big bubble, that prevails over more isolated stimuli.

Robustness to temporary obstruction In scenario D, two targets are present. The question is whether the model will immediately start changing target when the one it initially focused temporarily disappears, or keep the focus for a certain time. Here, there is a clear difference between accumulator-based models and the others. The time constant will ensure that the disappearance of the target will not be integrated instantly. Parameter τ can be tuned to control the update speed: the second instance of DNF, because it has a lower value of τ than the others, starts switching attention before the first target reappears.

Reaction time Scenarios E and F are useful to compare the time dynamics of either KF and all accumulator models. For KF, when a new stimulus appears, the activity will start shifting instantly. For discrete accumulators, the change is taken into account, but there is a delay before the maximum activity changes side. DNF is a mix of both: like KF, the spatial continuity allows for a gradual shift towards the newer stimulus, except a time lag is induced by the temporal dynamics of the differential equation, similarly to FFI and NLCA.

The convergence time of KF can be tuned to a small degree via its parameter q, but it does not give as much leeway as some accumulators and their time constant τ. However, FFI, NLCA and DNF behave differently when a switch occurs between two stimuli. FFI will lower its activity at the first unit and increase its activity at the other, symetrically to what it did before the switch. So the moment the model will actually change targets will depend only on the inputs (here, the delay between input switch and output switch is exactly the same as the duration for which the first stimulus was presented). For NLCA, it will depend mostly on its leakage term k. For DNF, one has to also consider the strength of lateral interactions.

Tracking speed This one is only relevant for topology-based models with a time dependency, i.e. KF^{G}, ${\mathrm{DNF}_1}^{\mathrm{G}}$ and ${\mathrm{DNF}_2}^{\mathrm{G}}$. The others will obviously track instantly a single moving target. Both KF and DNF can track the target with a minor delay. But DNF, depending on its parameters, might fail to track the target smoothly. With a high integration time τ and a small kernel, the simulus might shake off the current active bubble, so a new bubble ends up appearing at the new stimulus position from time to time.

Trajectory smoothing Again, only KF and DNF can smooth a trajectory that changes frequently. Three behaviors can be obtained depending on parameters q and τ respectively: follow the target faithfully, including in sharp turns (high q, low τ); round the turns (low q, medium τ); or converge to a seemingly average position ($q = 0$, very high τ).

A summary of these observations is given in table 4.4. DNF are by far the most versatile, which is consistent with their higher number of parameters. The downside is that

fitting them to achieve a specific task can prove to be difficult (see next chapter). An approximate algorithmic time for each model (not counting input processing or output aggregation) is also given in the table. Unsurprisingly, models with the least amount of steps, WS and WTA, are the fastest. However, this might be quite dependant to the (Python) implementation. For example, FFI is computed by multiplying the vector of inputs by a matrix containing 1 in its diagonal and $-w_-$ elsewhere. Matrix multiplication in `numpy`, Python's standard mathematical library, is slower than other operations, including 1D convolution, which is why FFI seems slower than NLCA or DNF despite its simpler design.

Table 4.4: Summary of model properties

Model	Able to select one stimulus and ignore the rest	Able to interpolate between stimuli	Able to priviledge a group of stimuli	Robust to temporary obstruction	Controllable speed of switch between stimuli	Able to track a target with little delay	Able to smooth trajectories	Suitable for sparse stimuli	Number of parameters	Mean computation time	Precisions
WTA	Y	N	N	N	N	a	N	Y	0	2E-6	a: not relevant
FL	Y	Y	N	N	N	a	N	Y	1	7E-5	
WS	N	Y	b	N	N	a	N	N	0	3E-5	b: depending on weight in the group
KF	N	Y	b	N	c	Y	Y	N	3	1E-4	c: starts instantly but can be slowed
FFI	Y	N	N	Y	N	a	N	Y	2	1E-4	
NLCA	Y	N	N	Y	Y	a	N	Y	3	9E-5	
DNF_1	Y	N	N	Y	Y	N	Y	N	7	1E-4	
DNF_2	Y	Y	Y	N	N	Y	Y	N	7	1E-4	

4.4 Discussion

Decision-making tasks can not all be achieved by a one-size-fits-all model. DNF appear to be the most versatile, failing only with sparse signals in a continuous domain, because it does not suffice to generate a peak of activity, and no interaction occurs. This is not be a very realistic use case, and it can be avoided by "Gaussianizing" the stimuli. On the other hand, their theoretical and computational complexity may not be warranted in every scenario. In competition tasks where topology is not relevant, accumulator models such as NLCA show similar properties to DNF for a lower cost. Finding a trade-off between conflicting stimuli can be done by either KF or FL, the latter being also able to switch between selecting the best (WTA) and interpolating between them (akin to WS) depending on their proximity. For tasks necessitating temporal filtering, KF might come as sufficient.

Models tested here are quite bare, and there is always room for refining and extending them. Parameters can be tuned to change behavior. We show two different examples of it with DNF, but another one would be memory: increasing w_+ sufficiently leads to

self-sufficient peaks to be formed, that stay in place even after the stimulus has disappeared. Furthermore, numerous extensions are available in the literature. WTA can be combined with a kernel to include neighbors in the aggregation. It can also be enhanced with iterative elements (winner takes most). FL has seen various implementations, most notably fuzzy inference systems by [Mamdani and Assilian, 1975] and [Sugeno and Yasukawa, 1993]. WS can be expanded with kernel methods such as SVM. KF has numerous extensions, most notably the extended Kalman filter, one of the most used estimation algorithms for nonlinear systems [Julier and Uhlmann, 2004]. DNF can be adapted to sparse inputs with a variation called sparse neural field [Quinton and Girau, 2010], though it is less robust. It can also be altered to incorporate predictive and active aspects [Quinton and Girau, 2011, Quinton and Goffart, 2018], which reinforce tracking abilities and robustness to distractor and occlusions.

One particular aspect of decision-making that is often overlooked is its relation to perception and cognition. More often than not, decision is more than a posthoc filtering of the model output: the data is already filtered inside the model through thresholding, attentional processes… And like decision drives action, action also impacts decision, through predictive aspects for example. The decision-making algorithm must be put in context of the cognitive system it belongs to. The choice between one strong stimulus and a big group of weaker stimulus, between attending the expected position of a stimulus and exploring unexpected ones, etc., depends on both the task and the system cognition. For instance, ignoring a distractor may be more important when the system is moving or acting towards a previously-selected target, than when it is still figuring out what to do. A decision-making system is often made of several components in perpetual interaction (e.g. extended KF and FL [Das et al., 2017]), and this is how more complex, interesting and robust behaviors may emerge.

So, models presented in this chapter can be seen as a building block of a more complex system. In robotics, when DNF are used, there are always multiple entangled ones. For example, in [Sandamirskaya, 2014], two DNF are used for object perception, one for object picking, and two more for motor control. The general idea is quite similar to different brain regions interacting with each other in any given task. This could be complemented by a learning aspect, mimicking brain plasticity. One could think of taking one model, e.g. a DNF, and tuning its parameters dynamically to adapt behaviors to the current needs. But that would put the burden of solving the task on the learning algorithm alone. What should be tuned dynamically, preferentially, are the mappings and pathways between multiple DNF representing different neural processes. However, this perspective will not be of use in this part of the thesis, as we focus on single multimodal perception tasks.

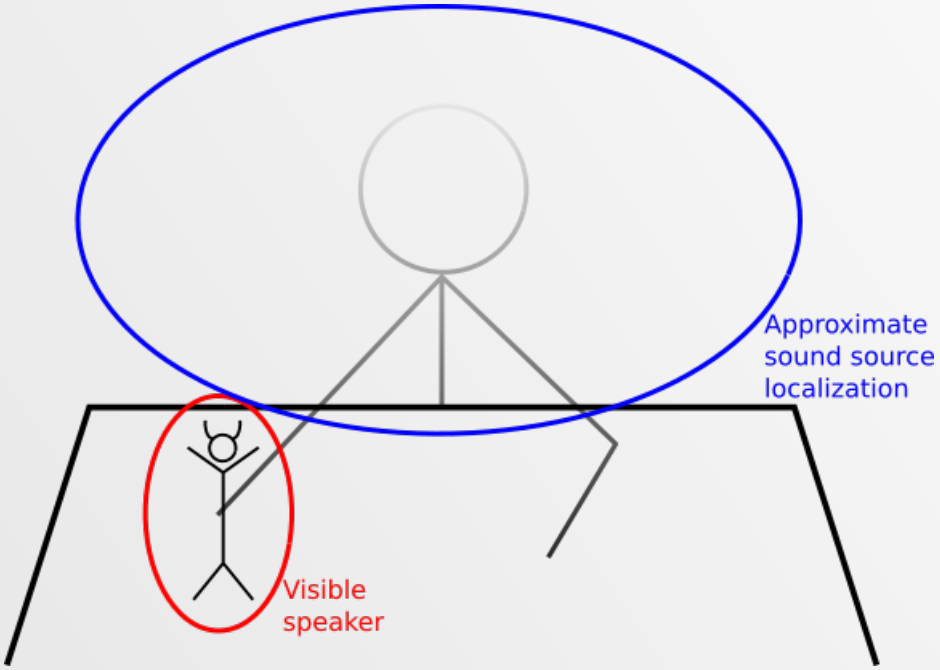
In particular, the nature of sensory modalities plays an important role in decision-making context. Real-life dynamic systems frequently receive multimodal inputs, and there are plenty of ways to merge them into singular, robust percepts [Durrant-Whyte and Henderson, 2016]. As we have seen, some algorithms are designed better for sparse or for topological data. We have also mentioned that noise is processed differently from one model to another. This is not neglectible, as the amount of noise in one modality weighs in on the reliability of the sensor, which in turns influences the weight that is given to it in the fusion. In the next chapter, we highlight the influence of noise in a multimodal merging task, using the ventriloquist effect as a benchmark and DNF as our main paradigm.

Interlude

Back to the story. Eve's conundrum is cut short by the game master, Alice, who invokes the Law of No Group Splitting in Dungeons. The four characters are suddenly attacked by a swarm of enemies coming from all sides, forcing them to flee into the pit where Bob's character disappeared. As they arrive in the basement, the stragglers quickly understand the reason behind Bob's distress. The place looks perfect for an ambush, and indeed, they are quickly greeted by the owner of the dungeon, the final boss, a demon overlord named Ypomni.

Alice brings out an oversized detailed figurine that will represent the enemy on the table, and she starts agitating it as she pronounces Ypomni's evil speech. The players are captivated. It is as if Ypomni themselves were speaking right below their eyes (even though their voice strangely resembles Alice's).

What the players are experiencing is actually a real illusion called the ventriloquist effect. They see perfectly the figurine, the movements of which are well synchronized with the speech. Alice is in the shadows at the end of the table, so her lip movements are not visible. In comparison, the sound source localization is far less precise, so the brain does not rely much on it in this specific occasion.



The same effect explains, for example, how we associate sounds with a person speaking on a television screen. The general theory is that the more reliable a modality is, the more it will capture a multimodal stimulus. In the upcoming chapter, we propose a new computational model of this effect.

Chapter 5

Application: A new computational model of the ventriloquist effect

Contents

The work presented in this chapter was motivated by the context of the project AMPLIFIER, in which this thesis belongs. Our colleagues have been carrying out new psychophysical experiments on the ventriloquist effect, taking into account the effect of saccades in particular. This gave us the initial objective to build a new neuro-inspired model of multimodal merging that was extendable to dynamic tasks. As we faced difficulties in tuning the model to experimental data, we furthered this work with a detailed analysis of parameter effect. The content of this chapter has been published in:
Forest, S., Quinton, J.-C., and Lefort, M. (2022). A dynamic neural field model of multimodal merging: application to the ventriloquist effect. *Neural Computation*, 34(8):1701–1726.

5.1 Introduction

Humans have versatile and diverse ways of perceiving the world around them. Senses provide a dense and continuous flow of data, yet our ability to process information is limited, so we need to select a subset of all available data in order to engage in adequate interactions with the environment. Performing relevant selection involves processes pertaining to (selective) attention.

Focusing on visual attention, human vision is constrained by the heterogeneous disposition of sensors on the retina, with a denser distribution near the center of the visual field (called fovea). As a consequence, humans will tend to gaze at objects of interest, in order to see them better. One outcome of this kind of overt attention is that it may trigger visual saccades towards objects located in the periphery of the retinotopic space. Because of its weaker resolution, saccades are less precise and more likely to be disturbed by artifacts.

That issue can be circumvented with the use of additional information from other modalities [Calvert et al., 2004]. For example, a sound congruent to a visual stimulus may guide saccades to this particular target [Frens et al., 1995, Kapoula and Pain, 2020]. Generally speaking, it is common to merge sensory data coming from multiple modalities. They might enhance each other [Meredith and Stein, 1986], complement one another [Newell et al., 2001], or even compete together to form an interpolation of different sensory inputs [McGurk and MacDonald, 1976, Alais and Burr, 2004]. These mechanisms depend on the relative reliability of the modalities, with factors including stimulus noisiness [Ernst and Banks, 2002], sensor precision [Witten and Knudsen, 2005], and possible top-down interference (such as selective attention; [Driver and Spence, 2004]). Studies on this topic vary from macroscopic (at a behavioral level) to microscopic (neurological) scale, but it is common for such insights to be shared across these two domains [Calvert et al., 2004, Alais et al., 2010].

Our aim is to build a computational model of multisensory integration that can be embedded in attention processes. We will focus on audiovisual merging especially.

5.1.1 Biological inspiration

One source of inspiration for our computational model is the superior colliculus (SC). It has been reported to integrate cues from multiple modalities, including visual, auditory and somatosensory [Wallace and Stein, 1996, Calvert et al., 2004], which makes it a relevant neural structure to be used as a reference for our model. It is also involved in the generation of motor commands such as saccades [Gandhi and Katnani, 2011]. However, please note that our purpose is not to build a biologically-accurate simulation of the SC, but rather get inspiration from the brain workflow, for which mesoscopic scale models of multisensory integration are available. Such scale should allow us to remain neurally plausible, as we later turn our attention to macroscopic observations and directly model behavioral data.

In previous works, the SC has already been used as a target of computational models of visual [Taouali et al., 2015] and multimodal [Casey et al., 2012, Bauer et al., 2015] perception. A common representation of a visual map in the SC is given by [Ottes et al., 1986], where the retinotopic space is mapped to the collicular space using a logpolar transformation. That transformation has been suggested to lie at the core of complex mechanisms of visual attention [Taouali et al., 2015], including saccades [Manfredi et al., 2009].

5.1.2 Computational model

Computational neural models of the SC exist in various forms, both for multisensory integration [Bauer et al., 2015, chapter 3] and for saccade generation [Girard and Berthoz, 2005]. One frequently used theoretical paradigm that encompasses both aspects, and that has been predominant when it comes to visual processing in the SC, is that of dynamic neural fields (DNF) [Marino et al., 2012, Taouali et al., 2015, Quinton and Goffart, 2018]. We have already described DNF in section 4.2.3.5. In this subsection, we make a recap on this model with a special focus on the SC and multimodal merging.

DNF describe the evolution of mean field potential over a continuous domain (usually simply called a map), for instance the average membrane potential of neurons in the intermediate layers of the SC [Trappenberg et al., 2001, Wilimzig et al., 2006]. While interactions at the microscopic scale may be of interest for many neural processes, focusing on neural fields at a mesoscopic scale helps to bridge the gap with behavioral data. This is not only useful to better understand adaptive functions found in living systems [Schöner et al., 2015], but also makes it possible to build artificial systems able to reproduce them (including decision-making and attentional capabilities based on noisy sensor data) and to implement them on robots (with topologies of sensors that differ from humans). Depending on their parametrization, DNF may for instance achieve selection or interpolation between several conflicting signals [Taouali et al., 2015], robust selective attention in presence of noise and distractors [Fix et al., 2011], working or long term memory of stimuli [Sandamirskaya, 2014].

DNF have long been used as models of visual attention [Fix et al., 2011] and (visuo)motor control [Wilimzig et al., 2006, Sandamirskaya, 2014, Quinton and Goffart, 2018]. However, the literature is scarcer when it comes to using DNF for multimodal fusion. [Ménard and Frezza-Buet, 2005] and [Lefort et al., 2013] have built models inspired on cortical maps, with a focus on joint self-organization rather than multimodal stimulus integration. [Schauer and Gross, 2004] have shown promising results with a bio-inspired DNF-based model of audiovisual integration. With very little preprocessing, they achieved a significant response enhancement when exposed to congruent visual and auditory signals, although they did not draw connections to known psychophysical phenomena, like we will.

5.1.3 Psychophysical reference

In this chapter, we will show that applications of DNF go as far as to account for well known psychophysical effects of multisensory integration. As an illustration of such possibilities, we will use the ventriloquist effect [Alais and Burr, 2004], as discussed in section 3.5. To sum it up, from a human participant viewpoint exposed to spatially incongruent visual and audio stimuli, the position of a stimulus is shifted towards the other, depending on which modality has the highest relative precision. The effect takes its name from ventriloquist shows, where spectators have the illusion that a puppet is speaking, while the sound is actually produced by the ventriloquist holding it.

We will draw on psychophysical data reported in [Alais and Burr, 2004], because their experimental paradigm and protocol can easily be replicated in silico, they provide extensive results in all conditions, and their paper is a seminal contribution to the field, with results that have not yet been challenged. One might notice that in their experiment, only the visual precision varied. However, by manipulating the relative precision between the two modalities, they showed the multiple sides of the ventriloquist effect (either vision capturing audition, the reverse, or an interpolation between both). We want our computational model to exhibit the diversity of behaviors linked to multimodal fusion, so this experiment constitutes an interesting showcase.

In addition to empirical data, we will also compare the performance of our model to optimal Bayesian integration, usually considered as the golden standard among formal and computational models of multisensory integration [Ernst and Bulthoff, 2004, Rohde et al., 2016]. However, note that we do not strive for a perfect quantitative fit of our model to the data. Indeed, even though optimization and sensitivity analysis will be combined to assess the ability of our model to robustly converge with behavioral data, our model enables a broad set of perspectives by building on past DNF models, of which the ventriloquist effect is only one illustration.

The remainder of this chapter is structured as follows. In section 5.2, we describe our computational model and its evaluation criteria in the context of the ventriloquist effect. We present the results in section 5.3, and discuss further on the capabilities of our model in section 5.4.

5.2 Method

5.2.1 General model

From a neurophysiological standpoint, the (deep) SC has been reported to receive projections from different modalities on a series of multimodal neural maps [King, 2004]. In this section, we first described how these maps are modeled, before turning to the projections they receive. An overview of our general model is given in figure 5.1.

5.2.1.1 Dynamic neural fields

Our model of a SC map activity is based on dynamic field theory [Schöner et al., 2015]. DNF model the evolution of the neural activity over time on each point of a topological space $\mathbf{X}$ that maps a portion of the brain. The mean field potential U at position $\mathbf{x} \in \mathbf{X}$ and time t is described by the following stochastic integro-differential equation:

$$\tau \frac{\partial U}{\partial t}(\mathbf{x}, t) = -U(\mathbf{x}, t) + I(\mathbf{x}, t) + \int_{\mathbf{x}' \in \mathbf{X}} W(\|\mathbf{x} - \mathbf{x}'\|)\, f(U(\mathbf{x}', t))\, \mathrm{d}\mathbf{x}' + \varepsilon \tag{5.1}$$

where τ is the time constant which determines the response timescale of the entire field, I is the input stimulation over the field and f is a non-linear activation function; as often chosen to simplify numerical simulations, we will use a ReLU function to approximate the mean firing rate of neurons [Quinton and Goffart, 2018]. The last term ε represents noise which, like the entire dynamic neural fields, can be interpreted at either a neurological (a sum of numerous variations of activity induced by external neurons) or psychophysical level (e.g. perceptual noise) ([Schöner et al., 2015], box 1.4, p. 36). Due to the variations being summed over a large population of neurons, white noise is often used, and ε is therefore sampled from a normal distribution $\mathcal{N}(0, \sigma_N)$.

Finally, the kernel approximating lateral interactions within the continuous population of neurons is defined by:

$$W(\Delta\mathbf{x}) = \lambda_+ \exp\left(-\frac{\Delta\mathbf{x}^2}{2\sigma_+^2}\right) - \lambda_- \exp\left(-\frac{\Delta\mathbf{x}^2}{2\sigma_-^2}\right) \tag{5.2}$$

with $\lambda_+ > \lambda_-$ and $\sigma_+ < \sigma_-$, thus giving rise to local excitation and more diffuse inhibition. In the case of visual attention models, with such constraints on parameters, and spatially coherent input stimulation reflecting the presence of localized objects within the visual field, the numerical simulation of the DNF equation will converge to a stereotypical peak of activity, filtering out noise [Fix et al., 2011, Quinton, 2010]. In the case of overt attention, it is then possible to directly project the DNF activity to control eye movements [Quinton

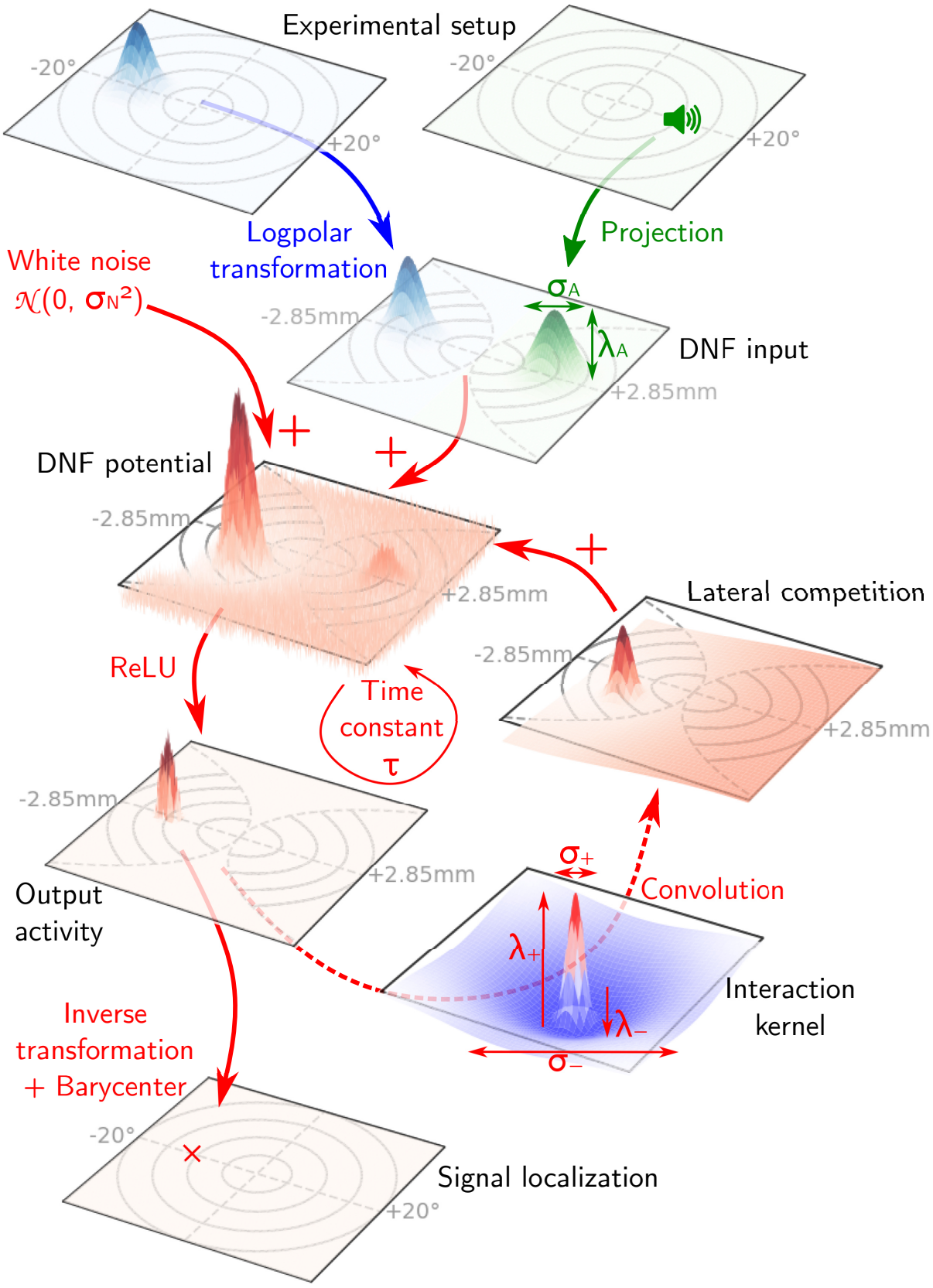


Figure 5.1: Visual representation of the audiovisual merging DNF model. Each rectangle represents a map, either in retinal space (shown with concentric circles) or SC (hourglass shape, obtained by performing a logpolar transformation on the visual map). The blue arrow and text relate to visual preprocessing, green to auditory. Steps and parameters from the model, other than preprocessing, are shown in red.

and Goffart, 2018], in agreement with visual fixations being correlated with a balance of activity in the SC [Gandhi and Katnani, 2011]. In our numerical simulations, we will simply estimate the stimulus position within the field as the barycenter of the field output $f(U)$ [Rougier, 2006].

The time course of field activity before convergence will not be the focus of this chapter, since we are mostly interested in the location of peaks after stabilization. Readers interested in activity evolution over time will find extensive insights in [Schöner et al., 2015] and an illustration of SC dynamics simulation in ([Taouali et al., 2015], figure 5).

5.2.1.2 Projections to the neural field

Empirical evidence supports that signals emanating from a common location in the environment, even through different modalities, will project to nearby locations in the SC [Wallace and Stein, 1996]. At the same time, the structure of the SC can be linked back to retinotopic space [Ottes et al., 1986]. Given these neurophysiological findings, we decompose the input I defined at each point of the DNF as the sum of a visual input I_V and an auditory input I_A. Although summing projections from different modalities introduces a strong assumption into the model, it is frequent in the literature [Sandamirskaya, 2014, Schöner et al., 2015].

The projection of visual stimuli from the retina to the SC has been modeled mathematically in the form of a logpolar transformation [Ottes et al., 1986]. Formally, a visual signal at a position (u, v) in the retinotopic space will be mapped to the SC at a position $\mathbf{x} = (x, y)$ given by:

$$\begin{cases} x = B_x \log\left(\dfrac{\sqrt{(u+A)^2+v^2}}{A}\right) \\ y = B_y \arctan\left(\dfrac{v}{u+A}\right) \end{cases} \tag{5.3}$$

A, B_x and B_y are constant parameters that originate from the literature [Ottes et al., 1986]. Their values are given in table 5.1.

As for the auditory inputs and to our knowledge, there is no mathematical formulation of their projection onto the SC. To avoid introducing additional model parameters or uninformed constraints, we thus simply aligned the audio stimuli to their spatially congruent visual counterparts, since we do not aim at modeling the learning of sensory maps in the current research work. As projections to the SC through complex neural pathways are usually quite distant from raw sensory stimulation, we generate population coded auditory inputs as gaussian blobs of amplitude λ_A and width (standard deviation) σ_A. While the gaussian blob associated to the auditory stimulation is directly projected without distortion to the SC neural map, a similar gaussian blob is generated for the visual stimulation yet transformed through equation (5.3) during its projection on the SC. Amplitude and width of the audio stimuli are added to the list of free parameters of the model, while visual amplitude is fixed (since redundant with λ_A) and visual width is driven by the experimental setup.

5.2.2 Application to the ventriloquist effect

Even with constraints imposed on projections to the DNF, the model of the SC presented in the previous section and recapped in figure 5.1 is designed to accomplish a variety of tasks related to audio-visual perception, attention or memory, building upon existing works on neural fields [Schauer and Gross, 2004, Sandamirskaya, 2014, Taouali et al., 2015]. In order to validate its capabilities for multimodal fusion, we here apply and test this generic model using an experimental paradigm associated with the ventriloquist

effect, this effect being largely documented, and human data available. We use the seminal work by [Alais and Burr, 2004], using human performance as ground truth for the evaluation of audio-visual fusion in our model. In their article, they reported detailed psychophysical results aggregated over hundreds of trials per condition and participant, with psychometric functions estimated in both unimodal and bimodal blocks of trials. For the latter, they relied on a fully crossed experimental design, manipulating various fusion-relevant parameters of the stimuli. Among other things, this makes their study particularly fit to replication using their data as a ground truth for computer simulations.

