Implication relative des traits de haut niveau et de bas niveau des stimuli dans la catégorisation, chez l'homme et le singe

Anne-Claire Collet

Université Fédérale
Toulouse Midi-Pyrénées

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par :

Université Toulouse 3 Paul Sabatier (UT3 Paul Sabatier)

Présentée et soutenue par :
Anne-Claire COLLET

le vendredi 12 février 2016

Titre :

Implication relative des traits de haut niveau et de bas niveau des stimuli dans la catégorisation, chez l'homme et le singe

---

École doctorale et discipline ou spécialité :

ED CLESCO : Neurosciences, comportement et cognition

Unité de recherche :

CERCO UMR 5549

Directeur/trice(s) de Thèse :

Rufin VANRULLEN

Jury :

Rapporteurs:

Pascal MAMASSIAN, Laboratoire des systèmes perceptifs (CNRS UMR 8248), Paris
Muriel BOUCART, Laboratoire de neurosciences fonctionnelles et pathologie, Université de Lille

Autre membre du jury:

Céline CAPPE, CERCO CNRS UMR 5549, Toulouse

Épigraphe pour un livre condamné

Lecteur paisible et bucolique,
Sobre et naïf homme de bien,
Jette ce livre saturnien,
Orgiaque et mélancolique.

Si tu n'as fait ta rhétorique
Chez Satan, le rusé doyen,
Jette ! tu n'y comprendrais rien,
Ou tu me croirais hystérique.

Mais si, sans se laisser charmer,
Ton oeil sait plonger dans les gouffres,
Lis-moi, pour apprendre à m'aimer ;

Ame curieuse qui souffres
Et vas cherchant ton paradis,
Plains-moi !... sinon, je te maudis !

Baudelaire, Les fleurs du Mal

Remerciements

Je tiens tout d'abord à remercier Rufin VanRullen, mon directeur de thèse, qui m'a encadrée pendant ces trois années. Bien que ma thèse ne se situe pas dans son domaine de prédilection (les oscillations…) il m'a toujours guidée et aidée dans mon travail. Ce fut pour moi un réel plaisir d'être supervisée par lui, même s'il a souvent été frustrant de constater à quel point sa réflexion était plus rapide que la mienne. Merci beaucoup Rufin !

Je remercie Céline Cappe, avec qui j'ai collaboré pour la troisième étude de cette thèse, pour ses précieux conseils et son enthousiasme !

Je remercie Denis Fize qui m'a acceptée en stage de M1 il y a maintenant 5 ans. C'est avec lui que j'ai fait mes premiers pas au Cerco, avec les singes. Et j'y suis restée…

Je me dois également de remercier tous les thésards et post-docs qui m'ont aidée durant ces trois ans, en commençant par Roger Koenig-Robert dont j'ai réutilisé la technique qu'il avait développée pendant sa thèse, et auprès de qui j'ai toujours pu trouver du soutien et des explications. Je remercie aussi Benedikt Zoefel pour son aide dans le domaine auditif (dans lequel j'étais et reste un peu néophyte), ainsi que Manuel Mercier pour les discussions très enrichissantes au sujet de l'intégration multisensorielle.

Je remercie Max pour son aide précieuse avec mes dispositifs expérimentaux, et pour tous les cafés partagés. Max, tu m'as sauvé la vie plus d'une fois !

Je remercie tous les étudiants du Cerco (passés ou actuels) : Fanny, Marissa, les 2 Marie, Marcello, Marina, Mehdi, Doug, Damien, Sasskia, Grace, Tracy, les deux Nico, Rasa, et tous les autres pour les moments de détente et de convivialité.

Et à tous ceux qui m'ont à un moment ou un autre donné un coup de main, qui ont partagé une pause, un café, un sourire : Merci ! Grâce à vous tous j'ai passé des années formidables au Cerco !

Enfin je remercie mes singes, mes bébés : Rox, Roucky, Dicky et Prosper. Sans eux, cette thèse n'aurait pas pu exister ! Plein de lipsmack !

Mais parce qu'une thèse ne se vit pas que dans un labo, je veux aussi remercier ma famille et mes amis :

A mes parents qui m'ont soutenue tout au long de mes études, c'est en grande partie grâce à eux que j'ai pu arriver jusque-là. A mon père qui m'a répété 1000 fois quand j'étais ado à quel point il était crucial que je parle anglais : Papa, tu avais raison !!! ça m'a bien servi… A ma mère qui a passé de longues heures au téléphone à écouter mes jérémiades quand j'avais un coup de mou. Merci !

A ma marraine Hélène, qui m'a toujours encouragée à m'accrocher à mes rêves, et qui m'a souvent aidée financièrement pour les réaliser. A elle qui est un modèle intellectuel depuis longtemps pour moi !

A ma cousine Irène, grâce à qui il y a huit ans j'ai pu faire un stage au GIN à Grenoble, et ainsi pour la première fois découvrir le milieu de la recherche. Une cousine, une confidente, un grand soutien ces dernières années.

A ma Loulou, bien que loin, toujours très présente dans ma vie depuis 25 ans !

A Eric, le voisin, avec qui j'ai passé deux années et demi géniales dans notre petit immeuble, avant qu'il ne m'abandonne lâchement ;-)

Au groupe des pink poleuz (et poleur) Maureen, Pauline, Charlotte, Olivia et Fred ! Mes années de thèse ont été l'occasion de découvrir un sport super, et des nanas (et Fred) encore mieux ! Sans vous, je n'aurais jamais été aussi sereine pendant ces trois ans !

Et à tous les autres que j'oublie (sur le papier) !

Mes années de thèse ont été parmi les meilleures de ma vie. Je me suis éclatée ! Et me voilà bien triste de les voir achevées…

Sommaire

Introduction 9

I) Définition de la catégorisation 9

II) La catégorisation visuelle 11

A) La catégorisation chez l'animal 11

B) La comparaison homme/singe 13

C) Le substrat neural de la catégorisation visuelle chez le primate : le cortex inférotemporal ? 14

D) L'aire IT et l'organisation des catégories chez l'homme et le singe 17

III) La catégorisation auditive 19

A) La catégorisation des sons en psychophysique chez l'homme 20

B) Corrélats neuraux de la reconnaissance des sons chez le primate et l'homme 22

IV) Les effets de contexte 24

A) Au sein d'une modalité 24

1) L'influence du contexte dans la reconnaissance d'images 25

2) L'influence du contexte dans la reconnaissance des sons 28

B) Multisensoriel : quand une modalité devient le contexte de l'autre 29

V) Les traits diagnostiques des catégories : haut niveau versus bas niveau ? 31

A) Définition haut niveau et bas niveau en audition et vision 31

B) Contribution des caractéristiques de bas niveau en catégorisation visuelle 32

1) Les caractéristiques de haut et bas niveau dans le domaine visuel 32

2) Le contrôle des caractéristiques de bas niveau en psychophysique 33

C) Caractéristiques acoustiques et catégorisation auditive 35

1) Les caractéristiques du son traitées par la cochlée 35

2) De la cochlée à l'aire A1 36

3) Les études en psycho-acoustique 37

D) Le contrôle des paramètres de bas niveau dans l'exploration du rôle du contexte 38

VI) Problématiques de la thèse 39

Chapitre 1 : Les corrélats neuraux de la reconnaissance d'objets dans une tâche de catégorisation. 41

I) Les corrélats neuraux de la reconnaissance d'objets chez l'homme 42

A) La création des séquences SWIFT 42

B) Protocole 43

C) Résultats 44

II) Les corrélats de la reconnaissance d'objets chez le singe Macaque 47

A) Introduction 47

1) La mesure de la conscience 47

2) L'électroencéphalographie chez le singe vigile 49

3) La catégorisation visuelle : tâche cognitive pour accéder à la perception consciente chez le singe 49

4) Les bases neurales de la catégorisation visuelle 50

5) Problématique et protocole 51

B) Matériel et méthodes 53

6) Le sujet 53

7) L'implantation des électrodes 53

8) Les stimuli 54

9) La tâche 58

10) Enregistrement des données 60

11) Analyse des données 61

C) Résultats 62

1) Résultats comportementaux 62
a) Performances globales 62
b) Performances sur les images nouvelles 62
c) Apprentissage 62
2) Résultats électrophysiologiques 63
a) Résultats généraux 63
b) Analyse par item unique 67
D) Discussion 69
1) Résultats généraux 69
2) Parallèle Homme/singe 69
3) Théorie de la perception consciente : Neuronal Global Workspace 71
Chapitre 2: L'effet de congruence contextuelle sur la catégorisation d'objets chez l'Homme et le singe Macaque 74
Chapitre 3 : L'importance de la congruence sémantique dans la catégorisation multimodale chez l'homme 104
I) Introduction 105
II) Matériel et méthodes 109
A) Les sujets 109
B) Les stimuli 109
1) Les catégories 109
2) La construction des séquences d'évènements 109
3) Les séquences visuelles 110
4) Les séquences auditives 111
C) Le protocole 113
D) Analyses 114
1) Vincentisation 114