5.2.2.1 Experimental data

For each bimodal trial, participants were exposed to a sequence of two presentations of audio-visual stimuli (conflicting and non-conflicting, in random order), and had to report which of them was perceived more leftward. In the non-conflict presentation, auditory information (1.5 ms sound click with position determined by the interaural time difference) and visual information (15 ms low-contrast Gaussian blob of controlled width, with standard deviation $\sigma_V \in \{2°, 16°, 32°\}$) were perfectly aligned with each other, but their eccentricity relative to the center of the participant's field of view was manipulated (from $-20°$ to $+20°$, as depicted on the horizontal axis of figure 1 of [Alais and Burr, 2004]). In the conflict presentation, stimuli were still aligned on the azimuthal axis, but an horizontal spatial discrepancy was introduced between the two, with the visual stimulus moving of $\Delta \in \{-5°, -2.5°, 0°, 2.5°, 5°\}$ (from left to right) and the auditory stimulus moving of $-\Delta$ (horizontal positions in figure 5.2).

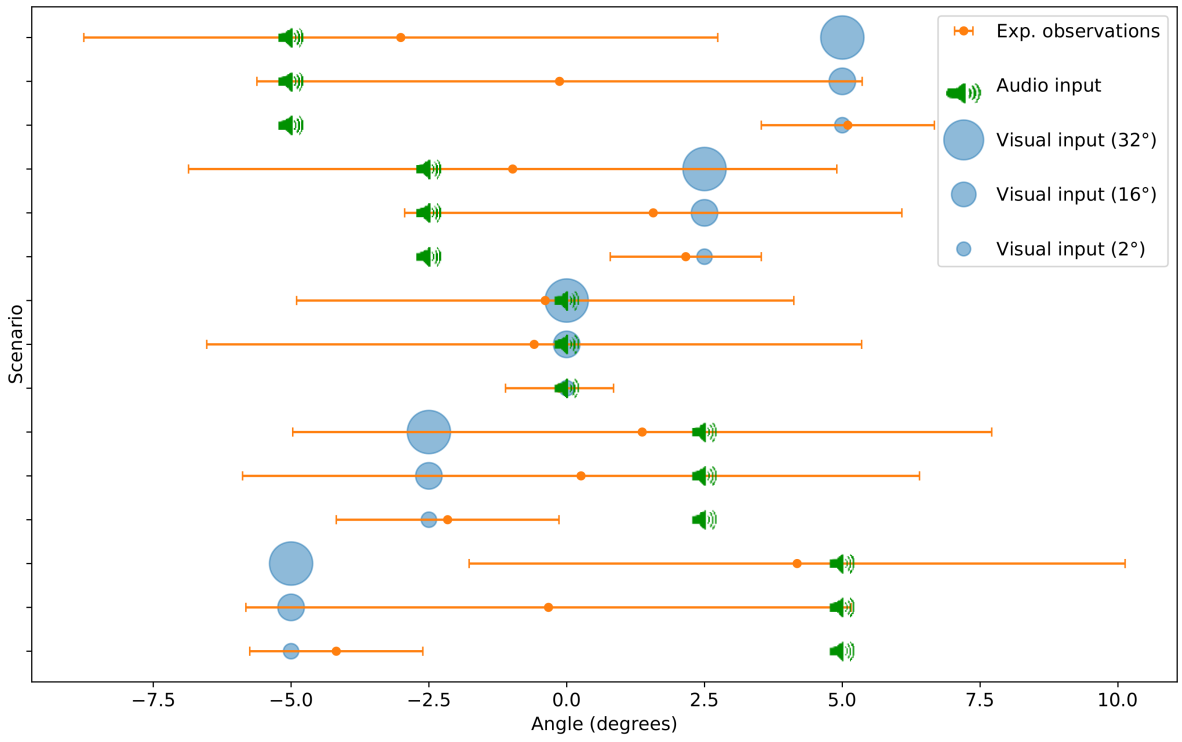


Figure 5.2: List of scenarios and experimental measures from [Alais and Burr, 2004]. In each line: The green speaker symbol gives the position of the auditory stimulus in the conflicting presentation. The blue circle of growing size gives the position of the visual stimulus, of width $\sigma_V = 2°$, 16° or 32° (not to scale). The measures of bimodal localization are represented by an orange error bar (mean ± SD).

As a consequence, we aim at replicating the psychometric curves (proportion of con-

flict stimuli perceived rightward as a function of eccentricity of the non-conflict stimuli) obtained in the 15 scenarios of the original study (3 visual precisions $\times$ 5 spatial distances). These psychometric curves were approximated by cumulative Gaussian functions (sigmoids with near-logistic shape; [Bowling et al., 2009]), thus reducing them to two parameters: median (also named point of subjective equality, equal to the mean for a Gaussian distribution) and standard deviation (accuracy). The Gaussian distributions associated to the unimodal and bimodal psychometric functions from [Alais and Burr, 2004] are reproduced on figure 5.2.

As a synthesis of their results, a thin visual stimulis ($\sigma_V = 2°$) captures the location of the merged signal given its high accuracy. When it is very wide (32°), the auditory stimulus does. In-between (16°), the merging is located between both. In addition, the higher the precision of the inputs (e.g. 2° visual stimulus), the lower the standard deviation of the human localization distribution after fusion, reflecting that auditory and visual information were taken into account in a statistically optimal manner [Rohde et al., 2016].

5.2.2.2 Model constraints and simulation

For this specific operationalization of the ventriloquist effect, all presentations happen on a single azimuthal axis: $y = 0$. While the version of our DNF model presented in section 5.2.1.1 could be used as a suitable model of two-dimensional maps in the SC, it introduces parameters that are not directly supported by empirical data from the selected study, and would simply make optimization and interpretation more complex. Committing to the principle of parcimony, we have therefore chosen to restrict our model to a unidimensional projection of the SC, reducing the computational cost of the simulations.

Whereas asking which stimuli were perceived as more leftward made sense experimentally to reduce task difficulty and prevent biases in responses, numerical simulations allow to directly estimate localization probability density functions. Yet given the noise and non-linearities from equation (5.1), we rely on the Monte Carlo method to sample the localization distribution under each condition through repeated simulation, and estimate summary statistics (mean and standard deviation of the empirical Gaussian distribution) for the conflict presentation alone. This means that the (static) inputs used in our model always consist of a bimodal signal, having a median location set at the fovea, and made of two unimodal components located opposite from each other. The non-conflict presentation is no longer necessary in this numerical setting. Since there is no generic analytical solution to this class of stochastic integro-differential equations, we rely on numerical resolution, which makes simulations computationally intensive and parameter estimation complex.

To correctly model the spatial distribution of stimuli used in the ventriloquist experiment, the simulated neural field covers angles from $-20°$ to $20°$ in retinal space (which, after the transformation of equation (5.3), corresponds to $\pm 2.85\,\text{mm}$ in SC) with a spatial resolution of 100 points ($\Delta x = 0.057\,\text{mm}$). Similarly, to ensure a correct approximation of the temporal dynamics of the multimodal fusion and guarantee convergence to a stable localization, we solve equation (5.1) using the Euler scheme with a temporal resolution of 100 iterations per second ($\Delta t = 0.01\,\text{s}$). All simulation constants are recapitulated in table 5.1. Algorithmically, the mean field potential (vector U) is initialized to zero and

updated by applying the following equation:

$$\begin{aligned}\forall k \in K, U(k\Delta x, t+\Delta t) = U(k\Delta x, t) \\ + \frac{\Delta t}{\tau}\Big(- U(k\Delta x, t) \\ + I(k\Delta x, t) \\ + \sum_{k' \in K} W\big(|k\Delta x - k'\Delta x|\big)\, f\big(U(k'\Delta x, t)\big) \\ + \varepsilon \Big)\end{aligned} \tag{5.4}$$

where $K = \{-50, -49, \ldots, 50\} = \{\frac{-2.85}{\Delta x}, \frac{-2.85+\Delta x}{\Delta x}, \ldots, \frac{+2.85}{\Delta x}\}$ and I can be decomposed according to section 5.2.1.2:

$$I(k\Delta x, t) = I_V(k\Delta x, t) + I_A(k\Delta x, t) \tag{5.5}$$

Table 5.1: Constant settings for all simulations. The values and descriptions of A, B_x and B_y are taken from [Ottes et al., 1986]. High spatial and temporal resolutions were chosen to prevent any qualitative impact on the results.

Constant	Value	Unit	Description
B_x	1.4	mm	x-axis scaling for the SC map
B_y	1.8	mm/°	y-axis scaling for the SC map
A	3	°	Shape of the mapping, relatively to $\frac{B_x}{B_y}$
Δt	0.01	s	Simulation time step
$\mathbf{X}$	[-2.85, 2.85]	mm	Spatial domain in SC
Δx	0.057	mm	Spatial discretization step

Given that we model a forced decision task (i.e. where human participants were asked to always answer even if they needed to guess), adequate parameters should always lead to the (quick) emergence of a stable activity pattern in presence of stimuli, usually under the form of a stereotyped peak of activity on the neural field. For an illustration, see results in section 4.3, and in particular $\text{DNF}_1{}^{\text{A}}$, $\text{DNF}_2{}^{\text{A}}$ and $\text{DNF}_2{}^{\text{B}}$ in figure 4.16. For this chapter, we will make sure that DNF parameters do not allow a double selection like in scenario $\text{DNF}_1{}^{\text{B}}$.

We can see that, given two similar but conflicting stimuli, the DNF will generate a prototypical peak of activity (an attractor in the dynamical system modelled by the set of differential equations), from which the barycenter can be used as the bimodal stimulus localization estimate, as developed at the end of section 5.2.1.1. The ensuing decision will either correspond to an interpolation between unimodal signals, or to the selection of the strongest one (barring random fluctuations not shown here). The choice between these two behaviors will depend on both the distance between the stimuli (as in section 4.3) and their relative precision (illustrated in the following result section, with much lower stimuli precision).

5.2.3 Evaluation

While our task is not limited to a quantitative fit to empirical data, we will use the differences between model outputs and psychophysical results as a performance metric, which allows an indirect comparison of numerical models using human behavior as ground truth. As all (human and simulated) localization distributions roughly follow a Gaussian profile, performance will be computed based on estimated means and standard deviations on all scenarios from figure 5.2.

5.2.3.1 Compared models

The seminal experimental results on which we rely were already accompanied by a MLE model [Alais and Burr, 2004]. It remains the dominant paradigm for multisensory integration [Rohde et al., 2016], to which we will compare. It explicitely relies on the hypothesis that the psychometric functions of visual and auditory stimuli are Gaussian cumulative distribution functions. The mean estimate and derived variance for their Bayes optimal combination are given by:

$$\hat{S}_{AV} = \frac{1/\sigma_V^2}{1/\sigma_V^2 + 1/\sigma_A^2}\hat{S}_V + \frac{1/\sigma_A^2}{1/\sigma_V^2 + 1/\sigma_A^2}\hat{S}_A \tag{5.6}$$

$$\sigma_{AV}^2 = \frac{\sigma_V^2\sigma_A^2}{\sigma_V^2 + \sigma_A^2} \tag{5.7}$$

where $\hat{S}_V$ and $\hat{S}_A$ are the mean estimates of the visual and auditory signals positions respectively (assumed to coincide with the actual position of the sources), and σ_V^2 and σ_A^2 their variances (derived from the unimodal psychometric functions, as described in [Rohde et al., 2016]). The Bayesian model differs by design from ours, insofar that it uses the unimodal performance to predict the bimodal behavior, whereas we fit our model directly on the bimodal scenarios, without prior knowledge of the unimodal variances.

In the case of our DNF model, for a given set of parameters allowing convergence to a stable localization decision through numerical resolution, each simulation should generate a single scalar output (between $-20°$ and $20°$ after projecting back to the visual space). By replicating such simulations, the Monte Carlo method therefore produces an approximate localization distribution in each condition. As the 15 generated distributions (one per condition) are expected to be roughly Gaussian and were tested against extreme observations (to prevent biaises in mean and standard deviation estimates due to statistical outliers), 50 simulations per condition were assessed as sufficient to extract accurate distribution parameters, and used as indices of model performance.

To test the usefulness of the logpolar transformation to correctly explain the experimental results for different eccentricities (confounded with varying degrees of audiovisual discrepencies), as well as to test the robustness of the DNF model to distortions in inputs projections, we will use two versions of our model: one where visual inputs go through a logpolar transformation following equation (5.3) (referenced as DNF+log in tables and figures); another where the transformation is replaced by an identity function (DNF+id), meaning $x = u$ and $y = v$. In the latter case, the DNF will operate directly on a visual map, i.e. $\mathbf{X} = [-20°, 20°]$, $\Delta x = 0.2°$, and the auditory inputs need no realignment.

5.2.3.2 Model parametrization

Following previous definitions and constraints, our model has eight free parameters (see table 5.2): six from the DNF equation, and two from our modeling of auditory inputs in the SC as a Gaussian blob. This is true for both versions (DNF+log and DNF+id), since the logpolar transformation parameters are constant and derived from the literature. The behavior of a DNF depends mostly on the shape of its interaction kernel W. Therefore, fusion performance can mainly be correlated to the four parameters λ_+, λ_-, σ_+ and σ_-. The dynamic and nonlinear nature of the DNF equation can make the dependencies very hard to comprehend, with strong interactions between parameters, especially when related to the kernel. Since we will also measure the variance of the model localization output, σ_N, which controls the amount of noise in the equation, will also play an important role; as well as τ, which controls the integration rate, and thus the weight of the noise compared to stimuli. Finally, while λ_A and σ_A do not intervene in the inner dynamics of the DNF, they

can also be tweaked as part of the audio preprocessing of the model. They do have some interaction with the other parameters, as the shape of the interaction kernel determines which shape of input signals will be favored.

To ensure a fair comparison of models, free parameters had to be adjusted to the multimodal merging task. Within the high-dimensional parameter space, meta-heuristics that were already applied to the optimization of DNF parameters (such as [Quinton, 2010]) did not prove to be robust enough in the case of our multimodal fusion scenarios and evaluation procedure. Indeed, we could not easily combine into a single optimization criteria our two metrics: mean multimodal localization and localization variance. Trying to tackle this multicriteria optimization problem on stochastic integro-differential equations also did not lead to acceptable Pareto-optimal sets of solutions.

Therefore, after a review of articles in the DNF literature, and extended preliminary simulations, we extracted for each parameter an interval in which suitable behavior was possible, and simply relied on an iterative and partial grid-search approach. Similarly to [Jenkins et al., 2021], we started by picking some expertise-driven parameter values, then analyzed model performance as a function of one or two parameters at a time. Keeping the best values found, we iterated over sets of parameters until convergence. In a way similar to a simplex algorithm, we obtained the parameter values in column "Selected" of table 5.2. We have found that a change in σ_A was sufficient at first sight to compensate most of the distortion of visual inputs by the logpolar transformation. Consequently, it is possible to switch between DNF+log and DNF+id and obtain results of the same order of magnitude, by tweaking σ_A and leaving other parameters intact.

Table 5.2: Model parameters. When one is fixed, its value is given in the "Selected" column. When one varies, either for exploration or visualization, it takes its values in the specified interval, discretized uniformly into 20 values. For DNF+log, values in italics have to be rescaled by a factor $\frac{2.85}{20}$ to accommodate for the change in field size from $[-20, 20]$ degrees to $[-2.85, 2.85]$ millimeters: while the transformation in the model is not linear, we use this field-wide rescaling to express all width and SD values in the same unit, opting for degrees. After the input is transformed, the DNF always operates on a regular space. σ_A has two different values for DNF+id and DNF+log respectively.

Parameter	Min.	Max.	Selected	Description
τ	0.05	0.5	0.15	Time constant
λ_+	0.1	1	0.425	Amplitude of lateral excitation
λ_-	0.05	0.2	0.15	Amplitude of lateral inhibition
σ_+	*0.2*	*2*	*0.85*	Width of lateral excitation
σ_-	*2*	*100*	*40*	Width of lateral inhibition
σ_N	0.5	5	2.8	Standard deviation of noise distribution
λ_A	0.1	2	1.1	Amplitude of auditory input
σ_A	*2*	*64*	20 | *26*	Standard deviation of auditory input

5.3 Results

Relying on the (locally) optimal parameters from table 5.2, this section first shows qualitative and illustrative behaviors of the DNF, before comparing performance between the different models described in section 5.2.3.1 (Bayesian, DNF+id, DNF+log), and then turning to a sensitivity analysis of the DNF model performance, studying the impact of pairs of parameters when keeping the others fixed. The objectives are to show

that good performance from either DNF model versions cannot be attributed to over-parametrization (and thus overfit to the experimental data), and to study the effect of parameters on the DNF behavior.

5.3.1 Evolution of field potential

As a way to showcase the behaviors of our models, we start by observing their dynamics in realistic experimental conditions, complementing the illustration of qualitative differences in DNF outputs based on stimuli distance in section 5.2.2.2. For this subsection, we will make tests using the DNF+id model, as its output can be directly read and easily interpreted in the topological space of the source stimuli. We use the parameters from the "Selected" column of table 5.2. The inputs in the second experimental scenario ($\Delta = -5°$, $\sigma_V = 16°$) and related model activity are given in figure 5.3.

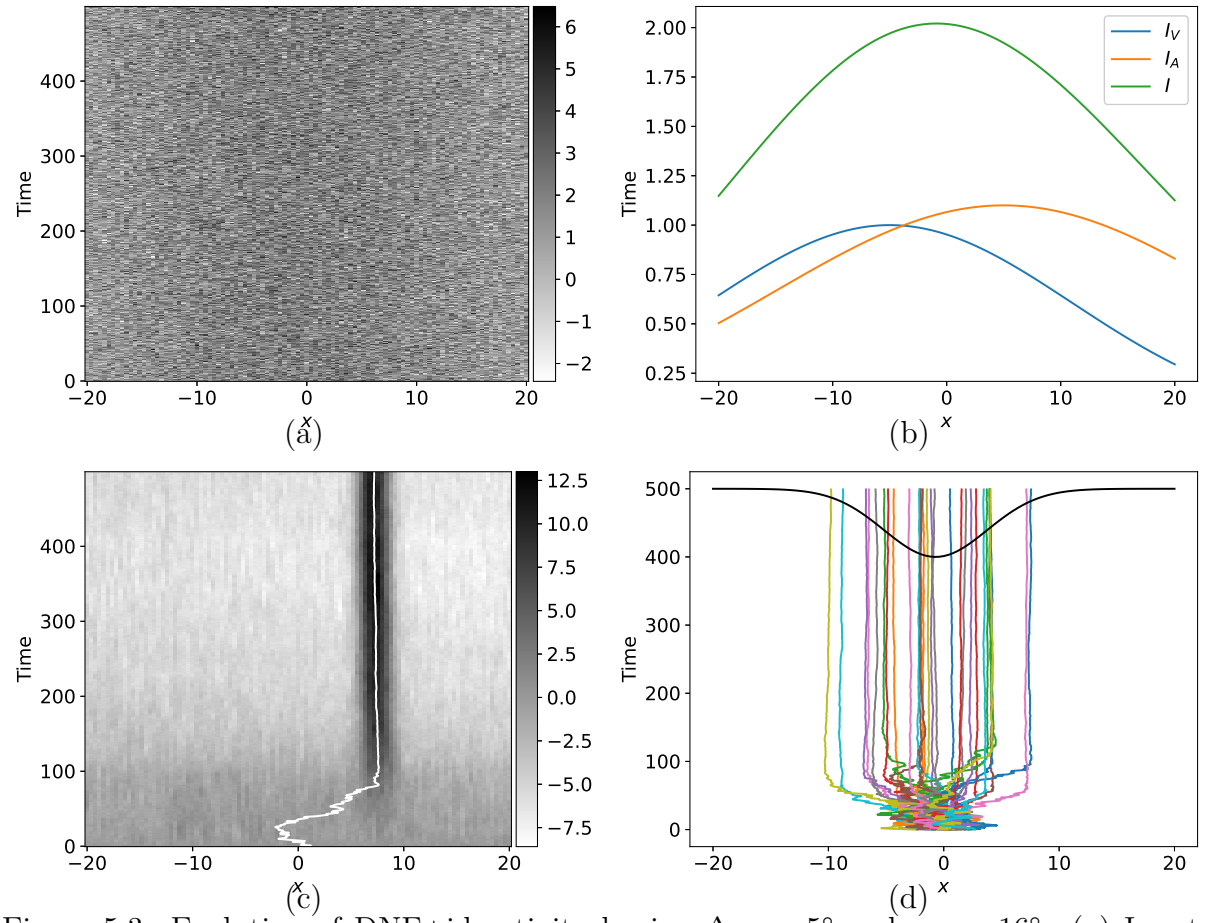


Figure 5.3: Evolution of DNF+id activity having $\Delta = -5°$ and $\sigma_V = 16°$. (a) Inputs summed with noise on neural field (x) over time. (b) Theoretical distribution of inputs in absence of noise. (c) Field potential U during one single run. The white line shows the evolution of the barycenter of field output $f(U)$. (d) Barycenters of DNF output for 30 other runs of the model. The black line shows the approximate Gaussian distribution obtained with the mean and SD of the final 30 positions.

As can be seen in subfigure (a), the amount of noise in the simulated data makes it almost impossible to distinguish the raw stimuli (b) with the naked eye. The evolution of DNF potential U is shown for one run of the model in subfigure (c). A peak forms at a seemingly random position, which is actually biased by the position of the stimuli. The underlying distribution of selected multimodal locations becomes apparent when the model is run multiple times (d). Some decisions do happen quite far from the source, which is consistent with stereotypical psychophysical studies, in which participants sometimes

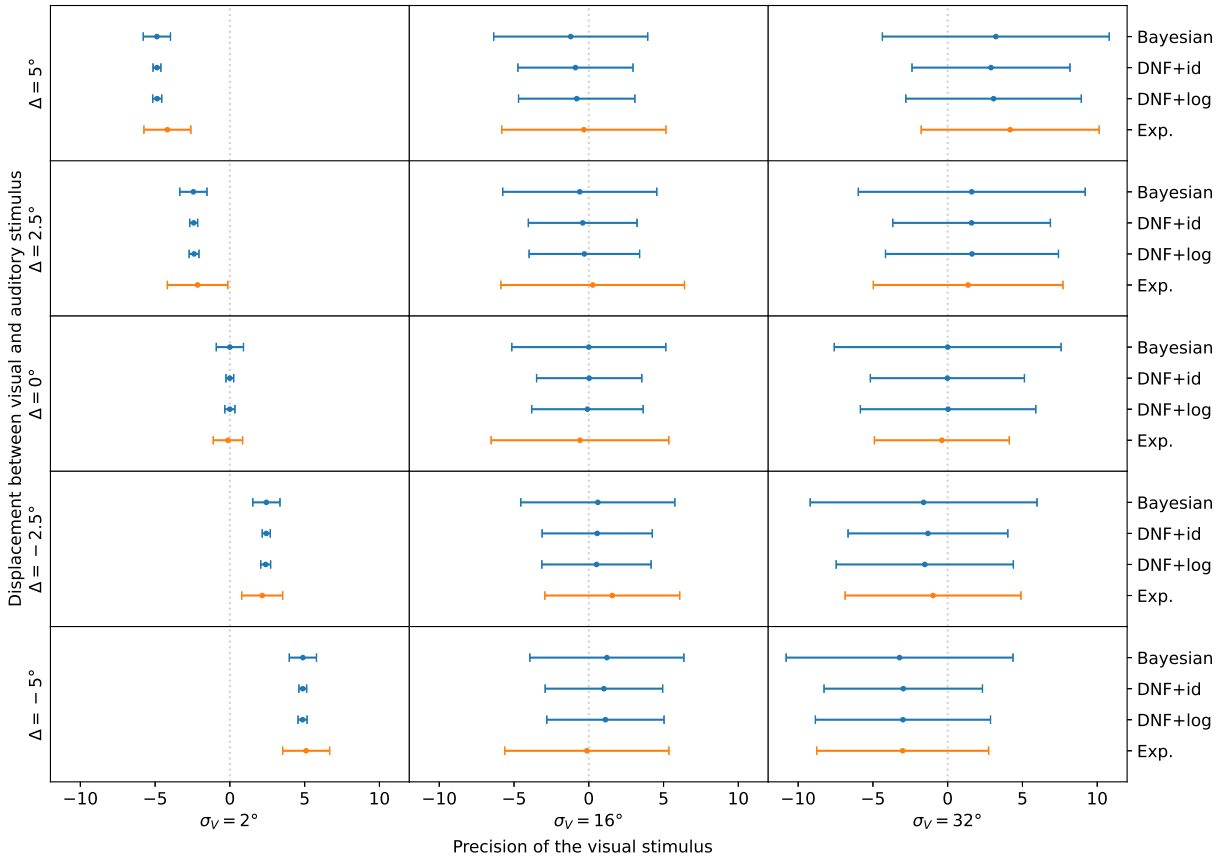


Figure 5.4: Experimental results of bimodal presentation (orange intervals, same as figure 5.2) and corresponding model outputs (in blue). For each error bar, the center dot represents the average localization, and the half-amplitude is the standard deviation.

realize extreme guesses. But the distribution of selected multimodal locations shows that on average, decisions are made in between the two stimuli. The mean and variance of this DNF output distribution are the summary statistics used for model evaluation.

5.3.2 Model evaluation

Given the aforementioned models, we simulated the experimental scenarios to compare with the psychophysical data. The results are summarized on figure 5.4. As a reminder, we observe two metrics: the mean localization of a bimodal presentation (center of the intervals on figure 5.4) and its standard deviation (half-amplitude of the intervals). To mitigate the influence of extreme observations due to the stochasticity of the model, and thus provide accurate estimates, results presented in this section have been aggregated over 2500 runs instead of 50.

The quality of fit varies between scenarios. For example, DNF-based models achieve better fits in scenarios 6, 14 and 15, while the Bayesian model fares better in scenarios 3, 11 and 13. The distances between model and experimental outputs are summarized in table 5.3. This shows a slight superiority of DNF+log over DNF+id, and a slight advantage of the Bayesian model when it comes to representing the localization variance only.

Meanwhile, DNF come with the ability to model complex dynamical behaviors and

Table 5.3: Comparison between our model with logpolar transformation (DNF+log), without logpolar transformation (DNF+id), and the reference Bayesian model, using root mean square error between simulated and experimental data over the 15 scenarios.

	Error between means	Error between SD
DNF+log	**0.626**	1.33
DNF+id	0.638	1.38
Bayesian	0.677	**1.28**

are closer to known neurobiological mechanisms. So it is worth noting that our model enables a versatile point of view of multisensory integration, for a quantitative fit similar to the classical model. In particular, our model can simulate observations on a smaller scale (one run is one human decision) than Bayesian models (mostly focusing on the global distribution of the results). Our model can simulate all random variations between observations, while staying faithful to important mechanisms of multisensory integration.

5.3.3 Parameter exploration

Our model already shows quantitative results comparable to the most standard modeling paradigm, but there are other useful properties that can be displayed. In this section, we will verify that performance is indeed consequent to our design choices, and not of overfitting. We will also show that there is still room for finetuning if one were to target some more specific criteria (such as a maximal fit of localization variance).

In order to emphasize parameter interactions in the most readable way, we have chosen to display the effects of two parameters at a time. In figures 5.5 to 5.7, six parameters keep the selected values mentioned in section 5.2.3.2, and two vary on a regular grid within the bounds given by table 5.2. We will only consider the DNF+log model from now on, our original and most complete version (even though similar analyses could be obtained with DNF+id).

We have found that depending on parameters, model behavior could fall into one of the following four categories. Only the first one is relevant to our simulation, the others will be masked in following figures.

1. For all scenarios, one single peak of activity emerges and stabilizes (often called a "bubble" in DNF literature). The rest of the field is inhibited thanks to lateral inhibition.

2. One bubble emerges but does not stabilize. The maximum potential increases indefinitely because of self-excitation. This is clearly implausible on a neural level.

3. No bubble emerges by lack of interaction, i.e. the term factored by W in equation (5.4) is negligible compared to the others. So the potential U will converge to an approximation of I. Two peaks will be observable when the stimuli are spatially discrepant, but they do not correspond to a bubble enhanced by self-excitation. The outcome is that the decision-making role of DNF goes missing, which falls far away from our objectives.

4. In scenarios where stimuli are far apart, two distinct bubbles emerge. This happens when there is not enough long-range inhibition for one bubble to take over the other. Psychophysically, that would account for an observer explicitely noticing that there are two distinct stimuli. [Alais and Burr, 2004] do not report this happening in their experiments.

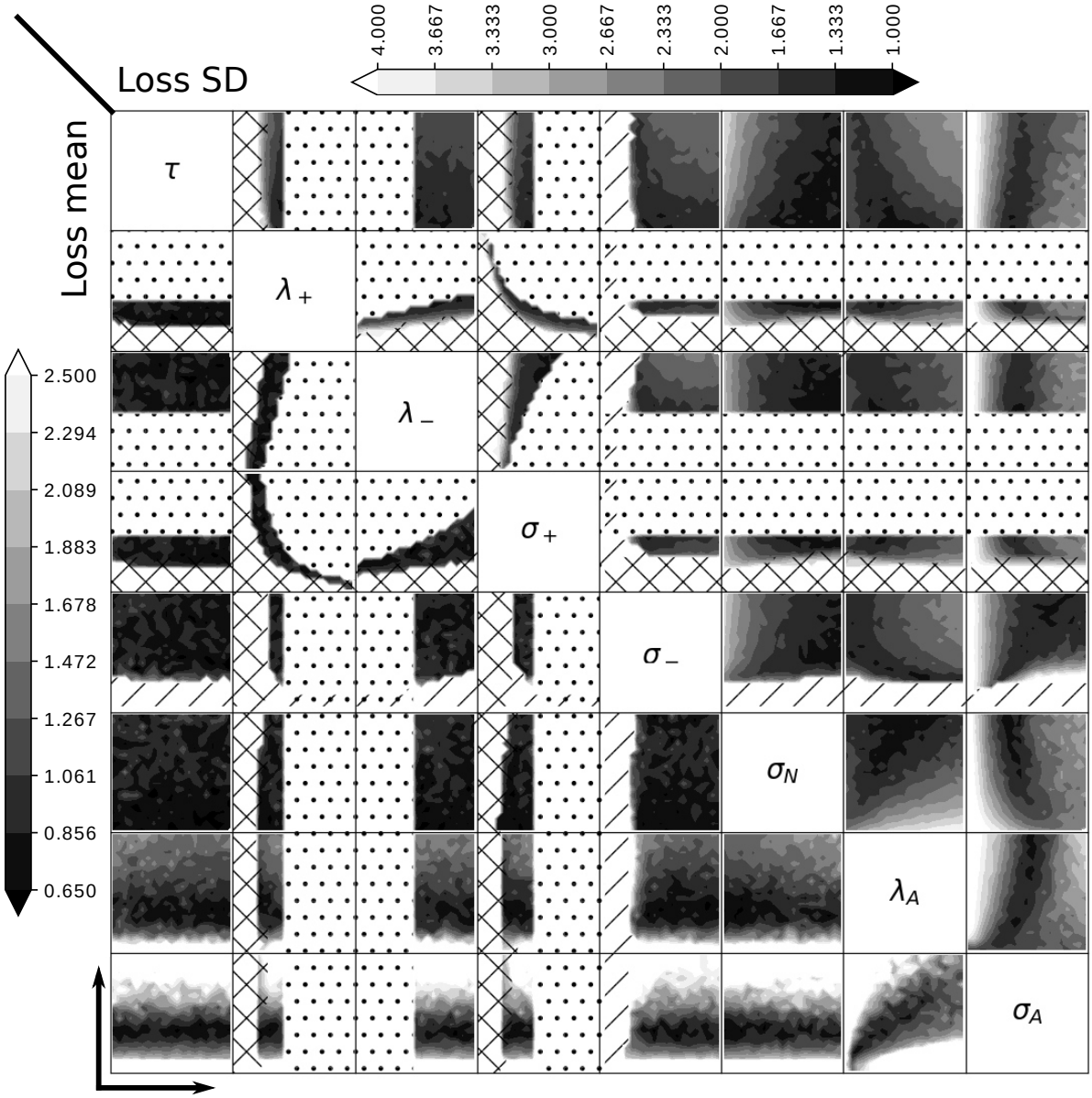


Figure 5.5: RMSE obtained by the DNF+log model depending on pairs of parameters. The bottom left triangular matrix is based on errors in mean localization of bimodal presentations, the top right one on their standard deviations. For each entry, the parameter labeled in row increases from bottom to top, and the parameter labeled in column increases from left to right. The blank areas filled with geometrical shapes designate parameter sets that fall out of scope of our simulation plan (cf. section 5.3.3). Dotted: no convergence, or overflowing activity (case 2). Hatched: more than one peak (cases 3 and 4). Crossed: no interaction (case 3).

5.3.3.1 Pairwise variations

Our first step is to make all 8 parameters of our model vary by pairs. The results are compiled in two triangular matrices (one for each error measure) in figure 5.5 (means bottom left, SD top right), of which each element contains a 2D regular grid. The bounds of each parameter are listed in table 5.2.

First, we can see that τ and σ_N have a strong effect on the localization standard deviation, and a slight effect on the mean localization. In general, increasing σ_N or decreasing τ would give moderately less reliable localization means, but more plausible standard deviations. This is coherent with our simulation paradigm: increasing σ_N means adding more noise, and decreasing τ means a quicker integration of new information through time, both increasing the weight of the noise relatively to the stable audio and

visual stimuli. We can also see that the mean localization is not completely smooth, and even less so for higher σ_N or lower τ. As a reminder, our results are by default aggregated over 50 runs for each parameter combination, for the purposes of smoothing the graphics. Fluctuations caused by extreme values are still expected, so it is consistent that they become more apparent when the amount of noise in the system is increased.

There is some predictible interaction between λ_A and σ_A. The graphs outline a parabola-shaped ridge, along which these parameters can evolve with little impact on the results. It is worth noting that an increase of σ_A can be compensated by an increase of λ_A. That is a characteristic of the DNF. The model is designed to select in priority stimuli whose profile match the positive part of the interaction kernel, which is very thin in the case of the selected parameters ($\sigma_+ = 0.85°$, or 0.12 mm after rescaling). When σ_A augments, the auditory stimulus strays further away from the thin template, and loses weight in the DNF integration. This loss of importance can be artificially compensated by an increase of λ_A.

Interaction kernel parameters λ_+, λ_-, σ_+ and σ_- have clear bounds. In a DNF, when a peak forms due to self-excitation, a minimum amount of inhibition is necessary for the system to stabilize. Too much excitation or too little inhibition will cause the peak to increase in amplitude indefinitely, which does not fit plausibly to any neural mechanism. On the contrary, too little excitation and no peak will form, no interaction will happen and the model will simply replicate its inputs as outputs. This is out-of-scope because it is impossible to generate a saccade or focus for fine-grained processing two stimuli that lie in different locations of the visual field. It is worth noting that λ_+ has an impact on the thresholds for λ_- and σ_+, and vice versa. That means that any of these parameters can be tweaked largely, as long as some ratios of excitation or inhibition are maintained. Interestingly enough, σ_- is less affected by the other three. The main use of this parameter is to ensure the presence of long-range inhibition, so it primarily needs to be sufficiently high. That is consistent with alternative implementations of DNF in the literature, where local inhibition in W is replaced by a constant global inhibition parameter, in situations where only one stimulus should be selected in the entire field [Schöner et al., 2015, Taouali et al., 2015]. This can be seen as a reduction of equation (5.2) with σ_- tending to infinity. Our model does not make this restriction: while a multi-selection is irrelevant in our application to the ventriloquist effect, we did not make the assumption of a unique selection in the entire SC.

5.3.3.2 Reducing the dimensionality of the parameter space

Some regular grids present ridges along which the two parameters vary while the model error stays approximately constant. This is particularly clear for the pair (λ_+, σ_+), allowing us to define a parametric curve on the optimal performance ridge which covers the whole range of parameter values. This curve is defined as a function of an abstract parameter p_+, with the grids and curves for the localization mean and standard deviation reproduced on figure 5.6. The use of p_+ allows us to check for interaction with other parameters, with one less dimension, and to cancel the effect of the local excitation parameters on the model error. The new grids made with p_+ are given in figure 5.7.

We can see that there are no interaction effects left, including between p_+ and λ_-. This confirms that the model behavior remains approximately invariant to its excitation parameters as long as as certain ratio is kept. Consequently, the number of parameters in our model could be decreased: for each value of σ_+ within a certain range, there is a value of λ_+ that achieves a similar fit.

The representation of figure 5.7 also makes clear the tolerable range of certain parameters, and the latitude in their tuning. Inhibition parameters have to exceed a certain

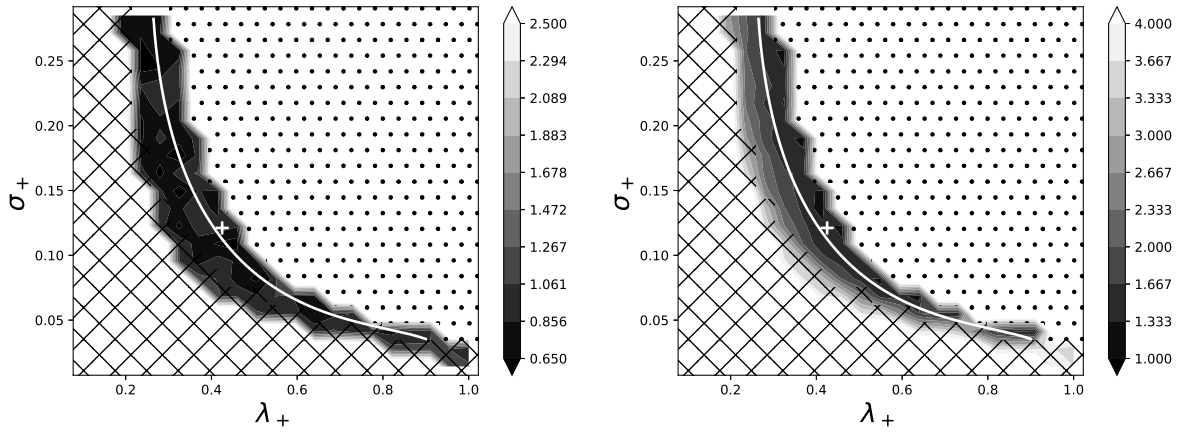


Figure 5.6: RMSE obtained by the DNF+log model depending on λ_+ and σ_+ (expanded from figure 5.5). The left graph is based on mean localization of bimodal presentations, the right graph on their standard deviations. The white cross indicates the default values used in the previous section. The white line shows the parametric curve that will be used for parameter reduction. The blank areas filled with geometrical shapes designate parameter sets that fall out of scope of our simulation plan. Dotted: no convergence, or overflowing activity. Crossed: no interaction (U replicates I).

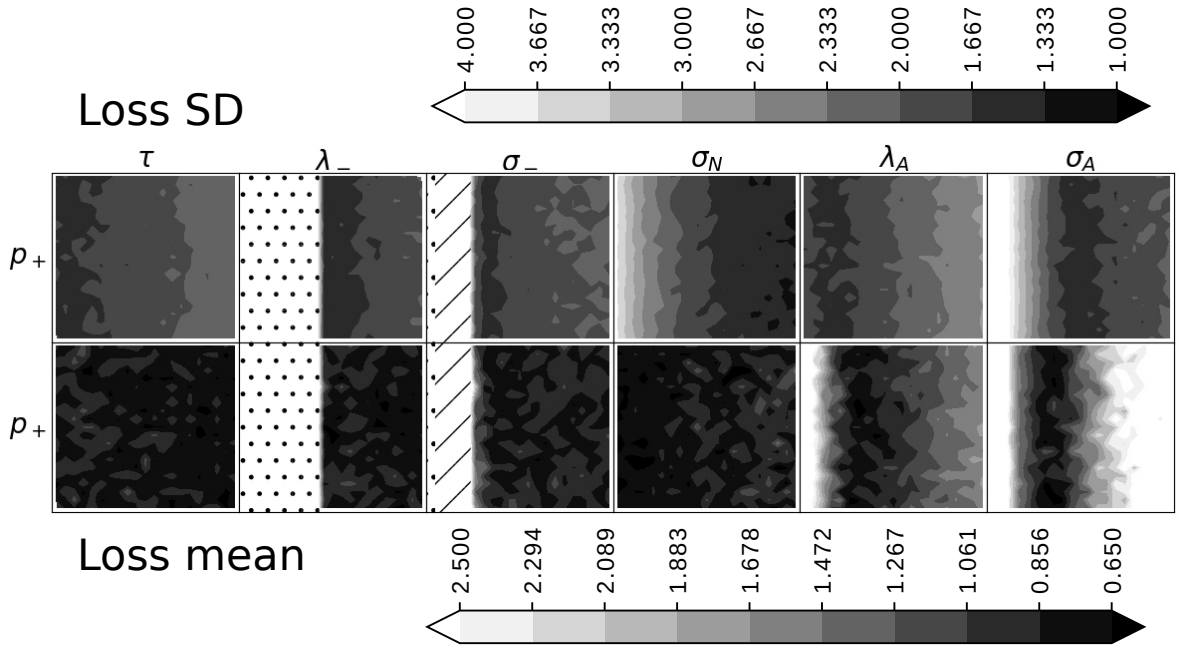


Figure 5.7: RMSE obtained by the DNF+log model depending on p_+ (from the parametric curve of figure 5.6) and other parameters. The bottom row is based on mean localization of bimodal presentations, the top row on their standard deviations. In each entry, the parameter labeled on the top increases from left to right. The bottom of a square corresponds to a low λ_+ and high σ_+, the top corresponds to a high λ_+ and low σ_+. See figure 5.5 for the rest of the legend.

threshold ($\lambda_- > 0.11$, $\sigma_- > 5°$), otherwise the self-excitation of the DNF will not be compensated, and the membrane potential U will increase endlessly. In addition, σ_- must be high enough (above approximately 30°) to ensure that only one peak is selected. We can see that a better fit in localization standard deviation can be attained by either decreasing τ or increasing σ_N, but at the detriment of the fit in mean localization. Similarly, λ_A and σ_A show vertical strips where the fit is maximal, but these strips do not coincide between

both error measures. Given our goal of reproducing in general aspects a psychophysical experiment, we have had to settle for a good quantitative fit in both criteria. But as we can see, if our objective was to fit either the mean localization or its standard deviation, performance could be increased substantially. There are no sharp ridges or spikes, and the local optima (see darkened areas on figure 5.7) are quite wide, so the parameter fitting would be relatively smooth, and the results we obtained in table 5.3 do not rely exclusively on finetuned values of many parameters.

In summary, there are several ways the number of parameters can be decreased. We have seen earlier that changes in λ_A and σ_A can compensate each other, so λ_A could be fixed arbitrarily, and some finetuning would be feasible with σ_A alone. σ_A determines, together with the kernel parameters, the relative weight each stimulus will have in the DNF. For an estimation of the mean localization of the bimodal signal, if we assume that λ_- and σ_- always remain above a necessary threshold, and that λ_+ and σ_+ are restricted to the parametric curve in figure 5.6, then we are left with only two free parameters: p_+ and σ_A. Remaining parameters intervene in the dynamic capabilities of our model (e.g. to predict response times) and its ability to explain some of the inter-observational variations.

5.4 Conclusion

Models of multimodal merging in psychophysics come predominantly from the Bayesian paradigm. We have shown, using the ventriloquist effect as an illustrative example, that it is possible to model such a task using a neurally-inspired, population-based dynamical system. The model we created conciliates known characteristics of the superior colliculus and the paradigm of dynamic field theory, reaching a quantitative fit comparable to the classical paradigm. The difference between the two models has to be examined at a more theoretical level, given that they operate at different levels of abstraction. DNF are meant to model neural dynamics [Amari, 1977]. While they do not constitute an exact simulation of neurons at a microscopic level, the behaviors that emerge from the dynamic system echo physically observable neural patterns at a larger scale, aggregating over thousands ot neurons. Bayesian models of multimodal fusion, on the contrary, were not derived to accurately relate to biological mechanisms (although fine-grained Bayesian models may be perfectly fit to model such mechanisms), but rather to estimate subjects' decision distributions at coarser spatiotemporal scales. Using the terminology from [Marr, 1982], the Bayesian model operates at the level of the computational theory, in that it describes the logic by which information coming from different sensory modalities will be integrated, without delving into the ways the inputs are represented or the algorithm is implemented. DNF models could be placed in the other two levels: either representation-algorithm, when the way inputs are transformed into a decision is described through mathematical equations; or hardware implementation, when we consider the discretized field where each neuron acts as a processing unit. Note that these levels are not mutually exclusive, and previous works have hinted at perspectives to analyze either Bayesian modeling [Ma et al., 2006] or DNF [Gepperth and Lefort, 2016] at the level of the other, among many attempts to explain Bayesian-compatible observations through operational models [Pouget et al., 2002, Weisswange et al., 2011, Parise and Ernst, 2016]. In any case, this different positioning does not preclude the ability of any of these paradigms to generalize to a wide range of tasks and mechanisms. Both make sense at their own level, although it can be argued that Bayesian modeling might be too broad to capture some of the most subtle behaviors that may emerge from neural interaction [Jenkins et al., 2021]. That additional precision of DNF comes at the cost of an extended parameter space.

It is worth noting that our choice of parameters is not detrimentally constraining.

There is some latitude in the parameter tuning, thus our modeling hypotheses do not particularly weaken the value of our results. In particular, there is flexibility in the shape of auditory inputs (the model does not rely on one specific pair of values (λ_A, σ_A)), and quantitative fit did not discriminate against the use of the logpolar transformation.

The relative freedom in model optimization opens up new simulation perspectives. First, there is room for additional parameters and tuning, not included in our current simulations as a first parsimonious approximation. For instance, in our model, as in many previous DNF models [Wilimzig et al., 2006, Fix et al., 2011], white noise is used while not spatially correlated. One could expect that spatially correlated noise (as used in [Taouali et al., 2015, Jenkins et al., 2021]) would help fit the variance better, especially in scenarios involving a very thin visual stimulus. Then, we have seen that the parameter dimensionality could be reduced (for example by removing σ_- and using global inhibition), and that some pairs of parameters could compensate one another in an optimization task (most notably, λ_+ and σ_+, τ and σ_N, λ_A and σ_A). Consequently, we have reason to believe that our model can be used to fit more demanding tasks. A hypothetical situation would be to simulate a bimodal perception task and fit both the signal localization and an observer's response time. One could then consider locking pairs of parameters on parametric curves (as we did with λ_+ and σ_+) for localization fitting, and use the newly freed dimensions (such as p_+) to fit for the additional constraints.

Indeed, our model has room for the integration of additional functionalities, and the first novelty brought by DNF stands in its dynamic properties. DNF are fully capable of integrating any kind of time-dependant signals (so long as they can be projected onto a topological map). Moreover, their inner dynamics may account for behavioral responses of a human during the perception process. For instance, the peaks of activity in the DNF can generate population-coded motor commands for visual saccades [Wilimzig et al., 2006, Quinton and Goffart, 2018]. While the experimental data we have used did not highlight any particular time-related merging effect, our model incorporates by design the groundwork for the modeling of new dynamic properties.

Additionally, we have seen that DNF are suitable when perceptive fields are not homogeneous across the map, as was showcased by the logpolar transformation. In that particular case, the expectation is that a visual stimulus that appears further away from the fovea will have an increased precedence in the audiovisual fusion. Indeed, in the periphery of the retina, the logpolar transformation will activate a smaller region of the multisensory map, and in our case the DNF matches thinner signals better. This situation is out of scope in the classical ventriloquist experiment, which centers on the fovea, with little eccentricity. This limitation in the experimental data may explain the lack of difference we found between DNF+id and DNF+log. But our simulation would still provide an interesting baseline for the modeling of eccentric audiovisual merging, especially with regards to saccade generation. A visual signal in the border of the field of view will be a likely target for a saccade, although (or, according to many models of saccade generation, because) it is seen less precisely. At the psychophysical level, how much this interferes with the general paradigm of multisensory integration (for which a less precise visual stimulus would actually be captured more easily by other modalities) is still an open question. However, on a computational level, our model reunites some of the keys to a common ground between multimodal fusion and active perception.

The incorporation of a logpolar transformation is a first step towards a generalization of neural maps for use by DNF. The topology derived from this particular transformation is theoretically adapted to the processing of audiovisual stimuli for saccade generation in the human SC. But, for other kinds of tasks, and other kind of agents, including robots, an adapted topology may be very different. We are used to placing stimuli in regular maps, but now that we have tested one case of an irregular, bio-inspired topology, with

acceptable results, we can wonder what other topologies could be tested and what impact they would have on fusion. In the experiments led in this chapter, this is especially relevant with regards to the auditory modality, which we had to project on a visually-grounded map with a strong hypothesis — namely, that auditory stimuli can be expressed as Gaussians in a regular 2D map. We made a choice that was convenient in this use case, and we had to parametrize the size of auditory stimuli in return. So, our follow-up question is, if we could capture the properties of the auditory modality into a new readapted topology, maybe its reliability, or any relevant irregularity, would be carried over; just like a SC-inspired topology gives different reliabilities to stimuli sensed in the fovea and in periphery.

Interlude

Back to the story. The final fight between the player characters and Ypomni has started, and has been going on for quite a while now. In fact, most characters are knocked out, with only Bob's hero left standing against the big boss, who is greatly injured. Bob, who plays an archer, announces he is going to shoot his last arrow, but in a final twist, Ypomni casts a spell that plunges the entire room into complete darkness. Bob, unimpressed, decides to target the boss by ear. Alice is puzzled: the rulebook specifies what ability check should be done in game in order to shoot a target in line of sight, but it does not say what to do when a character aims at the sound of a target. This sparkles a debate:

- Bob estimates that aiming at an auditory target is not harder than aiming at a visual target, so he should do the same check as usual.
- Alice argues that vision is much more reliable than audition because we can look at the target, and we see much better in the center of the field of view. Aiming by ear would be equivalent to aiming at a target seen from the corner of the eye. Thus Bob should have a malus on his check.
- Eve mentions that people are better at locating sounds horizontally than vertically. So aiming should not be harder as long as Ypomni does not crouch or jump.
- The rest of the players have fallen asleep.

In fact, both Alice and Eve are partially correct. Sounds are indeed located with a better precision on the azimuthal plane — the plane that aligns with the two ears of the listener — than in elevation. But auditory localization is still far less reliable than visual localization in normal conditions, especially when the perceiver can freely gaze at the target. (Unless, of course, vision is completely blurred or obscured.)

These irregularities are the new point of interest of the upcoming chapter. The objective is to learn new topologies that reflect the specificities of different modalities. These topologies will be used as a support of decision — in multimodal merging tasks in particular.

After another too long debate, the players and Alice agree on making a regular ability check with a small malus. Bob throws the die. His character shoots his last arrow.

To be continued.

Chapter 6

Learning topologies for fusion

Contents

The work presented in this chapter was initially intended to fill the gap of audio stimulus representation in the previous model, and replace it with a more biologically-plausible projection. It was quickly extended to any modality and any set of contraints. The expected challenge — to be able to use classical DNF in a learned, irregular, multimodal map — turned out to lead to some quite interesting properties. Most of the content of this chapter has been presented in: Forest, S., Quinton, J.-C., and Lefort, M. (2022). Combining manifold learning and neural field dynamics for multimodal fusion. *2022 International Joint Conference on Neural Networks (IJCNN).*

6.1 Introduction

When it comes to information processing and behavioral decision-making, the way we merge data coming from inputs of mixed nature is becoming increasingly important. Let us start with a toy example that we will follow for most of this chapter. A robot is placed in a room filled with objects, and is given a task, for example: "touch the alarm clock when it goes off". At first, the robot might be facing several objects resembling an alarm clock. Given the recent huge progress made in computer vision, it should have no difficulty recognising them. When a sound goes off, the robot should be able to locate its origin, but it is usually achieved with a low precision. Before taking an action, the robot has to select an object. Here, it should be the one clock-looking object that coincides most with the sound source localization. But how the modalities should be weighted depends not only on the task, but also on the reliability of the sensors. A perfectly visible alarm clock in the center of the field of view should not be preferred to a partly concealed clock in the corner of the camera when the sound seems to come from the sides, but at the same time, audition should not be heavily relied upon if, say, workers have been drilling holes in the contiguous room for the whole day. And then, even when it starts moving towards the right target, the robot should maintain its decision as the environment changes around him, and the strength of the stimuli may fluctuate or temporarily stop.

The task in this example faces multiple challenges, starting with two: the fusion of sensory modalities of different availability and reliability, and the selection of (and attention towards) a target. To tackle these problems, most of model nowadays are based on deep learning. In this chapter, we propose a novel approach based on dynamic neural fields (DNF), which we first described in chapter 4 and already applied to fusion in chapter 5.

One limit that previous DNF implementations have faced lies in the nature of the manifold they evolve on. Most applications in the literature assume the existence of an underlying regular topology, most often 1D or 2D. But it is hardly representative of the disparities in the sensory space, disparities which become crucial when performing multimodal fusion. We have started playing around this in the previous chapter by transforming inputs, but the DNF used after the transformation still evolved in a regular 1D map.

So let us take a look at the shape of stimuli perceived from the environment. The quantity of information available is huge, and the data an agent receives from its sensors is only a projection of it in a few given dimensions. Equipped with a standard camera, a robot will receive a projection of the part of the environment it is facing. This projection, that we called the sensory space in chapter 2, is in as many dimensions as there are pixels in the camera, but there is an evident underlying 2D topology (a first feature space) in it. With one microphone, the robot can detect sounds from anywhere around it, but it can hardly locate them. Two microphones may enable some 1D sound localization (in feature space) along the axis on which they are aligned, usually azimuthal (with interaural time/level difference), and even a bit of 2D or 3D by exploiting the shape of pinnae with a head-related transfer function (HRTF) [Argentieri et al., 2015]. We must first account for the specificities of each sensory modality before we create behaviors that exploit it at best. Additionally, we must find a way to match complementary information from different modalities, which in machine learning usually boils down to projecting stimuli onto a common manifold.

So, our first step will consist in learning unimodal manifolds. For this purpose, we will use growing neural gas (GNG) [Fritzke, 1995b], a manifold learning algorithm which is quite parcimonious in light of the possible complexity of the sensory space. Then, we will suggest an easy-to-implement solution to create a multimodal manifold suitable for fusion.

The main novelty of our work is that we will adapt DNF directly on this new topology, even though it lacks the regularity and low dimensionality of classical implementations. We will show that properties of DNF in selection and attention are compatible with such fabricated manifolds, and that this coupling allows new possibilities for multimodal fusion taking into account the relative resolution of the modalities.

6.2 State of the art

6.2.1 Manifold learning

Sensors provide high-dimensional samples of the environment, but sensory spaces can often be projected onto manifolds of lower dimension. Deep learning methods are particularly suited for learning such manifold (see [Bengio et al., 2013] for a review). For example, the last layers of a deep neural network have been shown to contain an intrinsic dimensionality that is smaller than the number of features in the data [Ansuini et al., 2019]. Dedicated methods such as variational autoencoders [Kingma and Welling, 2014] learn structured embedding in an unsupervised manner. As our focus in this chapter is the study of coupling between DNF and irregular multimodal manifold, we will use simpler methods (i.e. self-organizing neural networks) that will provide more control and insight for the study.

In self-organizing maps (SOM), e.g. the Kohonen model [Kohonen, 1982], each neuron represents a prototypical input in the high-dimensional sensory space, so that the input space is projected onto a neural lattice of fixed size and structure. SOM can have two limitations. First, the number of nodes is fixed, and thus might be insufficient to accurately map a complex intrinsic space. This is circumvented by variants that increase the map size at regular intervals until a given stopping criterion is reached: growing cell structure (GCS) [Fritzke, 1994], where nodes form a mesh of hypertetrahedrons (triangles in 2D), one of which get splits whenever a new node is added; growing grid (GG) [Fritzke, 1995a], where nodes are added by inserting (hyper)rows/columns in a rectangular grid; and others [Marsland et al., 2002, Van Hulle, 2012]. Second, the structure is fixed (e.g., a 2D triangular/rectangular/hexagonal grid). This means that the map dimensionality might not match the intrinsic dimensionality of the sensory space (e.g., fitting a cube on a 2D map). There are alternatives on this side too. In neural gas (NG) [Martinetz and Schulten, 1991], neurons are not arranged on a lattice, but are connected following a Hebbian-like rule, thus neurons with close prototypes are linked together. Eventually, the gas fills the input space in a way that matches the stimulus distribution. Similarly to GCS and GG, growing neural gas (GNG) [Fritzke, 1995b] is a derivative of NG, in which neurons are added (or deleted) at regular intervals.

Growing maps are still limited by the choice of the stopping criterion, often chosen as a maximum number of nodes, or a minimal accumulated error. They are also not adapted for online learning. Additional model variants set new conditions for nodes to be added or deleted on the fly: incremental growing grid (IGG) [Blackmore and Miikkulainen, 1993], where nodes are added at each iteration, but connections are created or deleted when the distance between nodes reaches certain thresholds; growing neural gas with utility criterion (GNG-U) [Fritzke, 1997], where nodes are removed whenever their utility (a measure of how much precision would be lost in the absence of this node) turns sufficiently low; grow when required (GWR) network [Marsland et al., 2002], where nodes are added only when the minimal distance between node prototypical input and drawn input exceeds a threshold; and others [Van Hulle, 2012]. The models presented here are summarized in table 6.1.