2) Probabilité cumulée de réponses 115
3) Théorie de la détection du signal et d' cumulé 115
4) Race model 115
III) Résultats 117
A) Résultats généraux : effet de congruence 117
B) L'intégration multi-sensorielle : le Race model 118
C) L'importance de la congruence sémantique 119
D) Les différences entre catégories 120
1) Taux de réussite et race model 120
2) d' observé et d' prédit par le race model 122
IV) Discussion 124
Discussion générale 128
I) Résumé des résultats 128
II) La contribution des statistiques de haut et bas niveaux chez l'homme et le singe .. 129
A) Le « haut niveau » peut-il tout expliquer chez l'homme ? 130
B) Les singes sont-ils davantage sensibles au « bas niveau » ? 134
III) Ouvertures 136
A) Apport des neurosciences cognitives comparées et la question de la généralisation 136
B) La catégorisation : un prétexte pour étudier la reconnaissance des objets 138
Références bibliographiques 141

Introduction

Nous évoluons sans cesse dans un environnement semé d'embûches, qu'il s'agisse de la jungle urbaine ou tropicale. Pour survivre, il nous faut identifier rapidement les menaces et adopter le comportement le plus approprié. L'ensemble de nos sens sont sollicités au quotidien pour nous fournir une description la plus complète et fidèle possible de ce qui nous entoure. Nous, humains, comme nombre de primates, sommes diurnes et trichromates. La vision est donc le sens que nous utilisons le plus communément pour analyser notre environnement. Toutefois, il n'est pas rare que la situation n'en permette pas l'usage, nous amenant alors à faire appel à d'autres sens… Imaginez, à un croisement, vous ne voyez pas bien, mais un vrombissement se rapproche rapidement, vous ne vous engagez pas ! Ou encore, vous êtes chez vous où tout semble en ordre, mais une forte odeur de gaz se fait sentir : vous sortez précipitamment ! En quelques dixièmes de seconde vos systèmes sensoriels ont donc transmis un signal à votre cerveau qui l'a identifié et vous a dicté la réaction appropriée. Ce ne sont là que quelques exemples, et le plus souvent plusieurs voies sensorielles sont impliquées simultanément dans la reconnaissance des objets (ou des risques) qui nous entourent. Il est donc indispensable de pouvoir combiner correctement ces différentes informations pour former une représentation cohérente du monde. De plus, dans des situations d'urgence, nos représentations mentales doivent être accessibles rapidement. Nul besoin d'identifier la source du danger en détail ! Peu importe que la voiture qui arrive à toute allure soit une Renault ou une Peugeot, ou bien même qu'il s'agisse d'un bus, d'un camion, d'une moto ! C'est un véhicule qui peut porter atteinte à notre intégrité physique, nous devons donc l'éviter.

I) Définition de la catégorisation

Notre cerveau a donc cette capacité à catégoriser les objets, c'est-à-dire à nous en donner une représentation globale. La catégorisation est le processus cognitif qui nous permet de classer et réunir les objets en groupes distincts. Au sein de chaque groupe, ou catégorie, les items partagent un certain nombre de caractéristiques communes, plus ou moins

abstraites. L'utilisation de catégories permet de réduire la demande en ressource cognitive impliquée dans la reconnaissance des objets, mais suppose des capacités d'abstraction et de généralisation. On distingue trois niveaux de catégories (Rosch et al., 1976, Thompson, 1995, Zayan and Vauclair, 1998) faisant intervenir un degré d'abstraction décroissant :

- **Le niveau super-ordonné** : A ce niveau catégoriel, l'hétérogénéité physique entre les membres est importante, il faut donc faire abstraction du physique pour extraire les caractéristiques générales de la catégorie. On peut ainsi dire que la catégorie « Animal » est au niveau super-ordonné.
- **Le niveau de base** : Il s'agit du niveau catégoriel le plus intuitif chez l'homme, qui correspond aux catégories sémantiques, le degré d'abstraction requis est moindre, les membres de la catégorie partagent beaucoup de traits physiques caractéristiques. La catégorie « Chien » est un sous-ensemble de la catégorie « Animal ». Les membres de cette catégorie sont tous des mammifères quadrupèdes, possédant une queue et ayant comme vocalisation caractéristique l'aboiement.
- **Le niveau subordonné** : A ce niveau, les membres de la catégorie présentent une forte homogénéité physique, le niveau d'abstraction requis est faible. La sous-catégorie « Cocker » se situe par exemple à ce niveau. Et l'on peut le décrire par des qualificatifs précis de taille, couleur, forme, etc. Toutefois, catégoriser au niveau subordonné peut requérir dans certains cas une certaine expertise dans le domaine.

Les catégories sont donc hiérarchisées selon un principe taxonomique. Si la variabilité est de mise aux niveaux super-ordonné et de base, elle est nettement moindre au niveau subordonné. Toutefois, même pour une reconnaissance d'objets au niveau subordonné, l'apprentissage « par cœur » ne serait pas efficace. En effet, en situation écologique, l'angle de vue, la luminosité, la position dans l'espace, le niveau de bruit, la distance par rapport à l'objet changent constamment. Nos systèmes perceptuels font donc face à un challenge permanent dans un environnement en perpétuel renouvellement.

II) La catégorisation visuelle

La catégorisation est donc une faculté qui peut sembler triviale de prime abord même si elle relève en réalité de processus cognitifs complexes ! C'est en simplifiant le monde que les chercheurs ont pu commencer à élaborer des théories sur les mécanismes sous-tendant cette aptitude. C'est d'abord dans la modalité visuelle que les capacités de catégorisation des hommes et des animaux ont été largement explorées.

A) La catégorisation chez l'animal

Comme pré-requis à toute investigation, il fallait d'abord s'assurer que nos amies les bêtes étaient douées de cette même faculté à organiser le monde en catégories.

Les premières observations de reconnaissance d'images par un primate non-humain ont été faites sur une femelle chimpanzé élevée comme un enfant par un couple d'américains (Hayes and Hayes, 1953). Après avoir observé un intérêt spontané de la jeune guenon pour des livres illustrés ou des catalogues, ces deux primatologues testèrent ses capacités à discriminer les images. Pour cela, ils choisirent d'utiliser pour chaque test des images d'objets appartenant à deux catégories distinctes au niveau de base, par exemple *chaise* versus *voiture.* Les images étaient imprimées sur des cartes, et il pouvait s'agir de photographies, de dessins réalistes ou encore de dessins contourés. Bien que la guenon ait fait quelques erreurs, ses performances (96% de réussite) ont démontré sans équivoque qu'elle était capable de catégoriser des images. Si cette étude démontrait pour la première fois la capacité d'un animal à reconnaître des images et à les classer selon un principe catégoriel, il s'agissait là des facultés cognitives du primate le plus proche de l'Homme. Ce n'est que 10 ans plus tard que des études similaires furent réalisées sur le pigeon, un animal très éloigné de l'Homme dans l'arbre phylogénétique supposé avoir des capacités cognitives largement inférieures. Herrnstein et Loveland (Herrnstein and Loveland, 1964) ont ainsi montré que des pigeons étaient capables de catégoriser des images sur la base de la présence ou de l'absence d'un être humain. Ces deux chercheurs ont prouvé que les pigeons étaient aptes à former des concepts qui ne varient pas en nature de ceux utilisés par l'Homme. Toutefois l'être humain, ayant accès au langage, est doté d'un esprit d'abstraction plus poussé et est alors capable de catégoriser plus finement les objets du monde qui l'entourent. Ces études ont ultérieurement été

approfondies par d'autres équipes, afin notamment de conforter ces premiers résultats mais aussi de déterminer les caractéristiques sur lesquelles ces animaux s'appuyaient pour catégoriser des images (Roberts and Mazmanian, 1988, Aust and Huber, 2001). Il en ressort que l'apprentissage de la règle de catégorisation n'est possible que si les stimuli renforcés positivement sont ceux qui contiennent la cible de la tâche. Par là s'entend qu'il est très difficile pour des pigeons d'apprendre à catégoriser des images sur la base de l'absence d'un stimulus, s'ils sont engagés dans une tâche de catégorisation go/no go « humain » versus « non-humain », ils ne peuvent apprendre la règle de catégorisation que si la réponse go est requise pour les stimuli contenant les figures humaines. La couleur n'apparaît par contre pas comme un critère crucial pour ces oiseaux.