Table 6.1: Non-exhaustive list of self-organizing models

Fixed structure (e.g., 2D)	**No fixed structure**	
SOM [Kohonen, 1982]	NG [Martinetz and Schulten, 1991]	Fixed number of nodes
GCS [Fritzke, 1994] GG [Fritzke, 1995a]	GNG [Fritzke, 1995b]	Nodes added/deleted at regular intervals
IGG [Blackmore and Miikkulainen, 1993]	GNG-U [Fritzke, 1997] GWR [Marsland et al., 2002]	Nodes added/deleted under conditions

6.2.2 Use in sensor fusion

Numerous articles have shown promising results in multimodal fusion using deep learning. Deep unsupervised learning can be used to project multimodal data on a low-dimensional manifold for use in robotics [Droniou et al., 2015]. Inputs can be mixed during neural network training to exploit the correlations between modalities [Yang et al., 2017]. [Jaegle et al., 2021] proposes a new type of deep neural network receiving multimodal inputs allocated through an attention module. Unfortunately, most of these works make the assumption that all multimodal data are related. Also, deep architecture are dedicated to one specific task and no generic architecture emerges [Ngiam et al., 2011]. We aim to create a new multimodal topology over which new dynamic properties could be applied, and self-organization offers solutions for a much lower cost [Ménard and Frezza-Buet, 2005, Johnsson et al., 2011, Lefort et al., 2013, Lallee and Dominey, 2013, Vavrečka and Farkaš, 2014, Parisi et al., 2017, Huang et al., 2019, Khacef et al., 2020, Gonnier et al., 2021].

SOM and their derivatives have long been used as models of multimodal fusion, but the ways modalities are combined can be very diverse. Map architectures can be divided in two categories. In the first, one SOM is trained for each modality, then all unimodal maps are connected depending on a special learning rule [Lefort et al., 2013, Khacef et al., 2020, Gonnier et al., 2021]. In the second, unimodal maps link to a new multimodal SOM [Ménard and Frezza-Buet, 2005, Lallee and Dominey, 2013] or NG [Vavrečka and Farkaš, 2014] that combines all information. Additional layers of SOM can also be considered to create a hierarchical flow of information [Johnsson et al., 2011, Parisi et al., 2017, Huang et al., 2019]. Additionally, models can be made more adaptive to time-dependant tasks with the help of GWR maps [Parisi et al., 2017, Huang et al., 2019]. Some of these models have already been proof-tested for visual, auditory and/or proprioceptive modalities on hardware setups [Johnsson et al., 2011, Huang et al., 2019] and robots [Lallee and Dominey, 2013, Gonnier et al., 2021].

After multimodal maps and/or interconnected unimodal maps have been learned, we need a paradigm to dictate the way perception will occur. Multimodal perception can be seen as a form of decision pondering sensory inputs of different reliability and relevance. We follow the architectural choice made in [Ménard and Frezza-Buet, 2005] and [Lefort et al., 2013], where dynamic neural fields (DNF) are used as the paradigm that governs fusion or segregation of stimuli in the multimodal topological space.

Reminder on DNF for fusion The vast majority of works using DNF assume the dynamics take place on a completely regular topology, e.g. a 2D lattice in the case of vision. However, there is no clear way of projecting two or more modalities onto the same lattice. In [Schauer and Gross, 2004] and in chapter 5 of this thesis, strong assumptions are made on the shape of stimuli in a modality (audio in our case) so that they fit in the topology of the other (vision). To mitigate this issue, [Lefort et al., 2013] propose using

separate manifolds for each modality, each learned by SOM, and apply DNF on each of them. Communication between modalities is ensured by a specific set of topographic connections.

The latter reference is actually one of the first to suggest using a learned manifold as the theater of neural dynamics. Otherwise, some attempts to alter the projection of inputs into the manifold have lead to satisfying results: [Taouali et al., 2015] and our chapter 5 successfully reproduce biological behaviors after applying a logpolar transformation to visual stimuli, which models the discrepancies in the resolution of the human retina [Ottes et al., 1986]. In [Lefort et al., 2013], the projections received by neurons are altered, although they are still organized in a rectangular lattice. Since DNF are strongly dependant to the topography, and usually rely on a symmetrical interaction kernel, one may fear that breaking the regularity of the underlying topology may make DNF completely unpredictable. Breaking the symmetry from the DNF side has been suggested before, either through asymmetrical kernels [Cerda and Girau, 2013] or through distortions of the topology by predictive reinforcements [Quinton and Girau, 2011], but these methods require some amount of learning as well.

An ensuing question would be how far from regular and/or rectangular/cubic can the underlying topology be for DNF to remain viable. If DNF could be made to operate on manifolds of unconstrained shape or dimension (easily accessible through GNG), then this would open the door to adding the properties of DNF to new applications, starting with new capabilities in multimodal fusion like the ability to take into account the different resolution and reliability of all modalities. To our knowledge, this has not been tested before. At best, suggestions have been made to approximate DNF activity using gaussian mixtures, sparsifying the space on which they operate to make them applicable in more complex topologies [Quinton and Girau, 2010]. Yet, this latter approach still relies on a continuous regular space on which the lateral connectivity kernel function and Gaussian mixtures can be defined, which remains a strong limitation when processing high dimensional inputs.

As an intermediate summary, we can see a need to learn a multimodal manifold that unites features from different modalities. Self-organization algorithms such as GNG are a parsimonious method of creating new topologies. Interesting fusion properties, such as stimulus selection, attention, or robustness to noise, can be brought by DNF, however the paradigm has not been tested on such irregular manifolds before. This is the purpose of the remainder of this chapter.

6.3 Methods

In this chapter, we use GNG to learn manifolds of the sensory space in each modality. We then assemble them into one multimodal graph, on which we use a DNF to produce behaviors that have, to our knowledge, never been implemented before on this kind of manifold. These three steps are summarized in figure 6.1 and explained in the next three subsections.

6.3.1 Unimodal topology learning

In this step, we process modalities separately. As our focus in this chapter is not on tuning the unimodal topology learning on a specific task, we use the standard GNG algorithm with its original parameter values, as described in [Fritzke, 1995b]. To summarize, GNG are trained by receiving a succession of randomly selected stimuli. Every time, the two neurons whose prototypical input match the stimulus best get a fresh connection. Then the best-matching unit (BMU) and its direct topological neighbors have their prototype

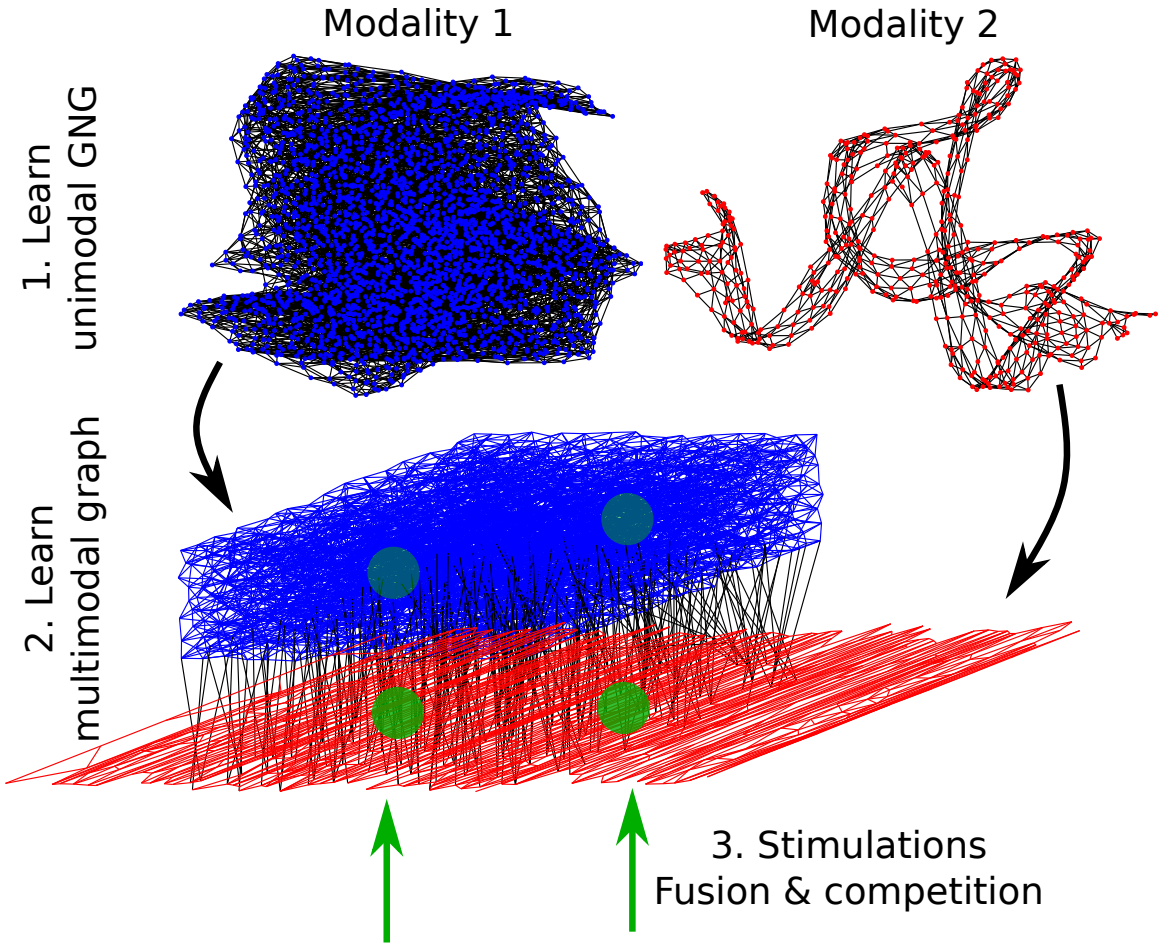


Figure 6.1: Recap of the steps taken in this chapter. 1. Learn a growing neural gas in each modality (resp. blue and red nodes with black connections). 2. Assemble them into one single graph by creating multimodal connections (new black connections between blue and red nodes). 3. Present stimuli and compute multimodal activity.

moved towards the stimulus. Connections that have not been updated in a long time are removed, and isolated neurons as well. Then at fixed intervals, a new neuron is inserted. Its prototypical input is placed at the middle of the most activated connection.

6.3.2 Multimodal topology learning

In the rest of this chapter, we will focus on bimodal architectures, but most of what follows is applicable to more than two modalities. As a reminder, bimodal architectures in self-organization literature often merge data in one of two ways: a multimodal map is created that receives information from the unimodal ones, or new connections are added between the unimodal maps, each having its own processing unit. We propose an intermediate solution that is the most parsimonious of all: we create a new bimodal graph that contains all nodes and edges from one modality, and all nodes and edges from the other. To create the crossmodal edges, we connect neurons of the two modalities that fire together, in an Hebbian-like manner. More precisely, we have tested two algorithms:

1. Draw a random multimodal input. If it lies in the sensory range (set in advance) of both modalities, find the BMU in each GNG and connect them (if they are not already connected). Repeat until a certain proportion of nodes have at least one crossmodal edge.

2. Browse through one of the gas. For each node, compute its coordinates back in

sensory space and connect it to the node that matches it best in the other gas. Repeat with the other gas.

Method 2 ensures that all neurons (where receptive fields overlap) have at least one neighbor in the other modality. This is a strong constraint, and according to our preliminary tests, it will not usually yield different results than the method 1, as long as the amount of random draws exceeds the total amount of neurons by a small factor. The only time the method 2 can be needed is when the manifolds have a significantly different resolution in some localities. In that case, random draws might leave a lot of neurons without crossmodal connections, making local dynamics quite unpredictable. Other than that, we advise using the method 1, as it is simpler both theoretically and computationally.

6.3.3 DNF processing

Once the bimodal graph is created, its associated neurons can be stimulated by sensory inputs (through their respective modality), and we can use DNF to select and attend to a stimulus (see chapter 4 for a description and chapter 5 for an application in multimodal fusion). In DNF, the distance between neurons plays an important role, as it determines whether they will excite or inhibit one another. Our model differs from others in the literature in that all neurons do not share a common coordinate system. So, we need to adapt the DNF equation, so that the distances are defined on the graph, and only that. We rely on the standard distance from graph theory, i.e. the number of edges on the shortest path between any two vertices.

In our model, each neuron is tied to a specific modality. So, the external input received individually will be modality-specific (although the rest of DNF operations will not be). To ensure that the total amount of external stimulation is independant from the local resolution of a modality, we will order all neurons of a modality by their proximity to the stimulus (using the euclidian distance in the coordinate system of that modality), and stimulate them descendingly according to their rank. For each neuron indexed k, given a stimulus indexed i, we note $r_{k,i}$ the rank of proximity between the prototypical input of k and the coordinates of i. The external stimulation I_k received by k is given by:

$$I_k = \lambda_{m,i}\, \mathrm{e}^{\frac{-r_{k,i}^2}{2\sigma_I^2}} \tag{6.1}$$

where $\lambda_{m,i}$ is the intensity of stimulus i with regards to k's modality m. A neuron can only receive external inputs from its own modality.

Next, we compute the evolution of activity in the graph over time. The following is completely modality-agnostic. The potential U_k of neuron k is initialized as 0 and updated incrementally by[1]:

$$\Delta U_k = \frac{\Delta t}{\tau}\left(-U_k + I_k + \sum_{k'} W(< k, k' >)\, f(U_{k'}) + h\right) \tag{6.2}$$

where Δt is the simulated time between steps, τ a time constant that determines the speed of DNF updates, f an activation function (ReLU), and h a negative resting level. $< \cdot, \cdot >$ designates the minimal distance in number of edges between two nodes in the bimodal neural gas, and W is a weight function expressed as:

$$W(\delta) = \lambda_+\, \mathrm{e}^{\frac{-\delta^2}{2\sigma_+^2}} - \lambda_-\, \mathrm{e}^{\frac{-\delta^2}{2\sigma_-^2}} \tag{6.3}$$

1. In this equation, only U_k is incremented over time, and the inputs I_k are static. However, none of our hypotheses prevent the inputs from being updated over time. We make this choice because dynamic inputs are not necessary for the results presented in this chapter.

with amplitudes $\lambda_+ > \lambda_- > 0$ and widths $\sigma_+ < \sigma_-$. W can be seen as a kernel shaped like a mexican hat [Amari, 1977]. So, this is essentially the same equation as in previous chapters, only the distance metric changes.

As we did before, we take a barycenter of the output $f(U)$ as an estimator of the position targeted by the model. While we are not supposed to know an euclidian topology in which the positions of GNG nodes can be averaged, we can still use the input data to interpolate a corresponding location in a 2D euclidian space for each neuron. We will do that for our visualization purposes, even though this interpolation should theoretically not be possible by default. Similarly, for the GNG, we will plot them by putting all nodes to their asserted location, only for visualization purposes.

6.4 New feature spaces

In this first section of results, we describe the newly-created multimodal topologies that will be used to support DNF operations (in section 6.5). For this chapter, we consider two modalities, vision and audition. That can correspond for example to a robot asked to locate a visual and/or audible stimuli. We test two kinds of inputs: one bio-inspired (SC, subsection 6.4.1), one robotic-inspired (using real HRTF measures, subsection 6.4.2).

So, the main difference between the setups is in the first step of our model, the generation of the unimodal manifolds (as described in subsection 6.3.1). For the GNG training, a stimulus location is drawn within the subspace of the environment that is accessible to the appropriate sensors. For example, a robot's visual perception might be restricted to the space in front of it, while its auditory range might be all around it. Input ranges are listed in table 6.2. Then, we simulate the information that would be received from the sensors if a real stimulus was sent from this position. The way they are preprocessed will be defined in each subsection.

We have set an upper limit to the number of neurons in the GNG. Otherwise, the resolution could become excessively high, increasing the computational cost for no valid reason. Once the limit is reached, the GNG is trained like a regular NG, except that nodes that have become irrelevant can still be removed and replaced. This is still more efficient than starting with all neurons and training a NG from the beginning.

Table 6.2: Ranges of inputs in the external environment

Section	Modality	X-range	Y-range	Z-range (if needed)
6.4.1	vis.	$[0, 90]$	$[-45, 45]$	–
	aud.	$[0, 90]$	$[-45, 45]$	–
6.4.2	vis.	$[-45, 45]$	$[-45, 45]$	$[0, 45]$
	aud.	$[-90, 90]$	$[-45, 85]$	–

6.4.1 Superior colliculus inspiration

Our first experimentation is inspired from observations in neurophysiology. Human visual perception is affected by the heterogeneous distribution of sensors in the retina, giving a higher resolution in the center of the field of view (the fovea) than in its periphery. This disparity can be observed in brain regions processing visual information, such as the superior colliculus [Ottes et al., 1986]. A mathematical model of the disparity between fovea and periphery, using a logpolar transformation, has been suggested by [Ottes et al.,

1986], and previous works have coupled it with DNF for visual [Taouali et al., 2015] and audiovisual processing (chapter 5).

Models of the superior colliculus are not only useful for computational neuroscience. While cameras used by robots are supposed to have a homogeneous resolution, they might happen to have blurry spots because of dirt or wear. Other modalities may also have a high variance in resolution. The logpolar transformation is one way among others to test these variations in a controlled setting. Additionally, even when the camera sensory space is perfectly regular, it has been suggested that adding a logpolar transformation on top of it could improve gaze control in robots [Manfredi et al., 2006, Manfredi et al., 2009].

6.4.1.1 Sensory space

In light of the aforementioned hypothesis, we take coordinates of a visual stimulus in a regular 2D space, and transform them following the logpolar transformation in [Ottes et al., 1986]. The new 2D coordinates are used as inputs for the visual GNG. Since we study the effect of variable resolutions in one modality, the other modality, audio, will be modeled as a regular 2D space as in chapter 5, so that it does not interfere with the analysis. Both GNG are given 1000 nodes maximum.

6.4.1.2 Feature space

Because there is a very localized difference of resolution near the fovea, we use method 2 (cf. subsection 6.3.2) to connect the two GNG. The resulting bimodal graph is shown in figure 6.2. Only edges are plotted here; there is a node at each intersection. For the visualization, visual nodes are positioned according to the reverse logpolar transformation of their features, and auditory nodes according to their raw coordinates. The unimodal GNG are superposed with a different color each.

As expected, the visual GNG has a much higher resolution around the fovea (0°), as can be presumed by the high density of nodes. It gradually decreases as the azimuth augments. On the contrary, the auditory GNG has roughly the same resolution everywhere.

Connections between neurons of different modalities are shown in red. For azimuths between 0° and approximately 30°, vision has a better resolution than audition: most nodes from the audio GNG are connected to multiple visual nodes. The trend is reversed for higher azimuths.

6.4.2 Robot perception

In this simulation, we will partly use real experimental data and show that DNF properties are still available in more complex sensory spaces. Our main change will be on auditory preprocessing. One way of performing sound source localization for robots is to compute a HRTF, a function that associates spectral features (caused by interferences on the signal by the head and pinnae) to source orientations [Argentieri et al., 2015]. Meanwhile, with the current progress on computer vision, we can assume that in most cases, there exists a vision preprocessing subsystem that outputs 2D or 3D position of objects in a relatively homogeneous map. So, we settle for a regular sensory space on vision side.

6.4.2.1 Sensory space

Data provided by [Algazi et al., 2001] includes head-related impulse responses of a robot equipped with artificial pinnae, to a sound located at different angles. Given an external stimulus position in 2D, we can interpolate the responses received by the two robotic ears.

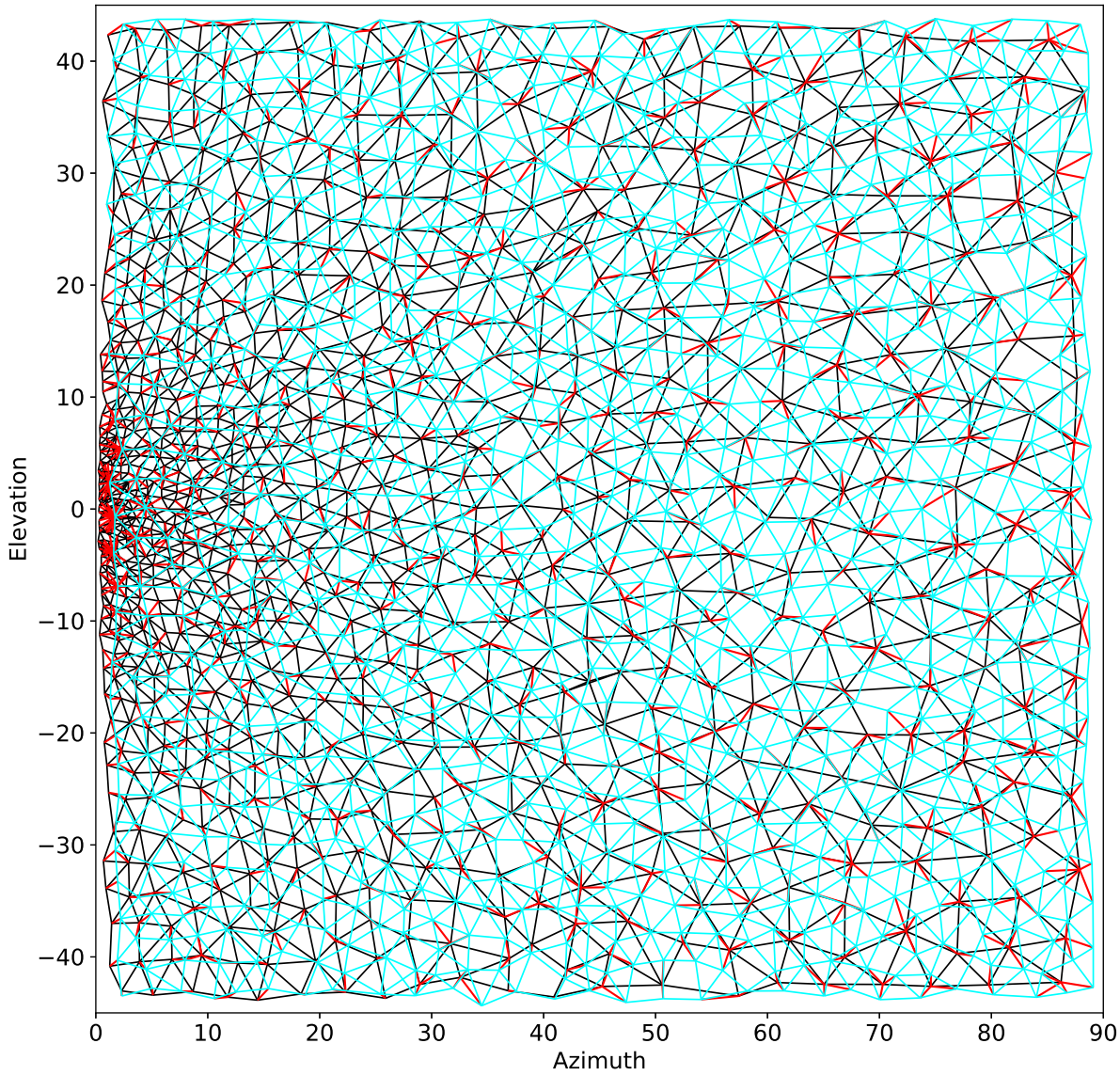


Figure 6.2: Representation of a bimodal graph in SC simulation. Edges are colored depending on the modalities of the neurons they connect. Visual-visual: black. Auditory-auditory: cyan. Visual-auditory: red.

We then compute their Fourier transform and make the difference between the ears to obtain a HRTF. In the end, each audio input is 100-dimensional.

For vision, we consider a robot with intact cameras and assume it can roughly estimate the 2D or 3D coordinates of an object in front of it.

We do not need visual and auditory perception to have the same range. Realistically, stimuli can be heard from more orientations than they can be seen (see table 6.2). To keep resolutions approximately balanced, we use respectively maximum 500 and 200 nodes for auditory and visual GNG.

6.4.2.2 Feature space

In 2D, the visual GNG is very similar to the auditory GNG in the previous section, which also directly received stimuli drawn from a regular 2D space. The new auditory one, however, has a distinct shape. Figure 6.3 shows what the GNG looks like after placing each node at the source location that would match its audio (100D) coordinates best. The graph appears to be stretched vertically.

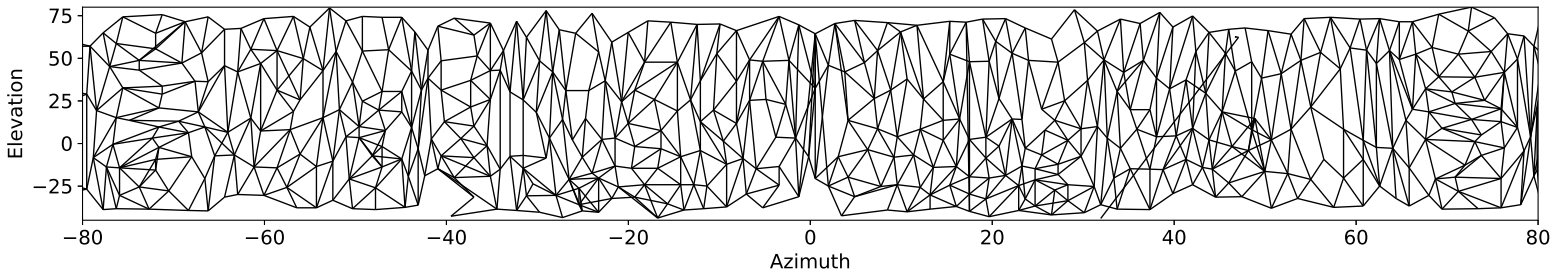


Figure 6.3: Auditory graph obtained from HRTF data. The 2D location of neurons is not known by the GNG, it has been interpolated from node coordinates in HRTF space, for visualization purposes only. Note that the x-axis and y-axis have different scales: the y-axis is compressed to make the tessellation more visible.

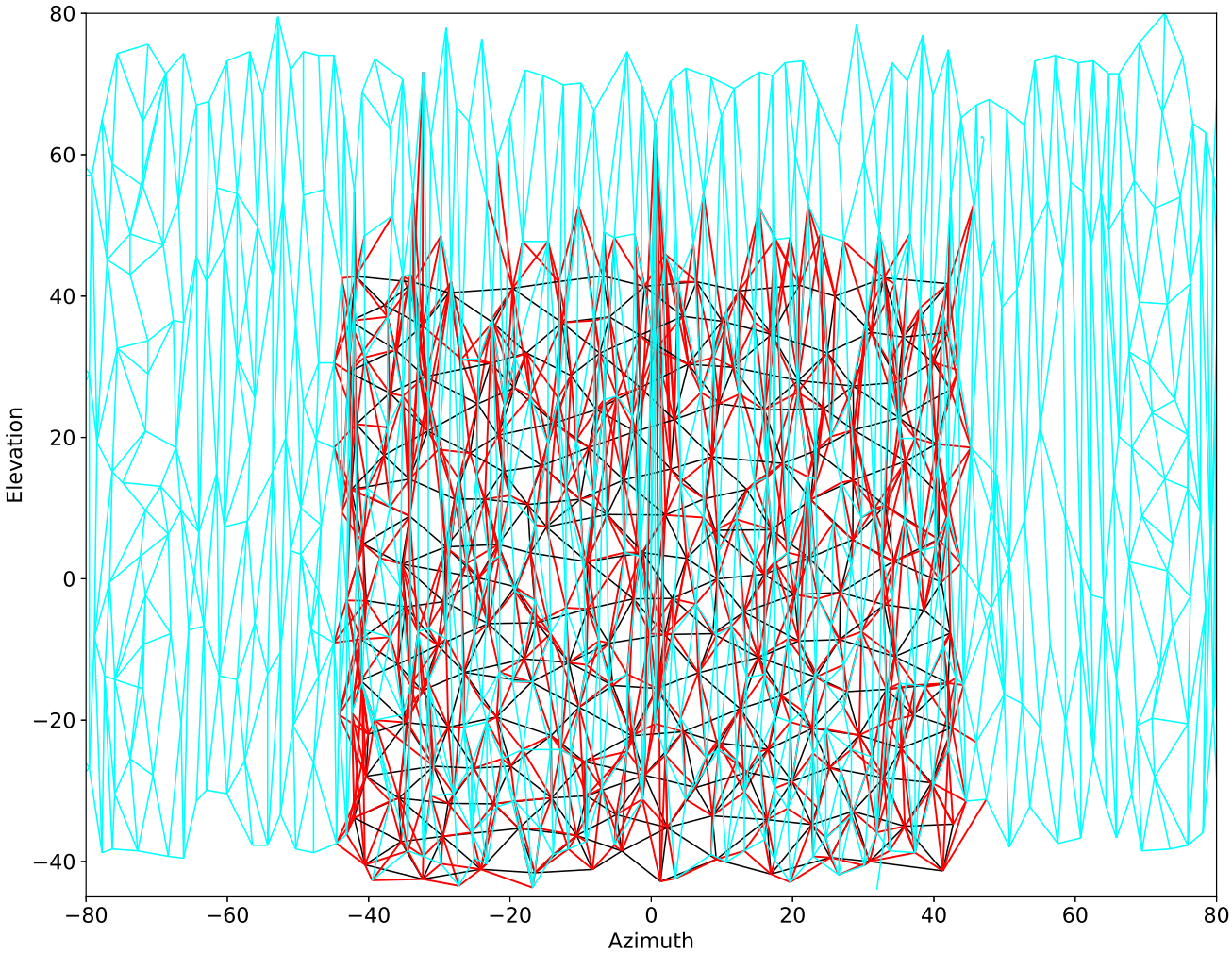


Figure 6.4: Bimodal graph obtained from HRTF data and regular 2D vision. Edges are colored depending on the modalities of the neurons they connect. Visual-visual: black. Auditory-auditory: cyan. Visual-auditory: red.

The graph obtained by connecting the GNG is given in figure 6.4. This time we use the first method (stimuli drawn randomly in the field of view), so a lot of auditory nodes are not connected to visual nodes, which is on purpose.

The same can be made using a 3D space for vision (adding depth perception for example). Since the visual space expands, and GNG are not advanced enough to reduce the dimensionality when the amount of possible inputs increases brutally, we also increase

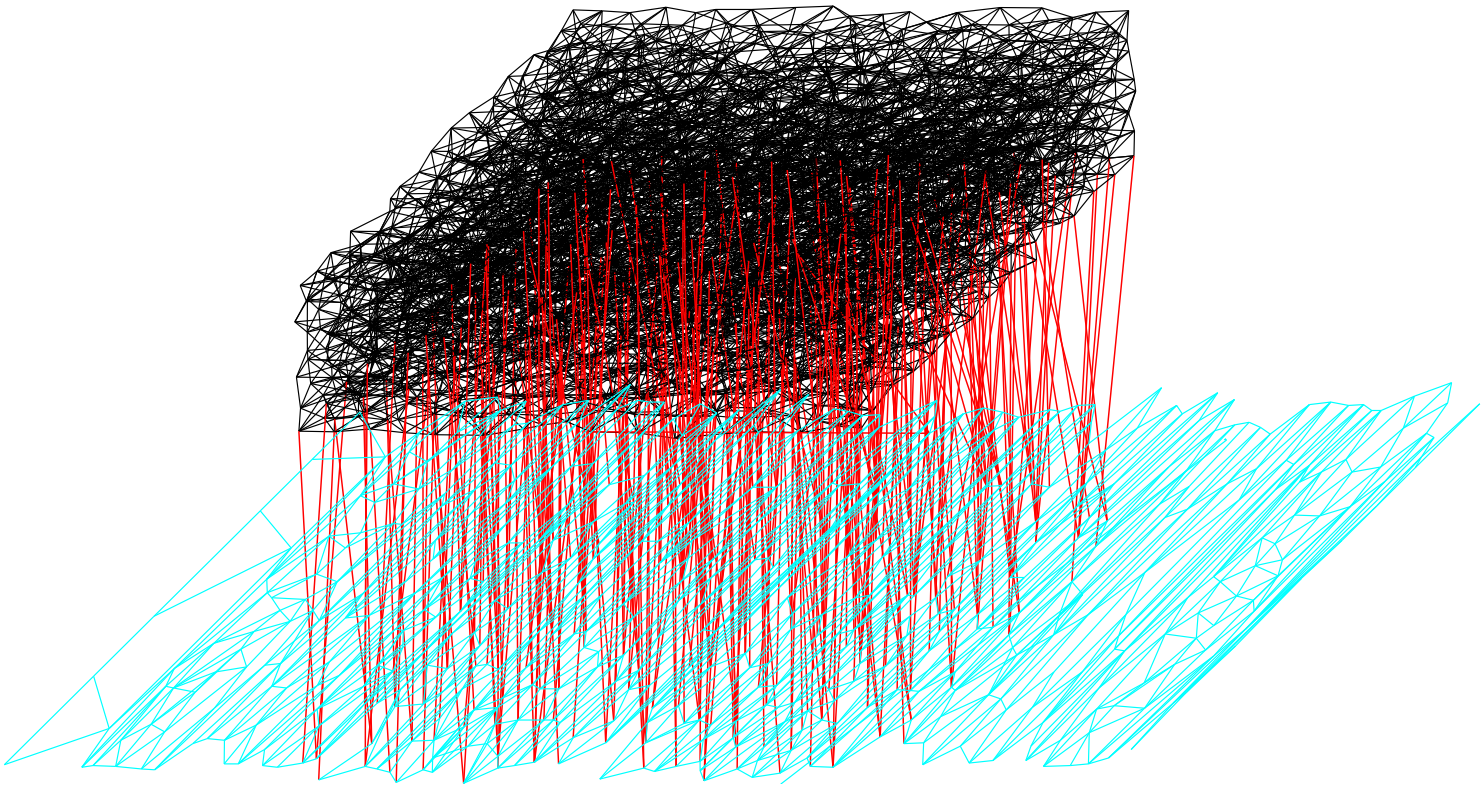


Figure 6.5: Bimodal graph obtained from HRTF data and regular 3D vision. For better readability, perspective is added, GNG are separated vertically, and visual-auditory edges connecting topmost visual nodes are hidden. Edges are colored depending on the modalities of the neurons they connect. Visual-visual: black. Auditory-auditory: cyan. Visual-auditory: red.

the number of neurons in the visual GNG to 3000. The resulting graph is given in figure 6.5.

6.5 Decision-making in multimodal topologies

This second section of results focuses on properties that remain, or appear, when using DNF for decision-making in the newly-created feature spaces. Input stimuli will be specified in each subsection, depending on the properties to showcase. For the same reasons, DNF parameters might need to be adjusted slightly from one setup to the next. All values are given in table 6.3.

6.5.1 Selection

This experiment is made on robot simulation (subsection 6.4.2). In order to test DNF selection, we put two separate stimuli A and B, separated both horizontally and vertically. Stimulus A has congruent audio and visual components, while B is not audible but visually more salient than A by 1 %. It is expected that A should be selected over B, as A is consistent over modalities. A and B have been picked arbitrarily to serve as an illustrative example, but the resolution is mostly the same in all the GNG. Results are synthesized in figure 6.6.

In the visual-only manifold, B largely takes precedence. A is mostly inhibited, with some (negative) residual activity left. This is expected, as B is more visibly salient, but it is worth noting that the 1 % difference between $\lambda_{\text{vis, A}}$ and $\lambda_{\text{vis, B}}$ matters. It is not shown

Table 6.3: Parameters used in our DNF implementation. Spread parameters are expressed in arbitrary unit that denotes the minimal number of edges that separate two neurons.

Parameter	Value			Meaning
	6.5.1	6.5.2	6.5.3	
				Simulation settings
Δt	0.01	0.01	0.01	Time step
σ_I	2.5	2/3/ . . . /20	2.5	Spread of stimulus
$\lambda_{\text{vis, A}}$	2	0	2	Strength of visual bottom/left stimulus
$\lambda_{\text{vis, B}}$	2.02	2	2.4	Strength of visual top/right stimulus
$\lambda_{\text{aud, A}}$	1.5	2	2.4	Strength of audio bottom/left stimulus
$\lambda_{\text{aud, B}}$	0	0	2	Strength of audio top/right stimulus
				DNF parameters
τ	0.1	0.1	0.1	Time constant
λ_+	0.55	1.1	0.4	Amplitude of lateral excitation
σ_+	1.5	0.2	2.5/3/3.5	Spread of lateral excitation
λ_-	0.3	0.9	0.3	Amplitude of lateral inhibition
σ_-	10	10	$+\infty$	Spread of lateral inhibition
h	-1	-1	-1	Resting level

in the figure, but we have tested swapping the intensity values, and A does take precedence in that inverted case. We are in a situation where both stimuli are considered equally by the DNF, and a very small difference in intensity is enough to bias the competition towards one or the other. This is a very standard observation in DNF literature, but it is still worth noting considering the topology is not entirely regular.

In the audio-only manifold, A is trivially selected, but we can see some loss of precision in elevation: the barycenter is found 7° higher than the actual stimulus. This is very consistent with the general lack of elevation-wise precision in auditory perception.

The precision is improved in the bimodal manifold. As would be expected, audiovisual congruent stimulus A is selected over visual-only B. But the barycenter is also closer to the actual stimulus position than in the audio-only case, meaning the visual elevation-wide better precision had a positive impact. Again, the enhanced multimodal precision is a classical observation in either neuroscience or machine learning, but it is worth noting that it persists when working with a complex underlying topology.

When we look more closely at the nodes around A, we can see than despite there being a lot of edges in all directions, a few neurons form a discernable bubble. It is interesting that these neurons come indiscriminately from both modalities. One could have feared an outcome where only visual neurons interact with each other, and auditory neurons, less regularly distributed, only serve to transmit a little bit of auditory stimulation. On the contrary, the crossmodal connections play an important part, so that the DNF does not leave out one modality for the other. When both are useful, both are used.

Selection with a superfluous dimension We redo the same experiments, but this time with sensory vision in 3D (adding non-relevant depth). Stimuli A and B are given the same depth, so that their distance in the external environment remains the same as before. According to our preliminary tests, the conclusion would be the same with stimuli of different depth. Results are displayed in figure 6.7. Only the visual-only and audiovisual conditions are shown, since the auditory-only condition is the same as before,

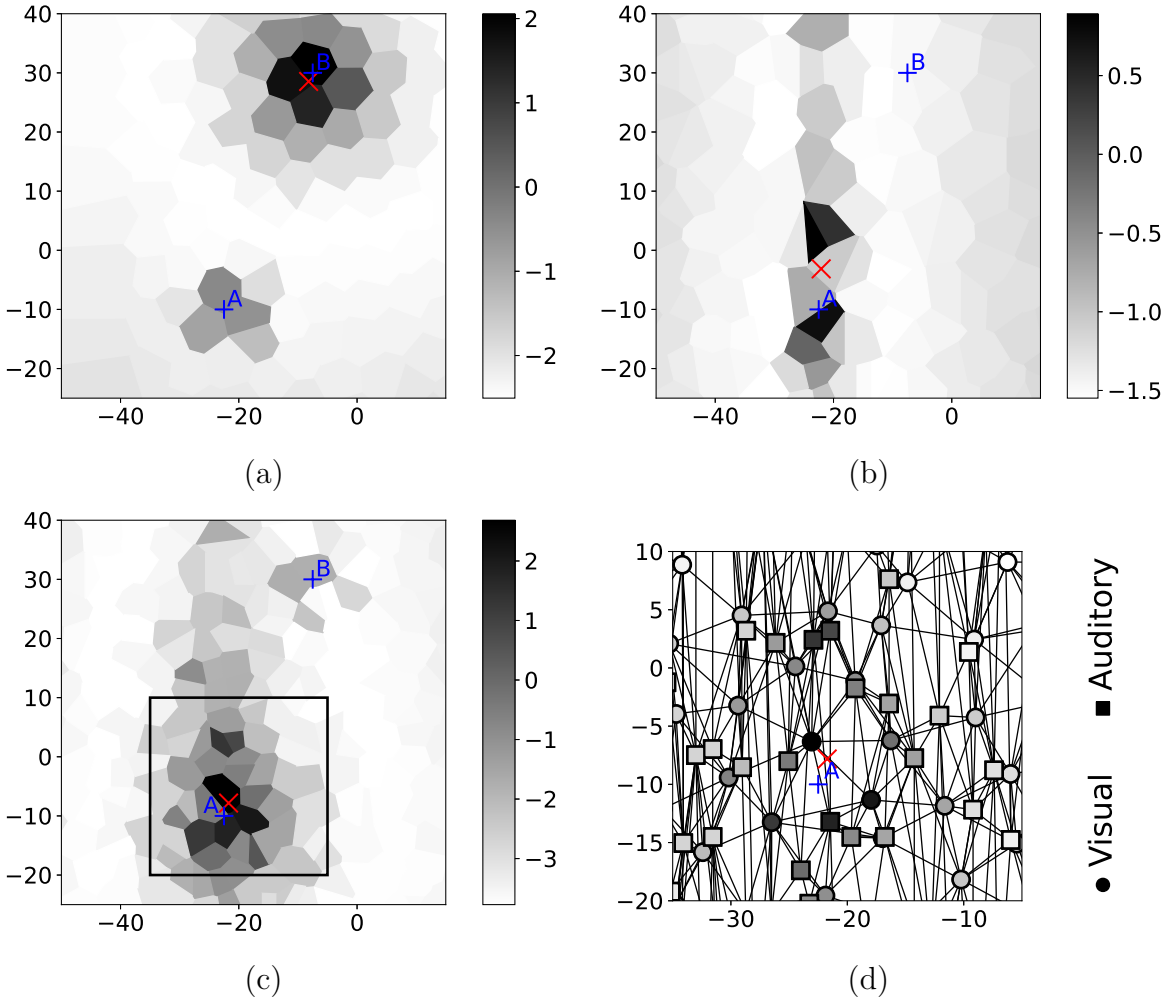


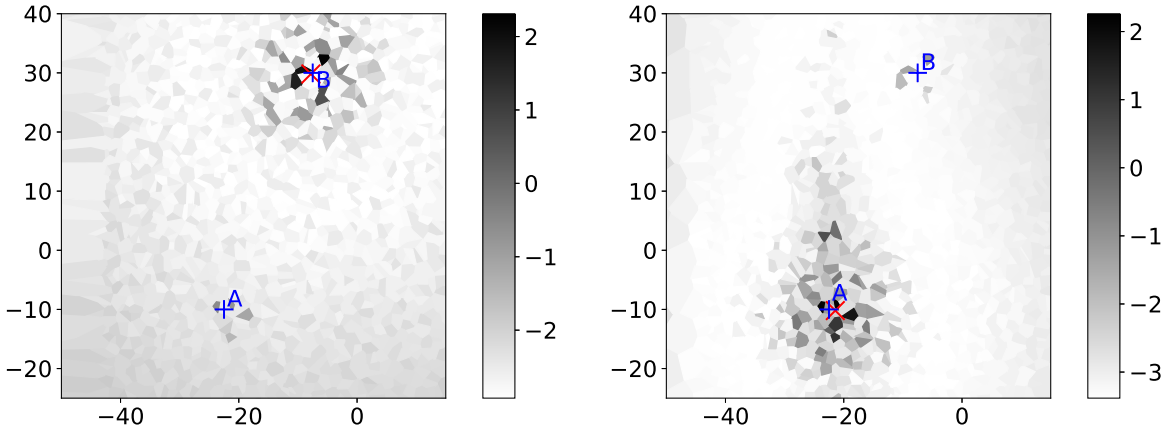
Figure 6.6: Results of stimulus selection by DNF in unimodal and bimodal GNG. These 2D depictions use neuron positions interpolated from the source data (for visualization purposes). Shades of gray reflect neuron potential U. Red crosses indicate the barycenter of output activation $f(U)$ in the reconstructed 2D projection. (a) Visual-only neural gas with two stimuli located at A and B, with B slightly more salient. Nodes are represented by Voronoi cells, edges connecting nodes are not represented. (b) Auditory-only neural gas, with only one input at A. (c) Bimodal neural gas. Its input is the sum of the ones used for (a) and (b). (d) Zoom on (c) around A, where all nodes and edges are represented.

and the zoom-in picture with edges is hardly readable. As a reminder, the visualizations are still made using x- and y-axes, meaning the new depth axis is completely flattened. These presentations are akin to looking at a cube from a side, hence the dense Voronoi tessellation and the scattered activity.

We find that the outputs are strikingly similar, i.e. a preference for multimodal consistent inputs and improving audio precision, despite a big increase in the number of neurons and edges, many of which are irrelevant to the task. This shows robustness of the model to distracting dimensions.

6.5.2 Merging incongruent stimuli

In previous experiments, we tested selection between one unimodal and one bimodal stimulus. Another experiment we can do in the same (2D) robotic setup is of selection between two conflicting unimodal stimuli, similarly to the ventriloquist effect. Here, we take an auditory stimulus A and a visual stimulus B that are spatially discrepant.



(a) (c)

Figure 6.7: Same as figure 6.6 with a supplementary dimension in the visual modality. The third dimension should be orthogonal to the plane used in this representation, and is flattened here.

Note that this setup is not viable for reproducing the ventriloquist effect as in chapter 5, for multiple reasons. 1/ This neural map is far less dense than the one used before. In order to retain the same degree of interaction between processing units, we would need far more nodes in the GNG. 2/ When we first modeled the ventriloquist effect, the precision in each modality could be controlled through stimulus size, and stimulus size only. Here, the precision also depends on the number of nodes in each GNG (fixed arbitrarily) and topology irregularities (out of our direct control). 3/ Running DNF in GNG is more computationally expensive than in regular lattices, since the usual efficient methods for computing convolutions are not available here. Running a series of DNF iterations with added noise would become quickly expensive.

Beside these technical issues, we expect that we could achieve similar qualitative results to chapter 5, with one caveat: When the topologies were chosen previously, they were homeomorphic to a rectangular grid, so all units had exactly 4 closest neighbors. With GNG, there are more connections, so the parameter space might change significantly. In particular, one would need less excitatory amplitude per unit. Also, the stochasticity in GNG creation would definitely increase the variance in signal localization.

Nevertheless, we can still test DNF selection by changing stimulus size, knowing that there remains an unquantified factor of topology shape and resolution. In this subsection, stimulus spread σ_I is split into modality-specific parameters $\sigma_{I,\text{aud}}$ and $\sigma_{I,\text{vis}}$, and takes integer values between 2 and 20. For all sets of values, we check whether the output of DNF comes closer to A (audio stimulus) or to B (visual stimulus). The results are given in table 6.4. Experimentally, the output always clearly favors either A or B, no in-between.

We can see a tendency, where the auditory stimulus takes precedence when it is more precise than the visual stimulus (figure 6.8, left), and vice versa (center). The separation between priorities is not entirely linear, which is to be expected given the irregularities in the topology. Some side-effects appear for very high values of σ_I, they are due to both stimuli being spread exaggerately and meeting at a random node (figure 6.8, right).

We can guess from figure 6.8, and from the low value of σ_+, that DNF form multiple very thin peaks before one takes over and inhibits the rest. That means than the more a stimulus is spread, the more peaks it will create in the competition phase, thus more precise stimuli have less internal competition, so less local inhibition, and are selected more easily. This is a consistent, albeit crude, way of implementing a ventriloquist effect. Cases of interpolation when modalities have equal reliability do not appear here, because

Table 6.4: Closest stimulus (A auditory or B visual) to the DNF output, depending on stimulus spreads $\sigma_{I,\text{vis}}$ and $\sigma_{I,\text{aud}}$

$\sigma_{I,\text{vis}}$	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20
20	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
19	A	A	A	A	A	A	A	A	A	A	A	A	B	A	A	A	A	A	A
18	A	A	A	A	A	A	A	A	A	A	A	A	B	B	A	A	A	A	A
17	A	A	A	A	A	A	A	A	A	A	A	B	B	B	B	A	A	A	A
16	A	A	A	A	A	A	A	A	A	A	A	B	B	B	B	B	B	B	B
15	A	A	A	A	A	A	A	A	A	A	A	B	B	B	B	B	B	B	B
14	A	A	A	A	A	A	A	A	A	A	B	B	B	B	B	B	B	B	B
13	A	A	A	A	A	A	A	A	A	A	B	B	B	B	B	B	B	B	B
12	A	A	A	A	A	A	A	A	A	B	B	B	B	B	B	B	B	B	B
11	A	A	A	A	A	A	A	A	B	B	B	B	B	B	B	B	B	B	B
10	A	A	A	A	A	A	A	A	B	B	B	B	B	B	B	B	B	B	B
09	A	A	A	A	A	A	A	B	B	B	B	B	B	B	B	B	B	B	B
08	A	A	A	A	A	A	B	B	B	B	B	B	B	B	B	B	B	B	B
07	A	A	A	A	A	B	B	B	B	B	B	B	B	B	B	B	B	B	B
06	A	A	A	A	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B
05	A	A	A	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B
04	A	A	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B
03	A	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B
02	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B

$\sigma_{I,\text{aud}}$

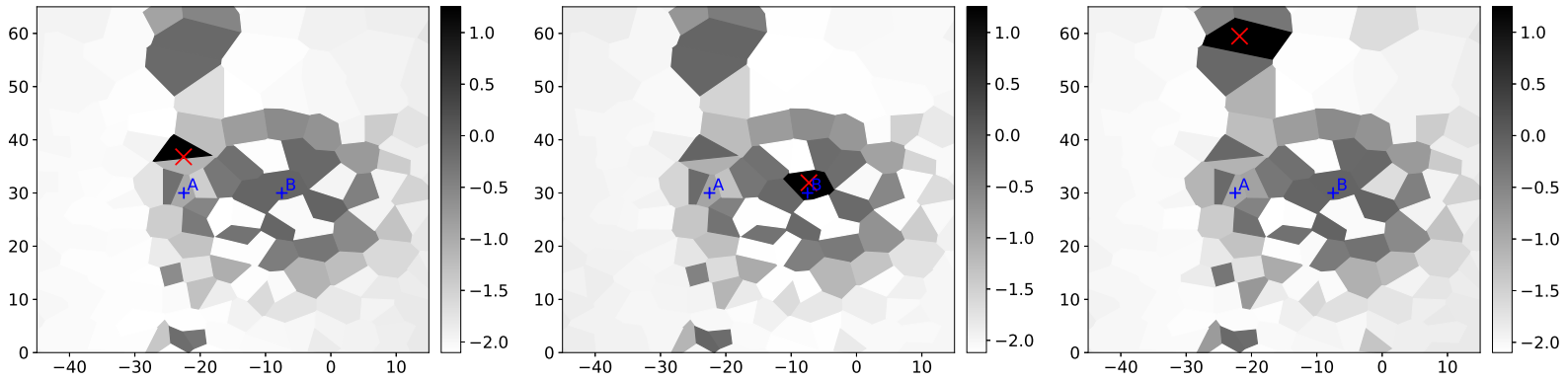


Figure 6.8: DNF activity in bimodal GNG given incongruent audiovisual stimuli (A auditory, B visual). In all three plots, $\sigma_{I,\text{vis}} = 17$. Left: $\sigma_{I,\text{aud}} = 12$. Middle: $\sigma_{I,\text{aud}} = 13$. Right: $\sigma_{I,\text{aud}} = 18$.

processing units are spread too sparsely. The chances for a node in the middle to be selected are quite low.

6.5.3 Effect of modality resolution

This experiment is made on SC simulation. We are interested in seeing what a DNF would select when confronted to conflicting bimodal stimulus depending on modality reliability. It is expected that near the fovea, vision is more reliable, so it should have a bigger weight in the fusion than audition. To test this, we put two separate stimuli A and B at a common azimuth x, and elevations $-5°$ and $5°$ respectively. Both stimuli can be both seen and heard, but A is 20 % more auditively salient than B, and B is 20 % more visually

salient than A.

When tested on a unimodal manifold, the DNF has no trouble selecting either A or B. Every time, the most salient stimulus in its respective modality has a higher chance of being selected. Occasionally, the DNF forms a bubble in-between the stimuli. This is mostly visible for higher azimuths in the visual GNG. The reason is that the resolution is so low that A and B are separated by only a few edges. The DNF does not have access to the corresponding inputs of its neurons viewed from the exterior. Thus, when viewed from inside the model, they are topologically very close to each other. So, the DNF treats the stimuli as if they were right next to each other, and merges them into a bubble of activity located at their center of mass.

In the bimodal manifold, the stochasticity in the creation of the GNG starts having an impact, as it may seemingly give a locally higher resolution to a modality when it is not expected. A might be selected instead of B, when B is more salient, just because B stimulates a region with fewer neurons or connections than average. To separate the random effect caused by the creation of the GNG, we create 50 bimodal manifolds, and test a run of DNF on 90 different azimuths for each of them. The results are aggregated in figure 6.9. As we suspect that the distance at which stimuli are merged depends on the width of the DNF kernel, and σ_+ in particular, we couple in our analysis the effect of resolution with the value of this parameter. We test three different values of σ_+ (table 6.3), represented by three different colors: green, red, blue from thinnest to widest.

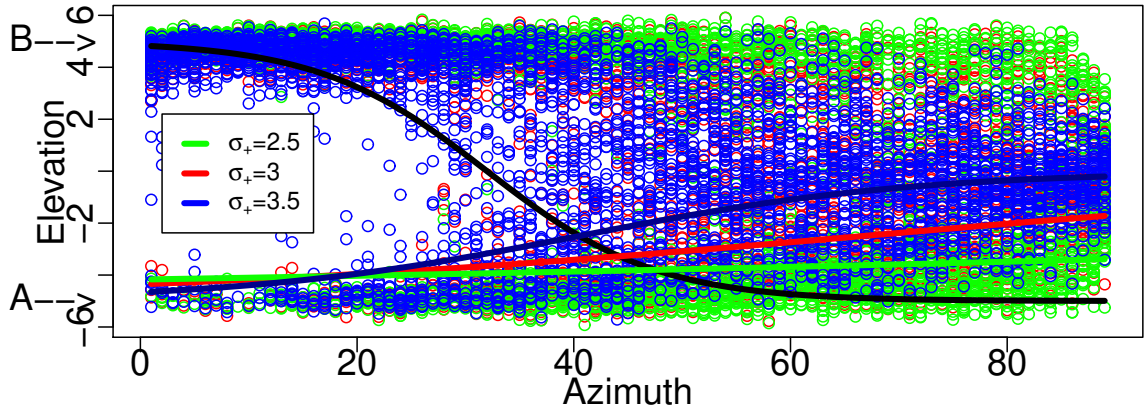


Figure 6.9: Statistical model of the modality priority change (in black) and the stimulus merging. One point represents the barycenter of the output of one of the 3 differently parametrized DNF (green: $\sigma_+ = 2.5$, red: $\sigma_+ = 3$, blue: $\sigma_+ = 3.5$), on one of the 50 randomized GNG, with two bimodal stimuli A and B at variable azimuth and elevations $\pm 5°$. The black curve shows a logistic regression of the switch between preferred stimuli. Colored curve show logisitic regressions of the stimulus merging effect depending on values of σ_+.

The curves represent the outcome of two mixed logistic regressions. The fit of the black curve is obtained after cancelling the merging effect, and shows a clear switch of preference from B to A centered on 32°. B is more likely to be selected than A when the visual modality is the most reliable, and vice versa. Logically, this effect is independant of σ_+ variations. The effect can be interpreted as the DNF automatically selecting a stimulus according to the most reliable sensor.

The fit of the colored curves are obtained by canceling the switch effect. We can see a convergence from $\pm 5°$ to 0° elevations, although for lower values of σ_+, the limit at 0° is

not reached before the end of the field of view. Only the lower curves are displayed but the effect is symmetrical.

The results show two trends. First, from the higher concentration of points at the 5° elevation in the leftmost part of the figure, we can see that B (visually stronger) is more often selected in lower azimuths than A. Then A is preferred for higher azimuths. Second, we see that the probability of A and B being merged (manifesting as an increasing concentration of points around 0°) increases with the azimuths. As we expected, the distance at which they are merged depends a lot on the value of σ_+. The larger the interaction kernel, the sooner the merging seems to happen.

6.6 Conclusion

Our model consists in two unimodal GNG, trained using the standard algorithm by [Fritzke, 1995b], then connected to form one new multimodal manifold. This manifold serves as a support for neural dynamics that are implemented according to the DNF paradigm [Amari, 1977] that we adapted for this purpose.

These DNF come with a wide range of properties. Focusing on multimodal fusion, we have showcased a sample of it, with a robust selection of stimuli, using the best information each modality had to offer, and filtering out irrelevant information. Additional properties could be explored in future works, such as memory, noise filtering, or conditional selection (e.g. by lowering excitation thresholds and adding a pre-stimulation to neurons coding for the red color, so that the robot only selects red objects). But we have shown that the interesting properties for multimodal fusion and selection were still available in a learned manifold.

Indeed, the main novelty of our work can be seen from two aspects. On one side is the use of neural dynamics in a multimodal manifold of unspecified dimensionality or regularity, a capability of DNF that has not been showcased before. The field applies on a learned manifold that is faithful to each unimodal sensory space, and is not hindered by irrelevant dimensions. On the other side is the creation of a multimodal topology, where the contribution of different modalities depends on their respective resolution, despite the manifolds being weakly constrained by a light algorithm. And this is made possible by the selective properties of the DNF.

Perspectives As we have seen when adding a dimension, the number of neurons in the GNG necessary to keep the same resolution, and consequently the computational cost of the model, may increase drastically when the sensory space is broadened. This would not be an issue with deep neural networks, that are very effective at finding intrinsic dimensions in data [Ansuini et al., 2019]. It would be interesting to see whether manifolds created by deep learning are also suitable vectors of neural dynamics. In active tasks involving active selection, DNF may well be seen as a lighter version of recurrent neural networks with a fixed convolution kernel. So, one could consider feeding into it the hidden neurons of a deep network, which contain a lower-dimension projection of the input sensory space. The only necessary modification would be to add connections between topologically-close neurons of the network.