Si des pigeons sont capables de catégoriser des images, il semble évident que les singes y soient également aptes puisqu'ils présentent des capacités cognitives supposées plus importantes. Les premières observations de catégorisation spontanée ont été réalisées par Seyfarth et collègues (Seyfarth, Cheney et al. 1980) sur le singe vervet en milieu naturel. Cette espèce de primates produit différentes vocalisations d'alarme en fonction du type de prédateur. Le groupe réagit alors en fonction du cris émis : lorsque le prédateur est un félin, le reste du groupe s'empresse de rejoindre les arbres ; lorsque c'est un oiseau de proie, les autres individus du groupe guettent alors le ciel et se réfugient dans les feuillages ; si la menace est reptilienne, les singes répondent alors au cri d'alarme en scrutant le sol. Les singes vervets catégorisent donc le type de prédateur, signalent vocalement et spécifiquement au reste du groupe la menace, et le reste des individus adopte alors des comportements de fuite et de vigilance adaptés. Parallèlement à ces observations de terrain, les neuroscientifiques se sont penchés sur les facultés de catégorisation visuelle des singes, en conditions expérimentales contrôlées. Dans les années 80, Schrier et ses collègues (Schrier et al., 1984, Schrier and Brady, 1987) ont démontré que des macaques étaient capables de catégoriser des images sur la base de la présence d'un être humain ou d'un singe dans des tâches de choix forcé. Ils se sont également intéressés aux critères diagnostic permettant aux macaques de réaliser correctement cette tâche. Ils ont ainsi altéré l'orientation de la figure humaine dans l'image, détruit par scrambling certaines parties du corps ou retourné le visage. Malgré ces altérations physiques des stimuli, les singes répondaient toujours à un niveau

significativement plus élevé que la chance aux images contenant un humain. Toutefois, leurs performances à la tâche de catégorisation s'en trouvaient diminuées. Lorsqu'on leur donnait le choix entre l'image intacte et l'image altérée, les sujets choisissaient de manière quasi systématique l'image intacte. Les singes avaient donc appris à reconnaître visuellement les traits humains dans une image, mais aussi la conformation correcte de ces traits. Comme chez le pigeon, les singes ont plus de facilité à apprendre la règle de catégorisation quand le stimulus renforcé positivement est celui qui contient l'objet à catégoriser (D'Amato and Van Sant, 1988). Si ces premières études ont eu l'indéniable avantage de montrer que les primates étaient capables de former des catégories, il n'en reste pas moins que les stimuli utilisés dans les tâches de catégorisation partageaient de très fortes similarités physiques, puisque le niveau catégoriel étudié était le niveau de base. La catégorisation de ces images ne requérait donc pas un haut degré d'abstraction. En 1998, Fabre-Thorpe et collègues ont exploré l'existence de concepts plus abstraits en testant deux macaques sur la catégorisation d'images contenant ou non de la nourriture, pour l'un, ou un animal, pour l'autre. Ces deux catégories sont vastes et comprennent des items variant drastiquement en termes de forme, couleur, taille : entre une barquette de frites et une pomme, la ressemblance ne saute pas aux yeux, de même qu'entre un criquet et une girafe… Pourtant les singes ont réussi à apprendre la règle de catégorisation et à la généraliser à de nouvelles images (Fabre-Thorpe et al., 1998). Ils ont donc formé des concepts abstraits et n'ont pas pu s'appuyer sur la similarité physique des stimuli.

B) La comparaison homme/singe

La capacité de catégorisation des primates est certes passionnante en soi mais son étude trouve aussi son intérêt dans la comparaison avec l'homme. L'expérimentation animale n'a de sens que si les résultats peuvent servir à la compréhension de notre espèce, et dans le cas qui nous occupe, dans la compréhension de nos facultés cognitives dont le substrat est le cerveau.

Dans leur étude publiée en 1988, Robert et Mazmanian ont comparé les facultés de catégorisation de l'homme avec celles de deux espèces animales : le pigeon et le saïmiri

(*saimiri sciureus*), petit singe du nouveau monde. Cette étude explorait chez ces trois espèces l'existence du concept « animal » à différents niveaux d'abstraction:

- Au niveau sous-ordonné, la catégorie cible était le martin-pêcheur, les distracteurs étant d'autres oiseaux.
- Au niveau de base, la catégorie cible était les oiseaux, les distracteurs étaient d'autres espèces animales.
- Enfin au niveau super-ordonné, la catégorie cible était les animaux et les distracteurs des objets.

Le résultat principal de cette étude était le très fort taux de réussite des trois espèces dans les trois niveaux de catégorisation. Toutefois des différences plus fines se dégageaient : ainsi le pigeon et le saïmiri étaient très à l'aise avec la catégorisation au niveau sous-ordonné alors que c'est le niveau qui posait le plus de difficultés aux sujets humains. Par ailleurs les 2 espèces animales montraient des performances amoindries pour la catégorisation au niveau de base alors que les hommes réussissaient parfaitement la tâche. Enfin au niveau super-ordonné, le singe et l'homme ont démontré une grande aisance alors que les pigeons étaient mis en difficulté.

Des années plus tard, Fabre-Thorpe et collègues (Fabre-Thorpe et al., 1998, FABRE-THORPE, 2003a) comparaient les performances des macaques et des sujets humains lors de tâches de catégorisation au niveau super-ordonné. Cette équipe a alors mis en lumière de fortes similitudes dans les processus mis en jeu, bien que les singes se soient révélés plus rapides et moins précis que les sujets humains.

C) Le substrat neural de la catégorisation visuelle chez le primate : le cortex inférotemporal ?

En plus d'être des animaux faciles à élever en laboratoire, les macaques sont des singes particulièrement indiqués pour la comparaison des capacités visuo-cognitives du fait de la forte similarité entre leur cortex visuel et le nôtre. Les chercheurs Hubel et Wiesel ont exploré depuis les années 1960 l'anatomie et la spécialisation des aires corticales du chat et du singe. Ils ont mis en évidence de remarquables parallèles anatomiques et

fonctionnels chez le macaque et l'homme. Il ont également proposé le modèle hiérarchique de la voie visuelle qui a révolutionné les neurosciences et ouvert de nouvelles perspectives (Hubel and Wiesel, 1968, 1970, 1974, 1977, Hubel and Wiesel, 1979). La voie visuelle ventrale se termine par le cortex inférotemporal (désigné IT dans le reste du manuscrit), selon le modèle hiérarchique de cette voie, et c'est dans cette aire que seraient encodées les représentations les plus complexes des objets.

A la même époque que Hubel et Wiesel, Gross a commencé à explorer le rôle de ce cortex inferotemporal dans la reconnaissance des objets chez le macaque. Ses premiers résultats ont montré que les neurones de IT avaient un champ récepteur très étendu (Gross et al., 1969), soutenant le modèle hiérarchique avec un accroissement de la taille des champs récepteurs de la voie visuelle ventrale. Quelques années plus tard, ce même chercheur montrait que les neurones de IT, en plus d'avoir de larges champs récepteurs, présentaient une sélectivité pour des formes spécifiques, certaines très simples (comme des cercles ou des rectangles), d'autres plus complexes (comme la forme d'une main). En outre, il a découvert que ces mêmes neurones présentaient une invariance d'activité pour la taille des stimuli (Gross et al., 1972). Ces résultats désignaient alors IT comme un candidat très probable de l'encodage de la représentation des objets et des catégories.

Dans les années 1990, les recherches sur le rôle du cortex inférotemporal s'intensifièrent, les travaux sont nombreux et nous ne pourrons pas en faire un inventaire exhaustif. Tanaka, par exemple, a montré que non seulement les neurones de IT ont une activité spécifique de certaines formes mais qu'ils sont de surcroît organisés en colonnes fonctionnelles, et qu'au sein d'une même colonne, les neurones présentent une sélectivité de forme similaire les uns avec les autres (Tanaka, 1996). Toutefois dans cette étude, les stimuli utilisés étaient assez simples et ne représentaient pas des objets réels. La même année, Logothetis mit en parallèle les capacités de catégorisation de l'homme et du singe ainsi que l'implication de IT dans leur réalisation. Il a alors proposé que la combinaison des activités des neurones de IT serait le substrat de la reconnaissance des objets et de leur catégorisation (Logothetis and Sheinberg, 1996).

Avec l'évolution des techniques d'imagerie, les études comparatives neurologiques, comportementales et cognitives entre l'homme et le macaque se sont multipliées. En 1999, Vogels (Vogels, 1999a, b) a corrélé les aspects comportementaux et neurologiques de la catégorisation chez le macaque rhésus. Il a mis pour cela en place une expérience de

catégorisation par choix forcé en saccade oculaire, soit “arbre” versus “non-arbre” ou “poisson” versus “non-poisson”. Cette fois, les stimuli utilisés étaient des images complexes représentatives du monde réel. Il s’agissait de photographies en couleurs d’arbres ou de poissons pour les cibles, les distracteurs de la tâche pouvaient être aussi bien des objets que des animaux, mais présentaient des caractéristiques physiques similaires (forme, texture ou couleur) aux items de la catégorie cible, interdisant aux singes de baser leur décision sur de tels critères. Après avoir démontré que les singes catégorisaient correctement les images, avec un taux de transfert important à de nouveaux stimuli, Vogels a réalisé dans la seconde partie de son étude des enregistrements unitaires de neurones dans le cortex inféro-temporal, aire terminale de la voie visuelle ventrale encodant la reconnaissance des objets. Il a trouvé dans cette aire bon nombre de neurones déchargeant spécifiquement pour une certaine catégorie mais pas pour d’autres. De plus, ces neurones avaient une activité très similaire pour tous les items appartenant à leur catégorie préférentielle quand ceux-ci étaient correctement catégorisés au niveau comportemental par le singe. Cette étude confirmait donc le lien direct entre la réponse comportementale et la réponse neuronale, désignant l’aire IT (inféro-temporale) comme le substrat le plus plausible de la catégorisation visuelle d’objets complexes.