With GNG, one may wonder whether manifold learning and neural dynamics can be done simultaneously. This could have tremendous applications in robotics, where hardware constraints mean that a model learnt on a robot may not function properly with another. We believe that DNF could actually help with the manifold learning, as they may help focus on relevant signals and ignore distractors. However, we foresee a few obstacles. First, GNG may not be adapted to online learning, as it assumes that each

new input is randomly drawn from an already known distribution, which we would not have here. Second, the parametrization of DNF is not trivial and may need to be slightly adapted on-the-fly as the network grows. We did a first pass on this in this chapter, but we did not have to consider the time-dependance of DNF with regards to the evolution of the manifold over time. This could be a challenge.

Even if the dynamics are not mixed with the learning, they still offer useful perspectives. For example, DNF have been suggested to model saccades [Quinton and Goffart, 2018]. Such active perception could have lots of uses. A robot could use head movements to place itself in the position where it perceives a selected signal best, like a human putting an ear forward to listen carefully. Or it could test gaze shifts to explore parts of its sensory space where its manifold is underdeveloped. Just like we cumulated manifold for sensory modalities, we believe a motor manifold could be learned. But instead of linking nodes from different modalities, one would have to link directed edges between sensory manifolds and the motor manifold. In the end, a movement of an object in the environment could be seen as an activity shift from a sensory node to another, which would be linked to the shift in motor positions that compensates it [Laflaquière et al., 2018]. Inversely, a saccade would be a path in a multisensory manifold from the most activated node to another, translated as a path in motor positions. The advantage of this workflow is that preliminary knowledge of the environment (e.g. to compute a barycenter of activity, like we did) would never be required at all.

Part III

Discussion

Chapter 7

Topologies and the burden of uncertainty

Contents

7.1 Introduction

Topologies are not usually the main focus in artificial intelligence (apart from manifold learning, although it is often more about dimensionality reduction than topologies). To put it bluntly, if you look at a localization task in 2D, you will often happen to find a convenient 2D map in which all informational units happen to respect a regular rectangular structure with regular spacing between units (i.e. a rectangular lattice). This is all the more convenient when you receive data from a 2D camera with regularly-spaced pixels of (allegedly) equal reliability (figure 7.1a). Unfortunately, this is not always the case. Cameras can show traces of wear or smears, and pixel attacks, to pick one example, show that it does not take much perturbation to completely overturn results from modern computer vision algorithms [Su et al., 2019]; and multimodal fusion does not get a preferential treatment. Take the Perceiver model [Jaegle et al., 2021], a multimodal extension of the Transformer (the current standard in many deep learning applications): inputs from different modalities are conveniently fused before training (figure 7.1c). This is surprising, since one of the main motivations behind multimodal fusion is for modalities to complete and enhance each other. That concern is usually addressed by fitting more parameters in a neural network model. But putting the actual maps into question is not a standardly explored research path.

The work in this thesis did not start with the aim to challenge the ground space of multimodal merging in artificial intelligence. But as the first contributions unfolded, so did the realization that topologies were vital to sensory integration — and often overlooked. The world does not hold sparse information, at least not on a perceptible level (figure 7.1b). Directions are not left *or* right, colors are not green *or* red… And sensors do not sample it sparsily either. What good is an algorithm giving unitary responses to unitary inputs? We get a list of classes that can be inferred from a list of pixels, whereas the most useful information may actually lie in the gaps between sensory samples (figure 7.1d). How wide the gaps are tells us about the reliability of the sensors, the edge

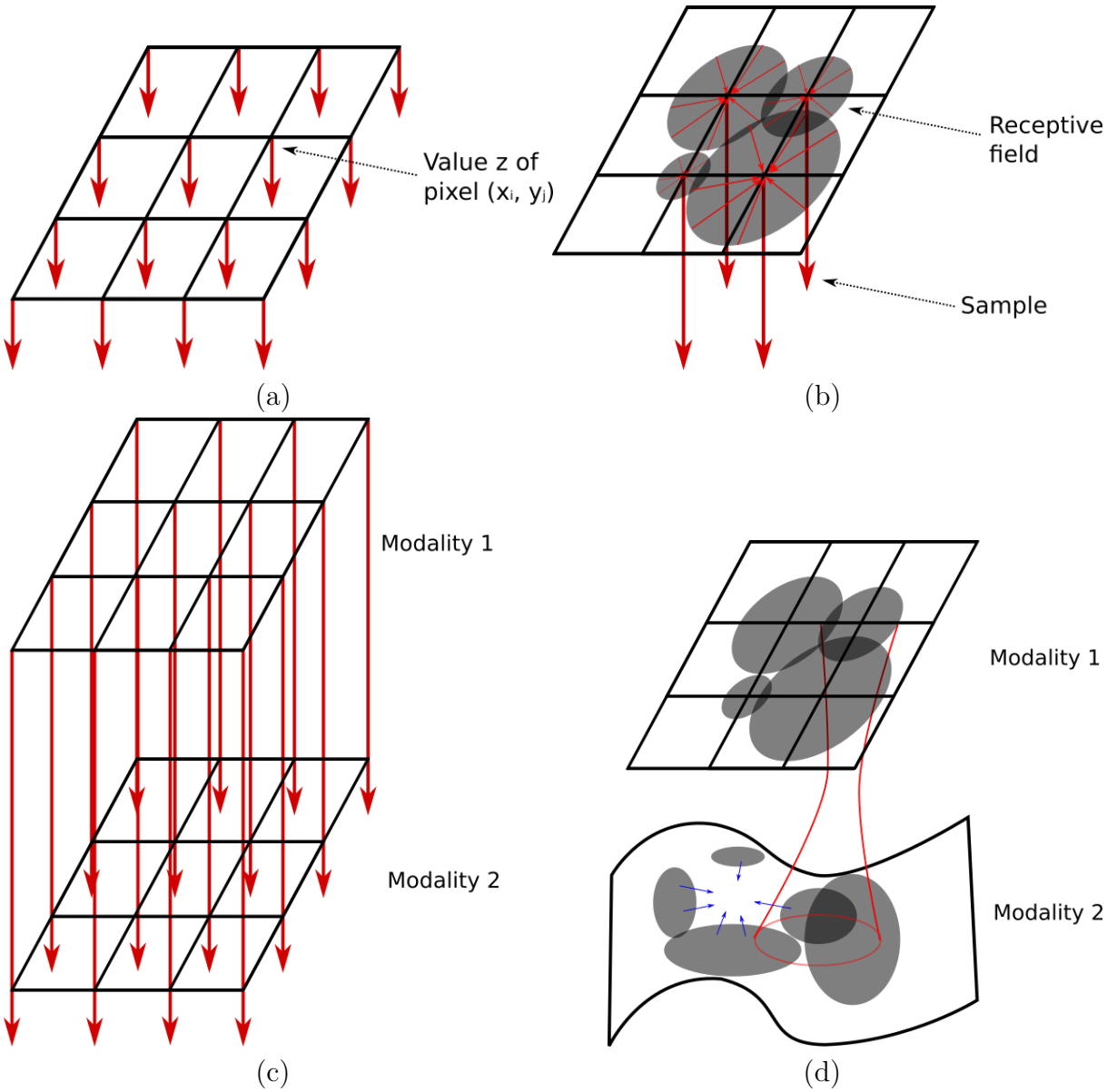


Figure 7.1: (a) Unimodal computational regular topology with discrete sampling. (b) Unimodal neuro-inspired topology projected onto a neural lattice. Each sample reflects an irregular aggregation of topologically-organized information about the world, and formed somewhere during sensory capture and processing (e.g. retina blur or receptive field). (c) Connection of multimodal maps where some degree of alignment between maps is assumed. (d) Connection of multimodal maps, having irregular topologies that do not necessarily align. A modality might have missing information (blue arrows), possibly leading to compensation mechanisms (exploiting intramodal and/or crossmodal information) or exploratory behaviors (e.g. saccades).

between exploitation and exploration, and the actions to undertake next. Only after the gaps (or regularities) have been properly taken into account, can we deal with feature integration.

The questions regarding topologies exceed by far the field of multimodal fusion, but this is the entry point that we have taken in this thesis, so let us review our contributions so far.

7.2 Contribution summary

In chapter 4, we reviewed a representative sample of decision-making algorithms and unified them in a common formalism. The taken point of view was meant to encompass multimodal merging among other tasks (i.e., what response to give to stimuli of contradicting modalities), having for objective to unify methods across fields ranging from neuroscience and psychology to robotics. Available properties include temporal dynamics (DDM and their family), spatial filtering (WS, FL), or both (KF, DNF). In our benchmark, DNF appeared to be the most versatile, but sensible differences could be observed in the way algorithms modeled uncertainty. Models like KF (or its cousin MLE) consider uncertainty as a statistical measure that reflects how reliable an observation is with regards to the perceptory noise of its respective modality, and how reliable its answer is. DNF do not represent uncertainty explicitely, as noise is directly integrated dynamically, as if it was an additional input made of summed perturbations from sensory and computational artifacts. So, the burden of representing modality reliability is left to the expert design of inputs, as well as the estimation of output uncertainty. From there, two challenges arise. First, assuming expertly-designed multimodal inputs, are DNF sufficient to achieve multimodal merging behaviors satisfying both neural modeling and artificial applications? Then, if we were to put aside expert supervision and put the burden of uncertainty back to the input space where it belongs, are DNF still capable of merging multimodal inputs in it?

Chapter 5 illustrates an answer to the first challenge, using the ventriloquist effect as a benchmark. This effect of audiovisual capture is reproduced qualitatively with a DNF, with results quantitatively comparable to MLE, the dominant paradigm in psychophysics. Again, it is worth noting that DNF model operate fusion at a different level than MLE. The latter treats the ventriloquist effect as a probability computation — the probability of an audiovisual stimulus being perceived at a given location knowing the psychometric function of its unimodal components. DNF model individual decisions, the aggregation of which fits the probability distribution psychophysically observed. This contribution shows the latitude in DNF parameter fitting, with graphical analyses that are quite novel to the literature of this paradigm. But the exploration of irregular topologies — this time through a logpolar transformation inspired from the superior colliculus — is not yet probant. Our hypothesis is that this limitation stems more from the data (no elevation and little excentricity) than the model.

The adaptation of DNF to irregular topologies is directly challenged in chapter 6. In this contribution, new topologies are learned through the use of GNG, a self-organizing method where the size and structure of the created manifold are left open. Unimodal topologies are created from sensory space, such as a promising 2D-like map from HRTF. Then multimodal topologies are obtained by connecting neurons from a GNG to another. Most importantly, behaviors that were linked to DNF in classical regular maps were reproduced in these new topologies. This opens the door to many developments that are discussed in the next chapter, but before that, let us fold back on the justifications behind the use of DNF.

7.3 Why dynamic neural fields?

This thesis has presented DNF as one decision-making algorithm applied to multimodal merging. We want to highlight that there is more to it: DNF make sense for multimodal merging specifically.

Perceptual decision-making mainly involves two processes: a bottom-up process in which the position and saliency of stimuli influence the decision, as we described in chap-

ter 4; a top-down process, in which pieces of information such as a given task are taken into account, and may play into which kind of algorithm is followed (from WTA to WS and everything in-between). Multimodal merging could be classified as a special type of perceptual decision-making, as it adds a new intermediate factor: reliability. Multimodality implies (not exclusively) that the inputs on which a decision is made have different natures, and thus, different reliabilities. Experiments on ventriloquism showcase an effect of stimulus reliability[1] on localization, relegating the known effects of saliency or top-down instructions to the background. And the former effect can be accounted for in three ways:

1. Reliability as an input. This is the starting point of MLE models: reliability is known as the inverse variance of a psychometric function, and the average estimate of the multimodal percept is computed from it. A limitation is that this model only fits to aggregations of trials, not explaining how each single one is generated. Whether the brain actually encodes variances is even disputed.

2. Reliability as a by-product of neural interaction. This is how DNF emulate the ventriloquist effect: with a small excitatory range, areas simulated by larger signals are locally submitted to more internal competition, putting them at a disadvantage when opposed to areas receiving a thinner stimulus. Whether this is how the brain compares reliabilities would be difficult to prove, as it would require measuring precisely the amount of inhibition sent between cortical microcolumns.

3. Reliability as a by-product of topologies. The way sensory inputs are projected into multimodal maps can partially be linked to the resolution of the sensor that brought them (see SC). So, comparing reliabilities amounts to comparing resolutions in sensory maps. This sounds quite plausible, and a proof of concept for a computational model is given in section 6.5.3.

In the human brain, there might very well be a bit of truth to these three hypotheses. And separating them should not be an easy task. In psychophysics, e.g. to study the ventriloquist effect, one can expect effects from stimulus saliency (brightness of visual spot, loudness of sound), task (locate the spot or locate the sound), sensory resolution, stimulus precision (spot width and blur), and perhaps also a modality-wide bias (getting used to trusting localization from the eyes). Untangling these effects empirically could prove very difficult.

From a computational perspective, however, most of these are achievable. The bottom-up process can be realized by DNF. We have seen in chapters 5 and 6 how DNF could play with stimulus precision and differences in sensory resolution. Top-down processes may not be implementable by one DNF alone, but a combination of them might be suitable, with different fields handling different memory, selection and interpolation subtasks. The only seemingly-unfeasible part would be the Bayesian idea of computing variances explicitely, which is only necessary if you adhere to a strict view of variance being encoded as a representation itself, and not a statistical expression of underlying neural interactions (that DNF may very well be a model of, cf. discussion in chapter 5).

So maybe this is what multimodal perception should look like: a myriad of processing modules connected and feeding into one another; some doing bottom-up filtering, some

1. Stimulus reliability should not be confused with modality reliability. For example, within vision only, a stimulus seen from the center of the retina is much more reliable than a stimulus seen at the edge of the field of view. So, it is safer to speak of stimulus reliability only. When we say that in spatial localization tasks, vision is often more reliable than audition for humans, it means that most visual stimuli are more reliable than most auditory stimuli.

comparing reliabilities, some injecting top-down features; some driven by DNF-like mechanisms, and possibly some by something else. Decision-making in the brain results from complex pathways which would be extremely hard to map in their entirety. Computational models can reproduce this at a lower scale. And this is where attention comes in. With all these modules in interaction with each other, there is too much information in the world for all of it to be treated in real time. Bottom-up attention (a seminal property of DNF) can help filter out distracting stimuli before they are processed by the rest of the system. Top-down attention may guide lower modules for more efficiency (and could be implemented by one DNF temporarily memorizing relevant features and sending excitation or inhibition to down-level DNF depending on whether or not they match these features[2]).

One of our first assumptions was that (multimodal) perception had to be active, and attention is an important part of it. It is convenient that DNF are very well suited for this, but that is not a coincidence: it is one of the reasons we picked this paradigm in the first place. That being said, we have not implemented the bigger part of active perception, which is to generate eye and body movements. Some perspectives on this are discussed in the next chapter.

2. One could think of this as a way to implement "levels of interest" towards different objects, as teased in chapter 1.

Chapter 8

Perspectives for active multimodal perception in robotics

Contents

8.1 Introduction

In this section, we propose some interesting research paths to further anchor our work in the field of active perception. We have proposed using DNF as a tool for covert attention, now we envision implementations of overt attention, using eye movements as a guiding example.

8.2 Active perception

One way to materialize active perception in DNF is to add predictive aspects in the model. In [Quinton and Girau, 2011], predictors pre-stimulate areas in which targets are expected to move. This can be extended to simulate neural excitation build-up in preparation for saccades [Quinton and Goffart, 2018].

Alternatively, one could define a set of rules dictating actions depending on DNF activity. GNG as used in chapter 6 make a good illustrative example. Previously, we have learned a topological map made of neurons that had a prototypical input in sensory space. The same way we created an audiovisual topology for robots, we could create a motor map, learned either from the set of motor commands or via proprioceptive feedback. Some inspiration can be found from the motor functions in the SC [Gandhi and Katnani, 2011].

Suppose there is a way to link directed edges in audiovisual topology to directed edges in motor topology — we do not have an implementation here, but we could imagine matching shifts in sensory perception to the articulary shifts that compensate them, as suggested in [Laflaquière et al., 2018]. For example, a shift in visual space towards one degree to the right can be compensated by an eye rotation of one degree to the left. Then we have a straighforward way of issuing motor commands: given a fixation unit (typically, the center-most neuron in the GNG), a saccade towards a stimuli can be given by the path from the best-activated unit (the neuron with highest potential in the DNF) to the fixation unit.

Proof of concept We do not pretend to have a full implementation of sensorimotor pathways, but we present a short demonstration here. For simplification, we take the 2D regular visual GNG from section 6.5 and assume having an exhaustive knowledge of the motor commands allowing shifts in input from one node to another. We start with a stimulation at the center of the field of view, attended by the same DNF as in section 6.5. The best-activated unit in this fixation phase is named n_{center}. We implement the following rule: "Every time the best-activated unit n_{best} changes, make a motor shift from n_{best} to n_{center}." Evolution of activity in two makeshift scenarios is shown in figure 8.1.

The left scenario illustrates a case of smooth pursuit. As the visual target slowly moves to the right, very small movements to the right are generated, pushing the target back to the center of the field of view. In the right scenario, the target moves abruptly to the right, leading to a saccade once the DNF has created a new peak. Interestingly enough, due to the low resolution of the visual feature map, the target does not arrive on n_{center} directly, causing a small corrective saccade later.

This is a very simple demonstration which is not meant to accurately represent the staggering complexity of eye movements as seen from neuroscience, but this is a good start. There could be room for making a realistic computational model of saccades in the SC, using the SC-inspired GNG of section 6.4.1, with the heterogeneous resolution potentially explaining switches between smooth pursuit and saccades, micro-saccades, corrective saccades, etc. Or practical explanations could be explored in robotics using tailored sets of rules.

8.3 Towards embodied cognition

In this section, we suggest perspectives for more advanced extensions of our work, reaching into the field of embodied cognition. These extensions, while very diverse, could in theory be combined. The common idea is to use the focus properties of DNF to enhance the learning of topologies, exploiting the constraints and regularities of the body of the learner (something we already started doing on a small scale with regular GNG).

DNF activation as a growing criterion A new step towards online topology learning would be to use the filtering and stabilization properties of DNF to learn GNG from temporally continuous data. Regular GNG are not suited for online learning, as successive inputs are usually spatially close to one another, meaning the graph will be easily attracted by recent presentations, and older presentations will be forgotten. DNF could be used to determine when to update GNG (when a peak of activity reaches a certain threshold), and/or to limit spatially the scope of the updates (using DNF activity as a pooling function). This method raises new challenges, either in the dynamics of the system (the GNG and DNF have two different time constants) or in the changing spatial resolution on which DNF have to be calibrated.

Modalities linked by DNF co-activation This would be a dynamic interpretation of the Hebbian rule (cells that fire together wire together). The idea is to run a DNF for selection on each modality, and once peaks of activity are formed on two modalities or more, connect neurons activated at the same time. This could be made simultaneously to GNG learning, allowing feature maps to calibrate on each other. Challenges similar to the previous point may be faced.

Predictive learning Once rules for DNF-led actions are settled, it would be interesting to see how this could affect GNG learning. One method, making use of predictive coding,

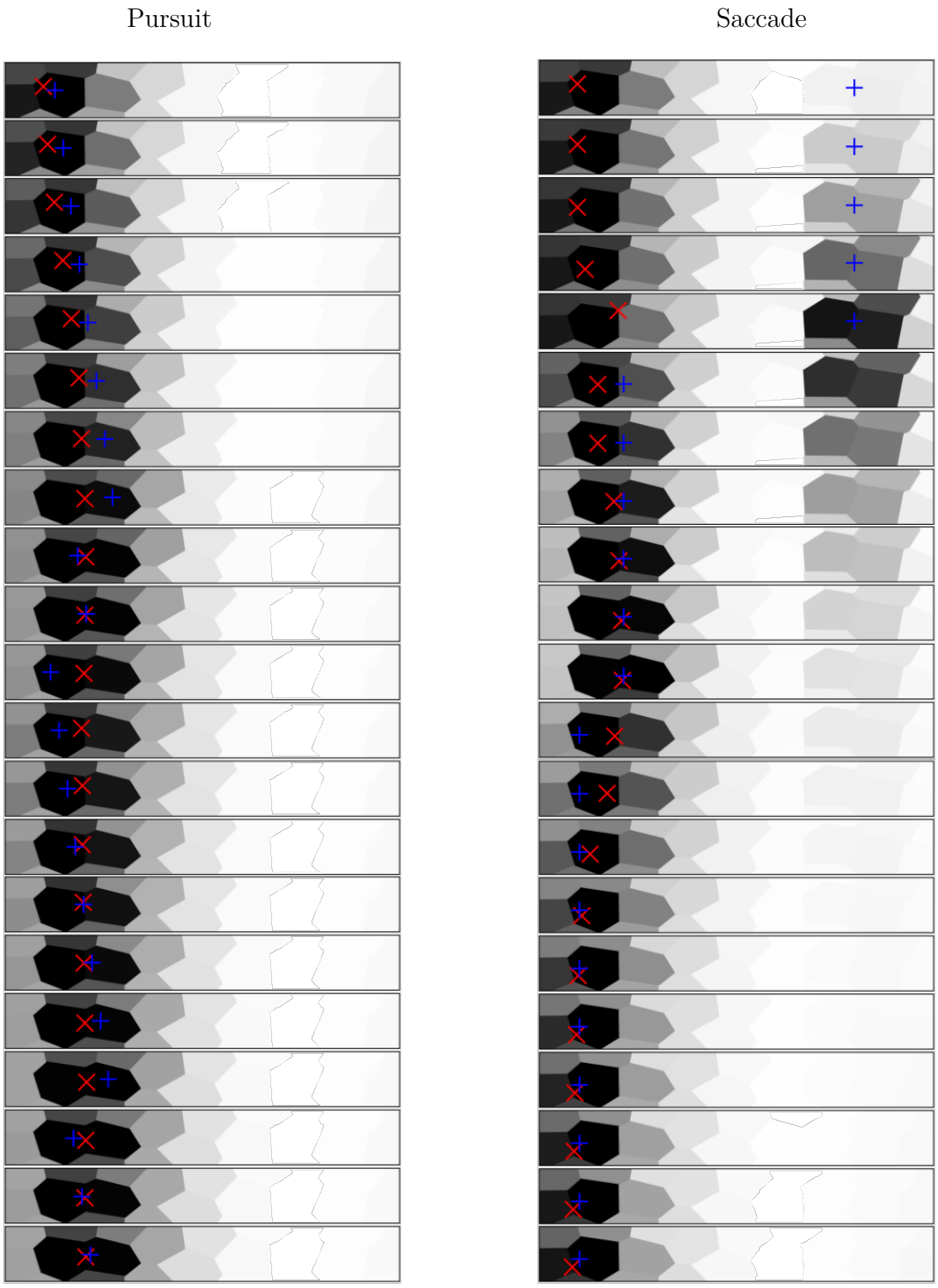


Figure 8.1: Evolution over time of DNF potential in a visual GNG. Representations are the same as in figure 6.6, only cropped differently. Snapshots were taken at regular intervals, with time advancing from top to bottom. The red × cross shows the barycenter of DNF activation. The blue + cross shows the position of the visual target. Left: After a fixation time (not shown), the target starts moving slowly to the right. Right: After a fixation time (not shown), the target shifts suddenly to the right.

would be to make predictions of the future best-matching unit every time a new motor command is triggered by DNF, connect predicted and actual best-matching unit, and use the accumulated prediction error as an indication of the need to add a new node in-between them. Indeed, higher prediction errors could be indicative of a too low resolution in an area. Motor directions could be generated according to the free energy principle, sending stimuli to the areas that are least well covered by the GNG, in order to minimize surprises in the long term.

Sensorimotor contingencies for open-ended learning This is the most ambitious perspective in the list. Again, we use DNF to trigger small motor commands, then a sensorimotor contingency is learned, associating the motor shift to the resulting perceptory shift. The trick is, if we assume that every motor change is equivalent to a sensory change, then we do not need to learn feature topologies from motor states and sensory states separately. On the contrary, we could say that if each node in the GNG is the prototype of a sensory input, then each edge is the prototype of a motor shift. The main difference compared to regular GNG is that edges are not created according to node co-activation, but instead store the outcome of a sensorimotor experience. One possible consequence is that some regions end up with more connections than others, precisely because DNF are involved: more connections mean more close-range excitation and chance for a DNF peak to emerge; more DNF peaks mean more motor commands in the vicinity, and more experience.

Now, and this is purely prospective, suppose for instance that you are learning from visual stimuli, and that you learn by making eye movements of random amplitudes in random directions. You will find out that the closer the stimuli are from the border of the field of view, the higher the chance that they disappear of your line of sight, making you lose information and miss experience. So, after a lot of development, if one area of your sensorimotor feature map should be more connected than the rest, it should be the center of your field of view. With this theory, even if you train robots equipped with perfectly regular vision, then they should have good reason to have an over-resolution (a virtual fovea) on their active feature topology. And if you add other modalities to the mix, such as audition with limitations that differ on azimuthal and elevation axes, then, maybe, you might end up with a topology of a brand new shape, like robots' own superior colliculus.

Conclusion

We have already concluded about the contributions of this thesis in chapter 7, and discussed perspectives in chapter 8. To sum up all this, and to avoid repeating what is written just pages above, we propose a final synthetic illustration in figure 8.2. Our contributions are framed in gray. Perspectives are in the red arrows.

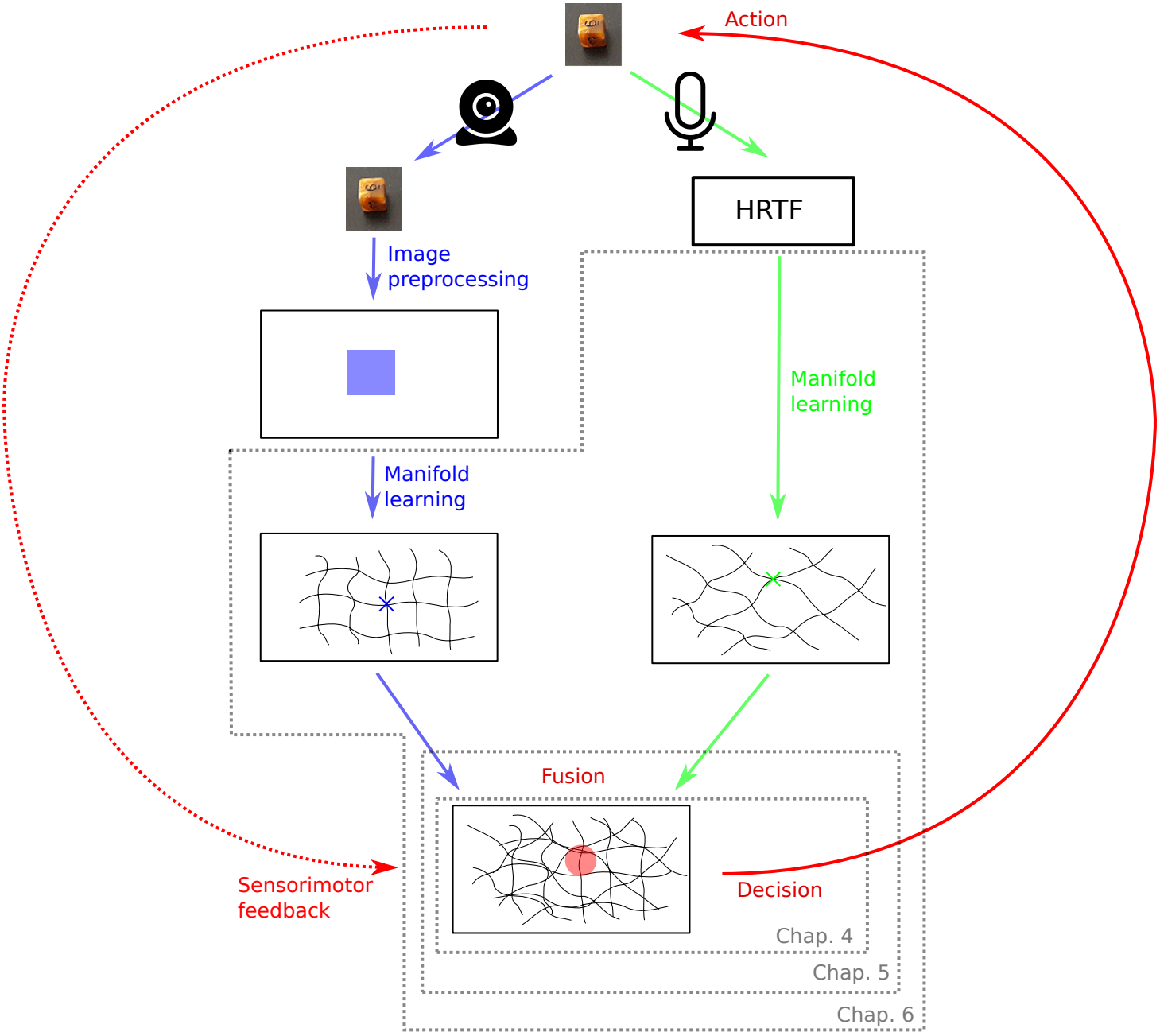


Figure 8.2: Illustrative synthesis of the scope of this thesis

Epilogue

This is the end of our guiding story. Bob's arrow finds its target and Ypomni draw their final breath, marking the conclusion of more than four years of campaign, and far too many nightly sessions of battle. Eventually, the campaign is a success, although the real achievement was mainly to arrive to this point in one piece.

But who knows, maybe this is not the end of Ypomni? Already, our players are planning the next games, with more challenges to come! Maybe some new players will be involved next time. And maybe, who knows, another lost soul will take over and start another multi-year campaign, starting where we have left off, but inevitably ending with the fateful return of Ypomni...

Publications

Articles

Forest, S., Quinton, J.-C., and Lefort, M. (2022). A dynamic neural field model of multimodal merging: application to the ventriloquist effect. *Neural Computation*, 34(8):1701–1726.

Available: https://hal.archives-ouvertes.fr/hal-03600794.

Forest, S., Villamar, J., Gautheron, L., Lefort, M., and Quinton, J.-C. (in prep.). Task dependence in decision-making algorithms: a review.

Conference

Forest, S., Quinton, J.-C., and Lefort, M. (2022). Combining manifold learning and neural field dynamics for multimodal fusion. *2022 International Joint Conference on Neural Networks (IJCNN)*.

Available: https://hal.archives-ouvertes.fr/hal-03693198

Posters

Lefort, M., Quinton, J.-C., Forest, S., Techer, A., Chauvin, A., and Avillac, M. (2018). Influence of eye-movements on multisensory stimulus localization: experiments, models and robotics applications. *Grenoble Workshop on Models and Analysis of Eye Movements*.

Available: https://hal.archives-ouvertes.fr/hal-01894621

Forest, S., Lefort, M., and Quinton, J.-C. (2019). Biological constraints in neural field models of sensor fusion. *Fourth International Workshop on Intrinsically Motivated Open-ended Learning*.

Available: https://hal.archives-ouvertes.fr/hal-03694078

Bibliography

[Alais and Burr, 2004] Alais, D. and Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration. *Current Biology*, 14(3):257–262.

[Alais et al., 2010] Alais, D., Newell, F. N., and Mamassian, P. (2010). Multisensory processing in review: from physiology to behaviour. *Seeing Perceiving*, 23(1):3–38.

[Algazi et al., 2001] Algazi, V. R., Duda, R. O., Thompson, D. M., and Avendano, C. (2001). The CIPIC HRTF database. In *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (Cat. No. 01TH8575)*, pages 99–102. IEEE.

[Allman et al., 2009] Allman, B. L., Keniston, L. P., and Meredith, M. A. (2009). Not just for bimodal neurons anymore: the contribution of unimodal neurons to cortical multisensory processing. *Brain topography*, 21(3):157–167.

[Alsius et al., 2005] Alsius, A., Navarra, J., Campbell, R., and Soto-Faraco, S. (2005). Audiovisual integration of speech falters under high attention demands. *Current Biology*, 15(9):839–843.

[Amari, 1977] Amari, S.-I. (1977). Dynamics of pattern formation in lateral-inhibition type neural fields. *Biological Cybernetics*, 27(2):77–87.

[Andersen et al., 2009] Andersen, T. S., Tiippana, K., Laarni, J., Kojo, I., and Sams, M. (2009). The role of visual spatial attention in audiovisual speech perception. *Speech Communication*, 51(2):184–193.

[Ansuini et al., 2019] Ansuini, A., Laio, A., Macke, J. H., and Zoccolan, D. (2019). Intrinsic dimension of data representations in deep neural networks. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

[Argentieri et al., 2015] Argentieri, S., Danès, P., and Souères, P. (2015). A survey on sound source localization in robotics: From binaural to array processing methods. *Computer Speech & Language*, 34(1):87–112.

[Arons, 1992] Arons, B. (1992). A review of the cocktail party effect. *Journal of the American Voice I/O Society*, 12(7):35–50.

[Arulkumaran et al., 2017] Arulkumaran, K., Deisenroth, M. P., Brundage, M., and Bharath, A. A. (2017). Deep reinforcement learning: A brief survey. *IEEE Signal Processing Magazine*, 34(6):26–38.

[Babaie-Janvier and Robinson, 2019] Babaie-Janvier, T. and Robinson, P. A. (2019). Neural field theory of corticothalamic attention with control system analysis. *Frontiers in Neuroscience*, 13:1240.

[Bajrami et al., 2015] Bajrami, X., Dërmaku, A., and Demaku, N. (2015). Artificial neural fuzzy logic algorithm for robot path finding. *IFAC-PapersOnLine*, 48(24):123–127.

[Bastos et al., 2012] Bastos, A. M., Usrey, W. M., Adams, R. A., Mangun, G. R., Fries, P., and Friston, K. J. (2012). Canonical microcircuits for predictive coding. *Neuron*, 76(4):695–711.

[Bauer et al., 2015] Bauer, J., Magg, S., and Wermter, S. (2015). Attention modeled as information in learning multisensory integration. *Neural Networks*, 65:44–52.

[Bellman and Zadeh, 1970] Bellman, R. E. and Zadeh, L. A. (1970). Decision-making in a fuzzy environment. *Management science*, 17(4):141–164.

[Bengio et al., 2013] Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828.

[Bertelson, 1999] Bertelson, P. (1999). Ventriloquism: A case of crossmodal perceptual grouping. In *Advances in psychology*, volume 129, pages 347–362. Elsevier.

[Bertelson et al., 2000] Bertelson, P., Vroomen, J., De Gelder, B., and Driver, J. (2000). The ventriloquist effect does not depend on the direction of deliberate visual attention. *Perception & psychophysics*, 62(2):321–332.

[Bitzer et al., 2014] Bitzer, S., Park, H., Blankenburg, F., and Kiebel, S. (2014). Perceptual decision making: drift-diffusion model is equivalent to a bayesian model. *Frontiers in Human Neuroscience*, 8.

[Blackmore and Miikkulainen, 1993] Blackmore, J. and Miikkulainen, R. (1993). Incremental grid growing: Encoding high-dimensional structure into a two-dimensional feature map. In *IEEE international conference on neural networks*, pages 450–455. IEEE.

[Bogacz et al., 2006] Bogacz, R., Brown, E., Moehlis, J., Holmes, P., and Cohen, J. D. (2006). The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological review*, 113(4):700.

[Bogacz et al., 2007] Bogacz, R., Usher, M., Zhang, J., and McClelland, J. L. (2007). Extending a biologically inspired model of choice: multi-alternatives, nonlinearity and value-based multidimensional choice. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 362(1485):1655–1670.

[Bosen et al., 2017] Bosen, A. K., Fleming, J. T., Allen, P. D., O'Neill, W. E., and Paige, G. D. (2017). Accumulation and decay of visual capture and the ventriloquism aftereffect caused by brief audio-visual disparities. *Experimental Brain Research*, 235(2):585–595.

[Bosking et al., 1997] Bosking, W. H., Zhang, Y., Schofield, B., and Fitzpatrick, D. (1997). Orientation selectivity and the arrangement of horizontal connections in tree shrew striate cortex. *Journal of neuroscience*, 17(6):2112–2127.

[Botvinick and Cohen, 1998] Botvinick, M. and Cohen, J. (1998). Rubber hands 'feel' touch that eyes see. *Nature*, 391(6669):756.

[Bowling et al., 2009] Bowling, S. R., Khasawneh, M. T., Kaewkuekool, S., and Cho, B. R. (2009). A logistic approximation to the cumulative normal distribution. *Journal of Industrial Engineering and Management*, 2(1):114–127.

[Braitenberg, 1986] Braitenberg, V. (1986). *Vehicles: Experiments in synthetic psychology*. MIT press.

[Broadbent, 2013] Broadbent, D. E. (2013). *Perception and communication*. Elsevier.

[Brooks, 1991] Brooks, R. A. (1991). Intelligence without representation. *Artificial intelligence*, 47(1):139–159.

[Buonomano and Merzenich, 1998] Buonomano, D. V. and Merzenich, M. M. (1998). Cortical plasticity: From synapses to maps. *Annual Review of Neuroscience*, 21(1):149–186.

[Buss and Spencer, 2018] Buss, A. T. and Spencer, J. P. (2018). Changes in frontal and posterior cortical activity underlie the early emergence of executive function. *Developmental Science*, 21(4):e12602.

[Calvert et al., 2004] Calvert, G., Spence, C., and Stein, B. (2004). *The Handbook of Multisensory Processes*. A Bradford book. MIT Press.

[Casey et al., 2012] Casey, M. C., Pavlou, A., and Timotheou, A. (2012). Audio-visual localization with hierarchical topographic maps: Modeling the superior colliculus. *Neurocomputing*, 97:344–356.

[Castanedo, 2013] Castanedo, F. (2013). A review of data fusion techniques. *The scientific world journal*, 2013.

[Cerda and Girau, 2013] Cerda, M. and Girau, B. (2013). Asymmetry in neural fields: a spatiotemporal encoding mechanism. *Biological cybernetics*, 107(2):161–178.

[Chabris and Simons, 2010] Chabris, C. F. and Simons, D. J. (2010). *The invisible gorilla: And other ways our intuitions deceive us*. Harmony, New York, NY.

[Chalmers, 2011] Chalmers, D. J. (2011). A computational foundation for the study of cognition. *Journal of Cognitive Science*, 12(4):325–359.

[Chen, 2011] Chen, S. Y. (2011). Kalman filter for robot vision: a survey. *IEEE Transactions on industrial electronics*, 59(11):4409–4420.

[Chen et al., 2020] Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

[Cynader and Berman, 1972] Cynader, M. and Berman, N. (1972). Receptive-field organization of monkey superior colliculus. *Journal of Neurophysiology*, 35(2):187–201.

[Das et al., 2017] Das, T. K., Harischandra, P. D., and Abeykoon, A. H. S. (2017). Extended kalman filter based fusion of reliable sensors using fuzzy logic. In *2017 Moratuwa Engineering Research Conference (MERCon)*, pages 58–63. IEEE.

[Deneve et al., 2001] Deneve, S., Latham, P. E., and Pouget, A. (2001). Efficient computation and cue integration with noisy population codes. *Nature neuroscience*, 4(8):826–831.

[Deutsch and Deutsch, 1963] Deutsch, J. A. and Deutsch, D. (1963). Attention: Some theoretical considerations. *Psychological review*, 70(1):80–90.

[Dietrich, 1994] Dietrich, E. (1994). Computationalism. In *Thinking Computers and Virtual Persons*, pages 109–136. Elsevier.

[Doki et al., 2015] Doki, K., Suyama, K., Funabora, Y., and Doki, S. (2015). Robust localization for mobile robot by K-L divergence-based sensor data fusion. In *IECON 2015 - 41st Annual Conference of the IEEE Industrial Electronics Society*, pages 2638–2643.

[Driver and Spence, 2004] Driver, J. and Spence, C. (2004). Crossmodal spatial attention: Evidence from human performance. In Spence, C. and Driver, J., editors, *Crossmodal Space and Crossmodal Attention*. Oxford University Press.

[Droniou et al., 2015] Droniou, A., Ivaldi, S., and Sigaud, O. (2015). Deep unsupervised network for multimodal perception, representation and classification. *Robotics and Autonomous Systems*, 71:83–98.

[Dubois et al., 2004] Dubois, D., Foulloy, L., Mauris, G., and Prade, H. (2004). Probability-possibility transformations, triangular fuzzy sets, and probabilistic inequalities. *Reliable computing*, 10(4):273–297.

[Dubois and Perny, 2016] Dubois, D. and Perny, P. (2016). A review of fuzzy sets in decision sciences: Achievements, limitations and perspectives. In Greco, S., Ehrgott, M., and Figueira, J. R., editors, *Multiple Criteria Decision Analysis: State of the Art Surveys*, pages 637–691. Springer New York, New York, NY.

[Durrant-Whyte and Henderson, 2016] Durrant-Whyte, H. and Henderson, T. C. (2016). Multisensor data fusion. In Siciliano, B. and Khatib, O., editors, *Springer Handbook of Robotics*, pages 867–896. Springer International Publishing, Cham.

[Ernst and Banks, 2002] Ernst, M. O. and Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, 415:429–433.

[Ernst and Bulthoff, 2004] Ernst, M. O. and Bulthoff, H. (2004). Merging the senses into a robust percept. *Trends in Cognitive Sciences*, 8(4):162–169.

[Fecteau et al., 2004] Fecteau, J. H., Bell, A. H., and Munoz, D. P. (2004). Neural correlates of the automatic and goal-driven biases in orienting spatial attention. *Journal of Neurophysiology*, 92(3):1728–1737.

[Fendrich and Corballis, 2001] Fendrich, R. and Corballis, P. M. (2001). The temporal cross-capture of audition and vision. *Perception & Psychophysics*, 63(4):719–725.

[Fix et al., 2011] Fix, J., Rougier, N., and Alexandre, F. (2011). A dynamic neural field approach to the covert and overt deployment of spatial attention. *Cognitive Computation*, 3(1):279–293.

[Frens et al., 1995] Frens, M. A., Van Opstal, A. J., and Van der Willigen, R. F. (1995). Spatial and temporal factors determine auditory-visual interactions in human saccadic eye movements. *Perception & Psychophysics*, 57(6):802–16.

[Frissen et al., 2012] Frissen, I., Vroomen, J., and de Gelder, B. (2012). The aftereffects of ventriloquism: the time course of the visual recalibration of auditory localization. *Seeing and perceiving*, 25(1):1–14.

[Friston, 2010] Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11(2):127–138.

[Friston and Kiebel, 2009] Friston, K. and Kiebel, S. (2009). Predictive coding under the free-energy principle. *Philosophical transactions of the Royal Society B: Biological sciences*, 364(1521):1211–1221.

[Friston et al., 2012] Friston, K., Thornton, C., and Clark, A. (2012). Free-energy minimization and the dark-room problem. *Frontiers in psychology*, 3:130.

[Fritzke, 1994] Fritzke, B. (1994). Growing cell structures—a self-organizing network for unsupervised and supervised learning. *Neural networks*, 7(9):1441–1460.

[Fritzke, 1995a] Fritzke, B. (1995a). Growing grid—a self-organizing network with constant neighborhood range and adaptation strength. *Neural processing letters*, 2(5):9–13.

[Fritzke, 1995b] Fritzke, B. (1995b). A growing neural gas network learns topologies. In Tesauro, G., Touretzky, D., and Leen, T., editors, *Advances in Neural Information Processing Systems*, volume 7. MIT Press.

[Fritzke, 1997] Fritzke, B. (1997). A self-organizing network that can follow non-stationary distributions. In *International conference on artificial neural networks*, pages 613–618. Springer.

[Gandhi and Katnani, 2011] Gandhi, N. J. and Katnani, H. A. (2011). Motor functions of the superior colliculus. *Annual Review of Neuroscience*, 34(1):205–231.

[Gepperth and Lefort, 2016] Gepperth, A. and Lefort, M. (2016). Learning to be attractive: Probabilistic computation with dynamic attractor networks. In *2016 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, pages 270–277.

[Gescheider, 1997] Gescheider, G. A. (1997). *Psychophysics: the fundamentals*. Lawrence Erlbaum Associates, New Jersey.

[Gibson, 1960] Gibson, J. J. (1960). The concept of the stimulus in psychology. *American psychologist*, 15(11):694–703.

[Girard and Berthoz, 2005] Girard, B. and Berthoz, A. (2005). From brainstem to cortex: Computational models of saccade generation circuitry. *Progress in Neurobiology*, 77(4):215–251.

[Goertzel, 2014] Goertzel, B. (2014). Artificial general intelligence: concept, state of the art, and future prospects. *Journal of Artificial General Intelligence*, 5(1):1.

[Goertzel et al., 2011] Goertzel, B., Pitt, J., Wigmore, J., Geisweiller, N., Cai, Z., Lian, R., Huang, D., and Yu, G. (2011). Cognitive synergy between procedural and declarative learning in the control of animated and robotic agents using the OpenCogPrime AGI architecture. In *Twenty-Fifth AAAI Conference on Artificial Intelligence*.

[Gold and Shadlen, 2001] Gold, J. I. and Shadlen, M. N. (2001). Neural computations that underlie decisions about sensory stimuli. *Trends in cognitive sciences*, 5(1):10–16.

[Gold and Shadlen, 2007] Gold, J. I. and Shadlen, M. N. (2007). The neural basis of decision making. *Annual Review of Neuroscience*, 30(1):535–574.

[Gonnier et al., 2021] Gonnier, N., Boniface, Y., and Frezza-Buet, H. (2021). Input prediction using consensus driven SOMs. In *2021 8th International Conference on Soft Computing & Machine Intelligence (ISCMI)*, pages 38–42. IEEE.

[Grieben et al., 2020] Grieben, R., Tekülve, J., Zibner, S. K. U., Lins, J., Schneegans, S., and Schöner, G. (2020). Scene memory and spatial inhibition in visual search. *Attention, Perception, & Psychophysics*, 82(2):775–798.

[Huang et al., 2019] Huang, K., Ma, X., Song, R., Rong, X., Tian, X., and Li, Y. (2019). An autonomous developmental cognitive architecture based on incremental associative neural network with dynamic audiovisual fusion. *IEEE Access*, 7:8789–8807.

[Jack and Thurlow, 1973] Jack, C. E. and Thurlow, W. R. (1973). Effects of degree of visual association and angle of displacement on the "ventriloquism" effect. *Perceptual and motor skills*, 37(3):967–979.

[Jaegle et al., 2021] Jaegle, A., Gimeno, F., Brock, A., Vinyals, O., Zisserman, A., and Carreira, J. (2021). Perceiver: General perception with iterative attention. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 4651–4664. PMLR.

[Jay and Sparks, 1987] Jay, M. F. and Sparks, D. L. (1987). Sensorimotor integration in the primate superior colliculus. I. Motor convergence. *Journal of Neurophysiology*, 57(1):22–34.

[Jenkins et al., 2021] Jenkins, G. W., Samuelson, L. K., Penny, W., and Spencer, J. P. (2021). Learning words in space and time: Contrasting models of the suspicious coincidence effect. *Cognition*, 210:104576.

[Jiang and Bernstein, 2011] Jiang, J. and Bernstein, L. E. (2011). Psychophysics of the mcgurk and other audiovisual speech integration effects. *Journal of Experimental Psychology: Human Perception and Performance*, 37(4):1193.

[Johnsson et al., 2011] Johnsson, M., Martinsson, M., Gil, D., and Hesslow, G. (2011). Associative self-organizing map. In Mwasiagi, J. I., editor, *Self Organizing Maps*, chapter 30. IntechOpen, Rijeka.

[Julier and Uhlmann, 2004] Julier, S. J. and Uhlmann, J. K. (2004). Unscented filtering and nonlinear estimation. *Proceedings of the IEEE*, 92(3):401–422.

[Jun et al., 2021] Jun, E. J., Bautista, A. R., Nunez, M. D., Allen, D. C., Tak, J. H., Alvarez, E., and Basso, M. A. (2021). Causal role for the primate superior colliculus in the computation of evidence for perceptual decisions. *Nature Neuroscience*, 24(8):1121–1131.

[Kaelbling et al., 1996] Kaelbling, L. P., Littman, M. L., and Moore, A. W. (1996). Reinforcement learning: A survey. *Journal of artificial intelligence research*, 4:237–285.

[Kalman, 1960] Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82(1):35–45.

[Kapoula and Pain, 2020] Kapoula, Z. and Pain, E. (2020). Differential impact of sound on saccades vergence and combine eye movements: A multiple case study. *Journal of Clinical Studies & Medical Case Reports*, 7:095.

[Khacef et al., 2020] Khacef, L., Rodriguez, L., and Miramond, B. (2020). Brain-inspired self-organization with cellular neuromorphic computing for multimodal unsupervised learning. *Electronics*, 9(10).

[Khansari-Zadeh and Billard, 2011] Khansari-Zadeh, S. M. and Billard, A. (2011). Learning stable nonlinear dynamical systems with gaussian mixture models. *IEEE Transactions on Robotics*, 27(5):943–957.

[Khosla et al., 2020] Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., and Krishnan, D. (2020). Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673.

[King, 2004] King, A. J. (2004). The superior colliculus. *Current Biology*, 14(9):335–338.

[Kingma and Welling, 2014] Kingma, D. P. and Welling, M. (2014). Auto-encoding variational bayes.

[Kingma and Welling, 2019] Kingma, D. P. and Welling, M. (2019). An introduction to variational autoencoders. *Foundations and Trends in Machine Learning*, 12(4):307–392.

[Klier et al., 2001] Klier, E. M., Wang, H., and Crawford, J. D. (2001). The superior colliculus encodes gaze commands in retinal coordinates. *Nature Neuroscience*, 4(6):627–632.

[Knill and Pouget, 2004] Knill, D. C. and Pouget, A. (2004). The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends in Neurosciences*, 27(12):712–719.

[Knudsen, 1982] Knudsen, E. (1982). Auditory and visual maps of space in the optic tectum of the owl. *Journal of Neuroscience*, 2(9):1177–1194.

[Knudsen and Brainard, 1991] Knudsen, E. I. and Brainard, M. S. (1991). Visual instruction of the neural map of auditory space in the developing optic tectum. *Science*, 253(5015):85–87.

[Kohonen, 1982] Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43(1):59–69.

[Krauzlis et al., 2013] Krauzlis, R. J., Lovejoy, L. P., and Zénon, A. (2013). Superior colliculus and visual spatial attention. *Annual Review of Neuroscience*, 36(1):165–182.

[Krishnapuram and Keller, 1992] Krishnapuram, R. and Keller, J. M. (1992). Fuzzy set theoretic approach to computer vision: An overview. In *[1992 Proceedings] IEEE International Conference on Fuzzy Systems*, pages 135–142. IEEE.

[Kustov and Lee Robinson, 1996] Kustov, A. A. and Lee Robinson, D. (1996). Shared neural control of attentional shifts and eye movements. *Nature*, 384(6604):74–77.

[Laflaquière et al., 2018] Laflaquière, A., O'Regan, J. K., Gas, B., and Terekhov, A. (2018). Discovering space—grounding spatial topology and metric regularity in a naive agent's sensorimotor experience. *Neural Networks*, 105:371–392.

[Lallee and Dominey, 2013] Lallee, S. and Dominey, P. F. (2013). Multi-modal convergence maps: from body schema and self-representation to mental imagery. *Adaptive Behavior*, 21(4):274–285.

[Lefort et al., 2013] Lefort, M., Boniface, Y., and Girau, B. (2013). SOMMA: Cortically inspired paradigms for multimodal processing. In *The 2013 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

[Lepora et al., 2012] Lepora, N. F., Fox, C. W., Evans, M. H., Diamond, M. E., Gurney, K., and Prescott, T. J. (2012). Optimal decision-making in mammals: insights from a robot study of rodent texture discrimination. *Journal of The Royal Society Interface*, 9(72):1517–1528.

[Lepora and Gurney, 2012] Lepora, N. F. and Gurney, K. N. (2012). The basal ganglia optimize decision making over general perceptual hypotheses. *Neural Computation*, 24(11):2924–2945.

[Lepora and Pezzulo, 2015] Lepora, N. F. and Pezzulo, G. (2015). Embodied choice: how action influences perceptual decision making. *PLoS computational biology*, 11(4):e1004110.

[Li et al., 2014] Li, M., Liu, F., Juusola, M., and Tang, S. (2014). Perceptual color map in macaque visual area V4. *Journal of Neuroscience*, 34(1):202–217.

[Lobo et al., 2018] Lobo, L., Heras-Escribano, M., and Travieso, D. (2018). The history and philosophy of ecological psychology. *Frontiers in Psychology*, 9.

[Ma et al., 2006] Ma, W. J., Beck, J. M., Latham, P. E., and Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nature neuroscience*, 9(11):1432–1438.

[Macaluso et al., 2016] Macaluso, E., Noppeney, U., Talsma, D., Vercillo, T., Hartcher-O'Brien, J., and Adam, R. (2016). The curious incident of attention in multisensory integration: bottom-up vs. top-down. *Multisensory Research*, 29(6-7):557–583.

[Mamdani and Assilian, 1975] Mamdani, E. H. and Assilian, S. (1975). An experiment in linguistic synthesis with a fuzzy logic controller. *International journal of man-machine studies*, 7(1):1–13.

[Manfredi et al., 2006] Manfredi, L., Maini, E. S., Dario, P., Laschi, C., Girard, B., Tabareau, N., and Berthoz, A. (2006). Implementation of a neurophysiological model of saccadic eye movements on an anthropomorphic robotic head. In *2006 6th IEEE-RAS International Conference on Humanoid Robots*, pages 438–443.

[Manfredi et al., 2009] Manfredi, L., Maini, E. S., and Laschi, C. (2009). Neurophysiological models of gaze control in humanoid robotics. In Choi, B., editor, *Humanoid Robots*, chapter 10. IntechOpen, Rijeka.

[Marino et al., 2012] Marino, R. A., Trappenberg, T. P., Dorris, M., and Munoz, D. P. (2012). Spatial interactions in the superior colliculus predict saccade behavior in a neural field model. *J. Cognitive Neuroscience*, 24(2):315–336.

[Marr, 1982] Marr, D. (1982). *Vision. A Computational Investigation Into the Human Representation and Processing of Visual Information.* MIT Press.

[Marsland et al., 2002] Marsland, S., Shapiro, J., and Nehmzow, U. (2002). A self-organising network that grows when required. *Neural networks*, 15(8-9):1041–1058.

[Martinetz and Schulten, 1991] Martinetz, T. and Schulten, K. (1991). A "neural-gas" network learns topologies. *Artificial neural networks*, 1:397–402.

[McCulloch and Pitts, 1943] McCulloch, W. S. and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133.

[McGurk and MacDonald, 1976] McGurk, H. and MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264(5588):746–748.

[Mendonça et al., 2015] Mendonça, C., Escher, A., van de Par, S., and Colonius, H. (2015). Predicting auditory space calibration from recent multisensory experience. *Experimental Brain Research*, 233(7):1983–1991.

[Meredith et al., 2020] Meredith, M. A., Keniston, L. P., Prickett, E. H., Bajwa, M., Cojanu, A., Clemo, H. R., and Allman, B. L. (2020). What is a multisensory cortex? a laminar, connectional, and functional study of a ferret temporal cortical multisensory area. *Journal of Comparative Neurology*, 528(11):1864–1882.

[Meredith and Stein, 1986] Meredith, M. A. and Stein, B. E. (1986). Visual, auditory, and somatosensory convergence on cells in superior colliculus results in multisensory integration. *Journal of Neurophysiology*, 56(3):640–662.

[Mnih et al., 2015] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. (2015). Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533.

[Moray, 1959] Moray, N. (1959). Attention in dichotic listening: Affective cues and the influence of instructions. *Quarterly journal of experimental psychology*, 11(1):56–60.

[Mountcastle, 1997] Mountcastle, V. B. (1997). The columnar organization of the neocortex. *Brain: a journal of neurology*, 120(4):701–722.

[Munhall et al., 1996] Munhall, K. G., Gribble, P., Sacco, L., and Ward, M. (1996). Temporal constraints on the mcgurk effect. *Perception & psychophysics*, 58(3):351–362.

[Ménard and Frezza-Buet, 2005] Ménard, O. and Frezza-Buet, H. (2005). Model of multimodal cortical processing: Coherent learning in self-organizing modules. *Neural Networks*, 18(5):646–655.

[Neisser, 2014] Neisser, U. (2014). *Cognitive psychology: Classic edition*. Psychology press.

[Newell et al., 2001] Newell, F. N., Ernst, M. O., Tjan, B. S., and Bülthoff, H. H. (2001). Viewpoint dependence in visual and haptic object recognition. *Psychological Science*, 12(1):37–42. PMID: 11294226.

[Ngiam et al., 2011] Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., and Ng, A. Y. (2011). Multimodal deep learning. In *ICML*.

[Norman, 1968] Norman, D. A. (1968). Toward a theory of memory and attention. *Psychological review*, 75(6):522–536.

[O'Regan, 1992] O'Regan, J. K. (1992). Solving the "real" mysteries of visual perception: the world as an outside memory. *Canadian Journal of Psychology/Revue canadienne de psychologie*, 46(3):461–488.

[O'Regan and Noë, 2001] O'Regan, J. K. and Noë, A. (2001). A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences*, 24(5):939–973.

[Ottes et al., 1986] Ottes, F. P., Gisbergen, J. A. V., and Eggermont, J. J. (1986). Visuomotor fields of the superior colliculus: A quantitative model. *Vision Research*, 26(6):857–873.

[Parise and Ernst, 2016] Parise, C. V. and Ernst, M. O. (2016). Correlation detection as a general mechanism for multisensory integration. *Nature communications*, 7(1):1–9.