D’autres équipes ont ensuite continué à chercher les aires corticales impliquées dans la catégorisation visuelle. Thorpe et Fabre-Thorpe ont proposé un modèle de la catégorisation visuelle traçant le trajet du signal nerveux de l’œil jusqu’à la réponse motrice chez le singe. Selon leur modèle, la catégorisation visuelle rapide serait un mécanisme principalement feed-forward, c'est-à-dire des aires corticales de bas niveau vers les aires de haut niveau. Ils ont ainsi proposé que l’information perçue par la rétine est envoyée vers les aires visuelles primaires qui la font ensuite remonter le long de la voie visuelle ventrale jusqu’à IT où sont stockées les représentations de haut niveau des objets. L’aire IT envoie ensuite l’influx nerveux vers le cortex préfrontal, aire où la décision catégorielle serait prise, qui ensuite, dans le cadre d’une tâche où une réponse motrice est attendue, donne l’information au cortex moteur (cf figure I.1).

Enfin les études neuropsychologiques chez l’homme apportent des arguments de poids en faveur de la théorie selon laquelle les neurones de IT supporteraient notre capacité à

catégoriser, puisque chez des patients lésés dans cette aire, il a été plusieurs fois reporté des agnosies visuelles catégorielles (Arguin et al., 1996).

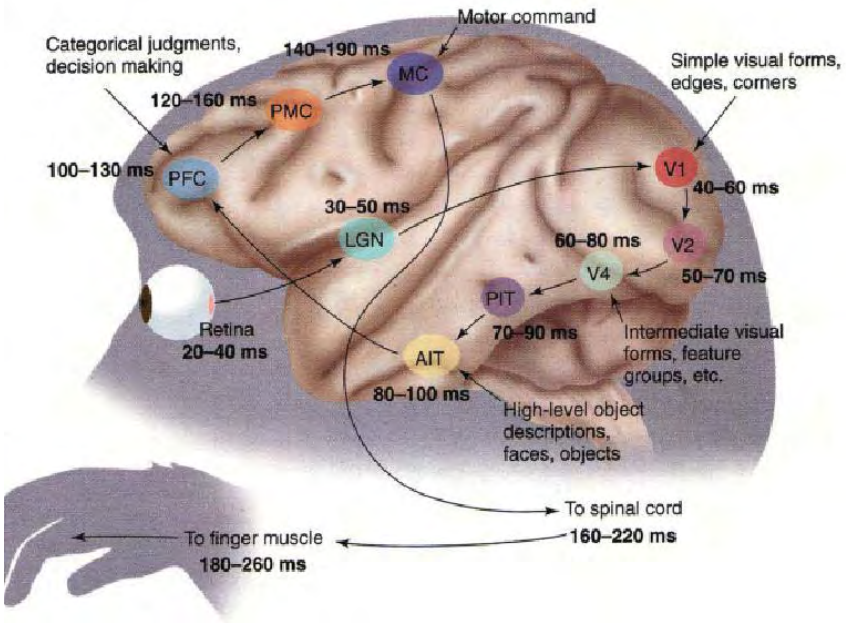


Figure I.1 : Chemin probable et latence de traitement de l'information visuelle chez le singe engagé dans une tâche de catégorisation visuelle rapide (Thorpe and Fabre-Thorpe, 2001)