[Parisi et al., 2017] Parisi, G. I., Tani, J., Weber, C., and Wermter, S. (2017). Emergence of multimodal action representations from neural network self-organization. *Cognitive Systems Research*, 43:208–221.

[Parr et al., 2022] Parr, T., Pezzulo, G., and Friston, K. J. (2022). *Active inference: the free energy principle in mind, brain, and behavior*. MIT Press.

[Plataniotis and Hatzinakos, 2000] Plataniotis, K. N. and Hatzinakos, D. (2000). Gaussian mixtures and their applications to signal processing. In Stergiopoulos, S., editor, *Advanced Signal Processing Handbook: Theory and Implementation for Radar, Sonar, and Medical Imaging Real Time Systems (1st ed.)*, pages 89–124. CRC Press.

[Pouget et al., 2003] Pouget, A., Dayan, P., and Zemel, R. S. (2003). Inference and computation with population codes. *Annual review of neuroscience*, 26(1):381–410.

[Pouget et al., 2002] Pouget, A., Deneve, S., and Duhamel, J.-R. (2002). A computational perspective on the neural basis of multisensory spatial representations. *Nature Reviews Neuroscience*, 3(9):741–747.

[Quinton, 2010] Quinton, J.-C. (2010). Exploring and optimizing dynamic neural fields parameters using genetic algorithms. In *The 2010 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7.

[Quinton and Girau, 2010] Quinton, J.-C. and Girau, B. (2010). A sparse implementation of dynamic competition in continuous neural fields. In *Brain Inspired Cognitive Systems 2010 - BICS 2010*, Madrid, Spain.

[Quinton and Girau, 2011] Quinton, J.-C. and Girau, B. (2011). Predictive neural fields for improved tracking and attentional properties. In *The 2011 International Joint Conference on Neural Networks*, pages 1629–1636. IEEE.

[Quinton and Goffart, 2018] Quinton, J.-C. and Goffart, L. (2018). A unified dynamic neural field model of goal directed eye movements. *Connection Science*, 30(1):20–52.

[Qureshi et al., 2018] Qureshi, M. S., Swarnkar, P., and Gupta, S. (2018). A supervisory on-line tuned fuzzy logic based sliding mode control for robotics: An application to surgical robots. *Robotics and Autonomous Systems*, 109:68–85.

[Radeau and Bertelson, 1974] Radeau, M. and Bertelson, P. (1974). The after-effects of ventriloquism. *The Quarterly journal of experimental psychology*, 26(1):63–71.

[Rasmussen, 2004] Rasmussen, C. E. (2004). Gaussian processes in machine learning. In Bousquet, O., von Luxburg, U., and Rätsch, G., editors, *Advanced Lectures on Machine Learning: ML Summer Schools 2003, Canberra, Australia, February 2 - 14, 2003, Tübingen, Germany, August 4 - 16, 2003, Revised Lectures*, pages 63–71. Springer Berlin Heidelberg, Berlin, Heidelberg.

[Ratcliff and McKoon, 2008] Ratcliff, R. and McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation*, 20(4):873–922.

[Rizzolatti et al., 1994] Rizzolatti, G., Riggio, L., and Sheliga, B. M. (1994). Space and selective attention. In Umiltà, C. and Moscovitch, M., editors, *Attention and performance 15: Conscious and nonconscious information processing*, pages 232–265. The MIT Press, Cambridge, MA, US.

[Rohde et al., 2016] Rohde, M., van Dam, L. C., and Ernst, M. O. (2016). Statistically optimal multisensory cue integration: A practical tutorial. *Multisensory Research*, 29(4-5):279–317.

[Rougier, 2006] Rougier, N. P. (2006). Dynamic neural field with local inhibition. *Biological Cybernetics*, 94(3):169–179.

[Rougier and Vitay, 2006] Rougier, N. P. and Vitay, J. (2006). Emergence of attention within a neural population. *Neural Networks*, 19(5):573–581.

[Roxin, 2019] Roxin, A. (2019). Drift–diffusion models for multiple-alternative forced-choice decision making. *The Journal of Mathematical Neuroscience*, 9(1):1–23.

[Russo and Ramponi, 1994] Russo, F. and Ramponi, G. (1994). Fuzzy methods for multisensor data fusion. *IEEE transactions on instrumentation and measurement*, 43(2):288–294.

[Sandamirskaya, 2014] Sandamirskaya, Y. (2014). Dynamic neural fields as a step toward cognitive neuromorphic architectures. *Frontiers in Neuroscience*, 7:276.

[Scarpina and Tagini, 2017] Scarpina, F. and Tagini, S. (2017). The stroop color and word test. *Frontiers in Psychology*, 8.

[Schauer and Gross, 2004] Schauer, C. and Gross, H. M. (2004). Design and optimization of Amari neural fields for early auditory-visual integration. In *2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No.04CH37541)*, volume 4, pages 2523–2528.

[Schöner et al., 2015] Schöner, G., Spencer, J., and DFT Research Group (2015). *Dynamic Thinking: A Primer on Dynamic Field Theory*. Oxford Series in Developmental Cognitive Neuroscience. Oxford University Press.

[Shams et al., 2002] Shams, L., Kamitani, Y., and Shimojo, S. (2002). Visual illusion induced by sound. *Cognitive brain research*, 14(1):147–152.

[Sigaud et al., 2011] Sigaud, O., Salaün, C., and Padois, V. (2011). On-line regression algorithms for learning mechanical models of robots: a survey. *Robotics and Autonomous Systems*, 59(12):1115–1129.

[Slutsky and Recanzone, 2001] Slutsky, D. A. and Recanzone, G. H. (2001). Temporal and spatial dependency of the ventriloquism effect. *Neuroreport*, 12(1):7–10.

[Sobrevilla and Montseny, 2003] Sobrevilla, P. and Montseny, E. (2003). Fuzzy sets in computer vision: An overview. *Mathware and Soft Computing*, 10(2/3):71–83.

[Srinivasan et al., 1982] Srinivasan, M. V., Laughlin, S. B., and Dubs, A. (1982). Predictive coding: a fresh view of inhibition in the retina. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 216(1205):427–459.

[Stein et al., 1989] Stein, B. E., Meredith, M. A., Huneycutt, W. S., and McDade, L. (1989). Behavioral indices of multisensory integration: orientation to visual cues is affected by auditory stimuli. *Journal of Cognitive Neuroscience*, 1(1):12–24.

[Sternberg, 1996] Sternberg, R. J. (1996). *Cognitive psychology*. Harcourt Brace College Publishers.

[Su et al., 2019] Su, J., Vargas, D. V., and Sakurai, K. (2019). One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, 23(5):828–841.

[Sugeno and Yasukawa, 1993] Sugeno, M. and Yasukawa, T. (1993). A fuzzy-logic-based approach to qualitative modeling. *IEEE Transactions on fuzzy systems*, 1(1):7–31.

[Sun et al., 1999] Sun, R., Peterson, T., and Merrill, E. (1999). A hybrid architecture for situated learning of reactive sequential decision making. *Applied Intelligence*, 11(1):109–127.

[Talsma, 2015] Talsma, D. (2015). Predictive coding and multisensory integration: an attentional account of the multisensory mind. *Frontiers in Integrative Neuroscience*, 9:19.

[Talsma et al., 2010] Talsma, D., Senkowski, D., Soto-Faraco, S., and Woldorff, M. G. (2010). The multifaceted interplay between attention and multisensory integration. *Trends in cognitive sciences*, 14(9):400–410.

[Taouali et al., 2015] Taouali, W., Goffart, L., Alexandre, F., and Rougier, N. P. (2015). A parsimonious computational model of visual target position encoding in the superior colliculus. *Biological Cybernetics*, 109(4):549–559.

[Tekülve et al., 2019] Tekülve, J., Fois, A., Sandamirskaya, Y., and Schöner, G. (2019). Autonomous sequence generation for a neural dynamic robot: Scene perception, serial order, and object-oriented movement. *Frontiers in Neurorobotics*, 13.

[Thériault et al., 2022] Thériault, R., Landry, M., and Raz, A. (2022). The rubber hand illusion: Top-down attention modulates embodiment. *Quarterly Journal of Experimental Psychology*, page 17470218221078858.

[Trappenberg et al., 2001] Trappenberg, T. P., Munoz, D. P., and Klein, R. M. (2001). A model of saccade initiation based on the competitive integration of exogenous and endogenous signals in the superior colliculus. *Journal of Cognitive Neuroscience*, 13(2):256–271.

[Treisman, 1960] Treisman, A. M. (1960). Contextual cues in selective listening. *Quarterly Journal of Experimental Psychology*, 12(4):242–248.

[Usher and McClelland, 2001] Usher, M. and McClelland, J. L. (2001). The time course of perceptual choice: the leaky, competing accumulator model. *Psychological review*, 108(3):550.

[Van Hulle, 2012] Van Hulle, M. M. (2012). Self-organizing maps. In Rozenberg, G., Bäck, T., and Kok, J. N., editors, *Handbook of Natural Computing*, pages 585–622. Springer, Berlin, Heidelberg.

[Van Wassenhove et al., 2007] Van Wassenhove, V., Grant, K. W., and Poeppel, D. (2007). Temporal window of integration in auditory-visual speech perception. *Neuropsychologia*, 45(3):598–607.

[Vavrečka and Farkaš, 2014] Vavrečka, M. and Farkaš, I. (2014). A multimodal connectionist architecture for unsupervised grounding of spatial language. *Cognitive Computation*, 6(1):101–112.

[Vickers, 1970] Vickers, D. (1970). Evidence for an accumulator model of psychophysical discrimination. *Ergonomics*, 13(1):37–58.

[Vijayakumar and Schaal, 2000] Vijayakumar, S. and Schaal, S. (2000). Locally weighted projection regression: An o(n) algorithm for incremental real time learning in high dimensional space. In *Proceedings of the seventeenth international conference on machine learning (ICML 2000)*, volume 1, pages 288–293. Morgan Kaufmann.

[Villalobos and Dewhurst, 2017] Villalobos, M. and Dewhurst, J. (2017). Why post-cognitivism does not (necessarily) entail anti-computationalism. *Adaptive Behavior*, 25(3):117–128.

[Vroomen et al., 2001] Vroomen, J., Bertelson, P., and De Gelder, B. (2001). The ventriloquist effect does not depend on the direction of automatic visual attention. *Perception & psychophysics*, 63(4):651–659.

[Wakileh and Gill, 1988] Wakileh, B. A. M. and Gill, K. F. (1988). Use of fuzzy logic in robotics. *Computers in industry*, 10(1):35–46.

[Wallace and Stein, 1996] Wallace, M. T. and Stein, B. E. (1996). Chapter 21: Sensory organization of the superior colliculus in cat and monkey. In Norita, M., Bando, T., and Stein, B. E., editors, *Extrageniculostriate Mechanisms Underlying Visually-Guided Orientation Behavior*, volume 112 of *Progress in Brain Research*, pages 301–311. Elsevier.

[Wallace and Stein, 1997] Wallace, M. T. and Stein, B. E. (1997). Development of multisensory neurons and multisensory integration in cat superior colliculus. *Journal of Neuroscience*, 17(7):2429–2444.

[Wang, 2002] Wang, X.-J. (2002). Probabilistic decision making by slow reverberation in cortical circuits. *Neuron*, 36(5):955–968.

[Weisswange et al., 2011] Weisswange, T. H., Rothkopf, C. A., Rodemann, T., and Triesch, J. (2011). Bayesian cue integration as a developmental outcome of reward mediated learning. *PLoS One*, 6(7):1–11.

[Wijeakumar et al., 2017] Wijeakumar, S., Ambrose, J. P., Spencer, J. P., and Curtu, R. (2017). Model-based functional neuroimaging using dynamic neural fields: An integrative cognitive neuroscience approach. *Journal of Mathematical Psychology*, 76:212–235.

[Wilimzig et al., 2006] Wilimzig, C., Schneider, S., and Schöner, G. (2006). The time course of saccadic decision making: Dynamic field theory. *Neural Networks*, 19(8):1059–1074. Neurobiology of Decision Making.

[Wilson and Cowan, 1973] Wilson, H. R. and Cowan, J. D. (1973). A mathematical theory of the functional dynamics of cortical and thalamic nervous tissue. *Kybernetik*, 13(2):55–80.

[Witten and Knudsen, 2005] Witten, I. B. and Knudsen, E. I. (2005). Why seeing is believing: merging auditory and visual worlds. *Neuron*, 48(3):489–496.

[Xu et al., 2018] Xu, J., Bi, T., Wu, J., Meng, F., Wang, K., Hu, J., Han, X., Zhang, J., Zhou, X., Keniston, L., et al. (2018). Spatial receptive field shift by preceding cross-modal stimulation in the cat superior colliculus. *The Journal of Physiology*, 596(20):5033–5050.

[Yang et al., 2017] Yang, X., Ramesh, P., Chitta, R., Madhvanath, S., Bernal, E. A., and Luo, J. (2017). Deep multimodal representation learning from temporal data. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5066–5074.

[Yuval-Greenberg et al., 2014] Yuval-Greenberg, S., Merriam, E. P., and Heeger, D. J. (2014). Spontaneous microsaccades reflect shifts in covert attention. *Journal of Neuroscience*, 34(41):13693–13700.

[Zadeh, 1965] Zadeh, L. A. (1965). Fuzzy sets. *Information and control*, 8(3):338–353.

Résumé de la thèse

Cette thèse porte sur la fusion multimodale sur des topologies artificielles dans un contexte de perception active. À titre d'exemple, les humains reçoivent des informations denses provenant de multiples capteurs et utilisent divers mécanismes pour sélectionner et se concentrer sur les signaux pertinents uniquement, par exemple en déplaçant le regard vers un objet pour mieux le voir. En raison des irrégularités dans les topologies sensorielles (cf. fovéa), les actions peuvent améliorer la perception, tandis que l'extraction et la fusion de données aident également à choisir le meilleur plan d'action. Les systèmes artificiels, par exemple les robots sociaux, font face à des besoins similaires (figure 8.3), malgré un ensemble de contraintes physiques qui leur est propre.

Cette thèse propose des modèles computationnels pour l'IA, en s'inspirant entre autres d'études en neurosciences impliquant le colliculus supérieur (*superior colliculus*, SC), une structure sous-corticale impliquée dans la génération de saccades vers des stimuli visuels, auditifs ou multisensoriels. Une attention particulière est portée sur l'influence des topologies de la perception, c'est-à-dire les régularités et irrégularités des espaces des descripteurs jouant un rôle dans la prise de décision. Dans le SC, une topologie visuelle est connue, mais pas de topologie auditive ou multimodale à proprement parler. À défaut de les modéliser avec fidélité, il sera nécessaire de générer de nouvelles topologies qui respectent les avantages et restrictions des espaces sensoriels propres à cette structure.

Première contribution : Revue et uniformisation d'algorithmes de prise de décision

Pour sélectionner des informations à partir de signaux multiples dans un contexte dynamique et multimodal, il faut trouver un moyen de générer des décisions fiables et robustes. La prise de décision en général a été abordée à la fois en psychologie et en robotique, via de nombreux algorithmes différents : *drift-diffusion model*, *leaky competitive accumulator*, estimation de maximum-vraisemblance côté psychologie ; filtres de Kalman et logique floue côté robotique ; de rares modèles faisant déjà la passerelle entre les deux, comme les champs neuronaux dynamiques (*dynamic neural fields*, DNF).

Une de nos contributions est de passer en revue et comparer ces algorithmes, en soulignant leurs propriétés spatio-temporelles, y compris la fusion, l'attention sélective, la mémoire, etc. Après un travail d'uniformisation des algorithmes, issus de domaines différents avec des conventions différentes, la revue fait ressortir des propriétés parfois communes entre les modèles : capacités d'adaptation entre sélection et interpolation pour la logique floue et les DNF, de lissage temporel pour les filtres de Kalman et les DNF, et de réactivité ajustable pour les modèles bio-inspirés à base d'accumulateurs (dont les DNF). Les DNF apparaissent comme le modèle le plus versatile, au prix d'un coût computationnel relativement élevé et d'une paramétrisation plus complexe.

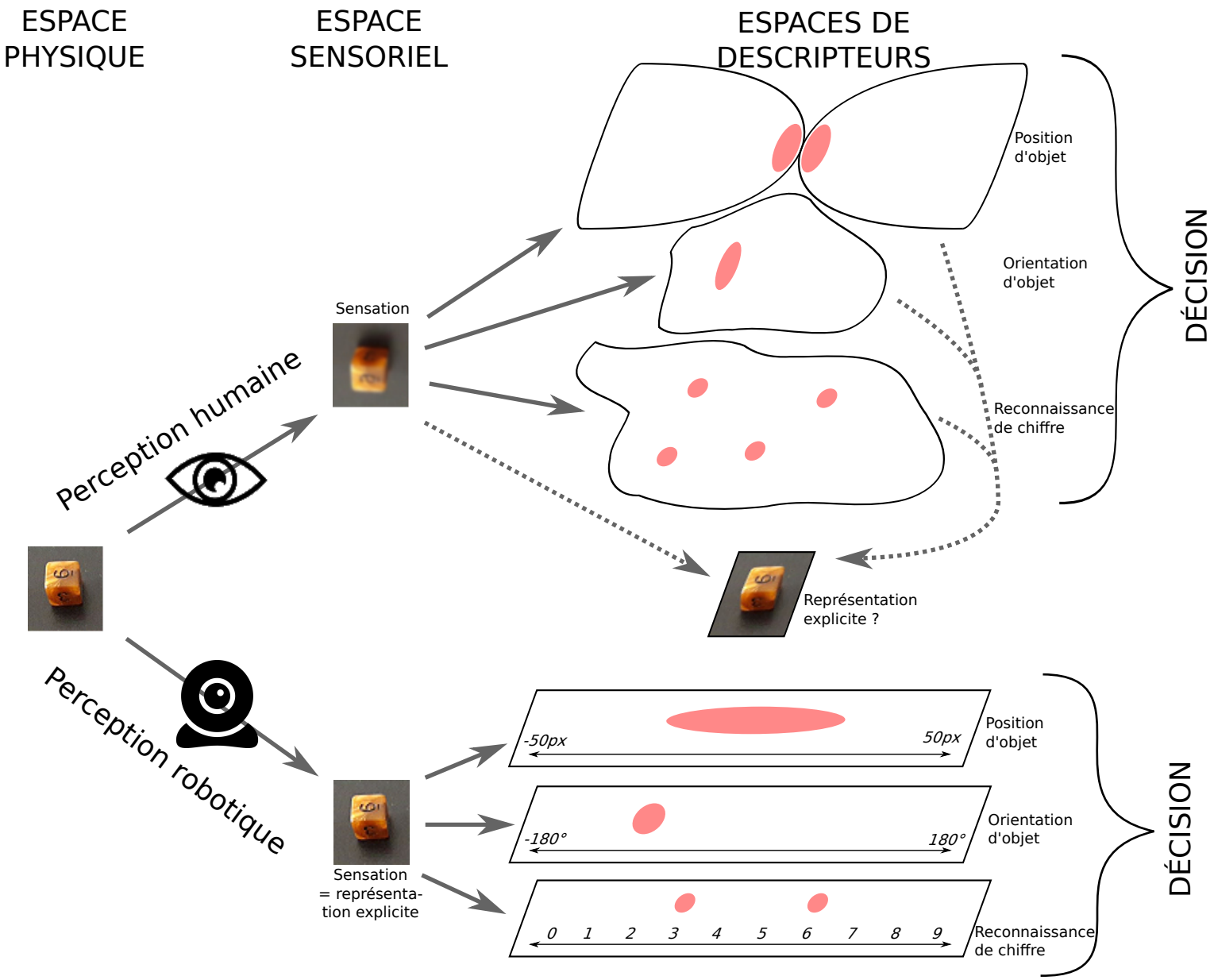


FIG. 8.3: Parallèles et différences entre perception biologique/humaine et perception artificielle/robotique. Par des processus complexes, les sensations sont interprétées sous forme de descripteurs encodés dans des cartes (neuronales ou computationelles) dédiées. Ces descripteurs jouent dans la prise de décision, qui en retour modifie le stimulus ainsi que l'intégration des percepts. L'existence dans le cerveau humain d'une représentation explicite de l'espace physique perçu ne fait pas l'unanimité, car la conscience qu'a l'humain de son espace perceptible peut être expliquée comme une aggrégation de connaissances de plus ou moins haut niveau. Cela motive notre positionnement qui ne repose pas tant sur des représentations figées que sur des topologies dédiées à certaines tâches et actions.

Deuxième contribution : Modèle bio-inspiré de fusion multimodale, application à l'effet ventriloque

En particulier, les DNF présentent des caractéristiques très intéressantes, notamment l'attention et la fusion de données en fonction de la distance et de la précision des stimuli. Dans cette deuxième contribution, nous utilisons ensuite le DNF comme un outil de filtrage et de fusion au sein d'un modèle neuro-inspiré de fusion multimodale. Ce modèle trouve notamment son inspiration dans le SC, une région subcorticale impliquée dans la génération de commandes de mouvement des yeux, et dont on sait qu'elle reçoit des signaux issus de modalités différentes. Une transformation dite logpolaire permet de modéliser la projection de signaux perçus par la rétine sur le SC, une projection non linéaire car les signaux plus proches de la fovéa activent plus de capteurs.

Nous montrons comment le modèle de fusion peut s'appliquer pour modéliser de manière réaliste des occurrences de l'effet ventriloque, un effet psychophysique de capture de localisation de stimuli audio ou visuels (le plus fiable capture l'autre ; ou une interpolation est faite s'ils ont la même fiabilité). Les résultats obtenus sont qualitativement semblables à d'autres modélisations plus classiques mais moins détaillées faites par estimation de maximum-vraisemblance. Une attention particulière est portée sur le choix des paramètres, et une analyse de sensibilité est faite pour montrer la marge de manœuvre existante dans l'optimisation potentielle de ceux-ci. Une étude de cette ampleur de l'effet des paramètres n'avait encore jamais été produite dans le domaine des DNF.

Troisième contribution : Apprentissage de topologies combiné à la fusion

Puis, afin d'étudier plus en détail le rôle des topologies sur ces tâches cognitives, une dernière contribution montre que les DNF conservent leurs propriétés dans des cartes topologiques irrégulières apprises. Dans cette expérience, les topologies sont apprises via un gaz neuronal croissant afin d'extraire les dimensions intrinsèques de l'espace sensoriel. Ensuite, une carte visuelle est jointe à une carte auditive pour tester des cas d'attention dans une nouvelle topologie multimodale.

En particulier, une expérience à partir de données de signaux auditifs captés par des robots, de haute dimension, produit une topologie sous-jacente de localisation 2D, dont la forme est très cohérente avec les modèles qualitatifs de localisation auditive. La fusion audiovisuelle se fait également de façon très cohérente, avec une favorisation de la modalité la plus précise, et une sélection des stimuli congruents en priorité sur les stimuli incongruents, avec une amélioration de la précision.

Conclusion et perspectives

La figure 8.4 propose une synthèse illustrative des contributions de cette thèse. En zoomant progressivement, nous avons d'abord une boucle de perception (colonne centrale) et d'action (flèche de droite). La perception sert à la prise de décision, qui motive l'action et modifie la perception[1]. Les sensations brutes sont pré-traitées par des mécanismes non développés dans cette thèse, pour obtenir des percepts à placer dans des topologies faites sur mesure. La contribution III précise une manière de créer des topologies à la fois unimodales et multimodale. Une topologie multimodale est utilisée dans une tâche

1. Des retours sensorimoteurs (abordés dans le manuscrit en perspectives) peuvent aussi influencer la fusion et la décision indirectement.

de localisation de stimuli audiovisuels. La fusion y est adaptée du modèle développé en contribution II, qui a été vérifié en modélisant avec succès l'effet ventriloque, proposant un nouveau modèle informatique de ce processus de prise de décision étudié en psychologie. La prise de décision est en principe modélisée sous différents prismes en fonction des domaines (psychologie / IA) et des objectifs des chercheurs. La contribution I unifie ces différents points de vue et propose une comparaison, sous un formalisme commun, de différents algorithmes de prise de décision.

Les perspectives autour de ces travaux s'articulent autour de deux axes principaux. D'une part, des modèles plus complexes peuvent être envisagés, en remplaçant les gaz neuronaux croissants par des modèles d'apprentissage profond. Ainsi, un modèle de fusion multimodale pourrait s'appliquer à des tâches de haut niveau telles que la reconnaissance d'émotions, en cumulant les traitements visuel (traits du visage), auditif (timbre de la voix) et linguistique (contenu des paroles).

D'autre part, les topologies nouvellement créées peuvent servir de support à des actions, par exemple des mouvements des yeux en direction d'un stimulus. Les arêtes au sein d'un graphe peuvent être traduites comme des commandes motrices permettant de faire la transition entre deux états perceptifs. En ajoutant à la croissance des gaz neuronaux des règles inspirées du codage prédictif, les actions d'un agent artificiel pourraient prendre directement part à la création des topologies. Ce modèle pourrait alors servir de tremplin à une forme de cognition incarnée.

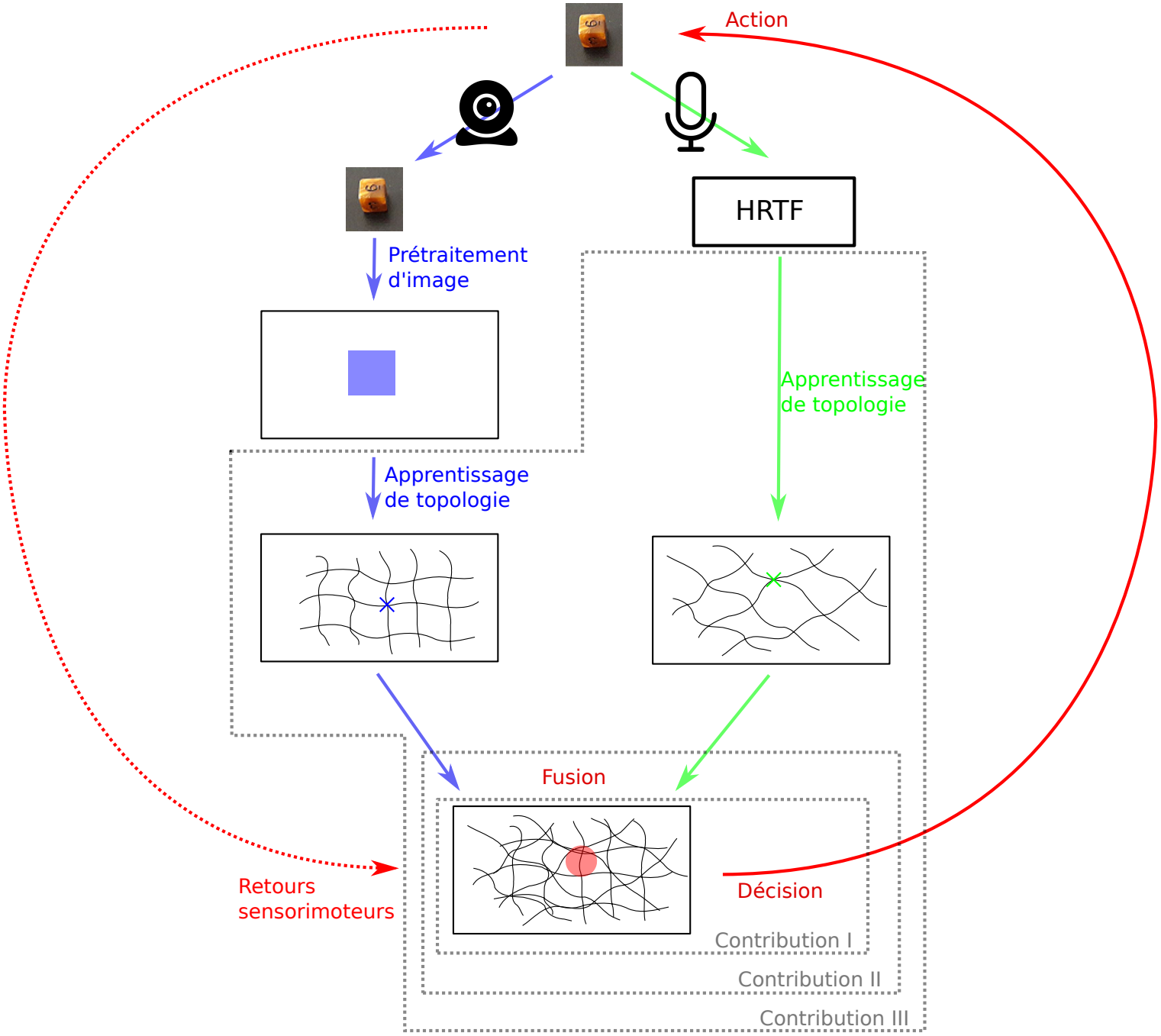


FIG. 8.4: Synthèse des mécanismes abordés dans chaque contribution