D) L'aire IT et l'organisation des catégories chez l'homme et le singe

Vers la fin des années 2000, des chercheurs tels que Kriegeskorte et Kiani ont comparé les catégories formées dans IT chez l'homme et le singe. Pour ce faire, Kriegeskorte et collègues (Kriegeskorte et al., 2008a) ont enregistré l'activité simultanée de plus de 300 neurones de IT chez deux macaques (674 neurones enregistrés au total) lors d'une tâche de fixation passive durant laquelle 92 images d'objets étaient présentées. A partir de la distribution de l'activité évoquée par ces 92 stimuli, ils ont construit des matrices de dissimilarité qui permettent de grouper entre eux des stimuli générant une activité similaire. Des sujets humains ont été soumis à la même tâche de fixation, durant laquelle leur activité cérébrale était enregistrée par IRM fonctionnelle. Les stimuli ont ensuite été

regroupés suivant la même technique des matrices de dissimilarité. Les résultats sont édifiants (cf figure I.2) ! Ils montrent en effet que les items appartenant à une même catégorie évoquent des activités similaires et, chose plus importante, que les regroupements catégoriels sont très semblables entre l'homme et le singe ! Il y a ainsi chez les deux espèces une frontière nette entre les items animés et les items non-animés. De plus, il y a une subdivision au sein de la catégorie "animé"entre les faces humaines et animales ainsi qu'entre les corps entiers ou les parties du corps. Au sein de la catégorie "non-animé", les subdivisions entre les objets naturels et artificiels sont moins nettes, et ce, chez les deux espèces.

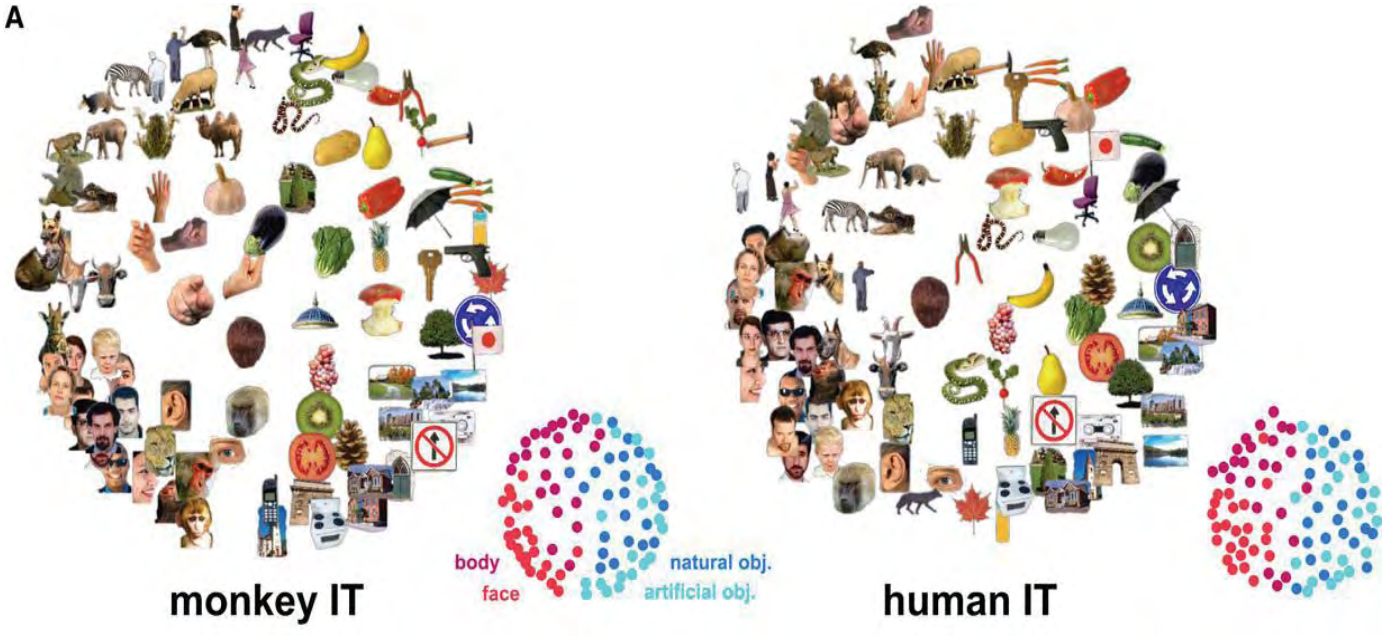


Figure I.2: Arrangement des stimuli reflétant la similarité des patterns d'activation qu'ils génèrent dans l'aire IT du macaque et de l'homme (tiré de Kriegeskorte et al. 2008).

L'année précédente, Kiani et collègues (KIANI et al., 2007a) avaient également enregistré des neurones dans l'aire IT du macaque et proposé un arbre taxonomique des catégories reconstruit à partir de la similarité des distributions d'activité des neurones. Grâce à cet arbre, nous pouvons noter que les catégories formées dans le cortex inférotemporal du singe correspondent remarquablement aux catégories sémantiques élaborées par l'homme.

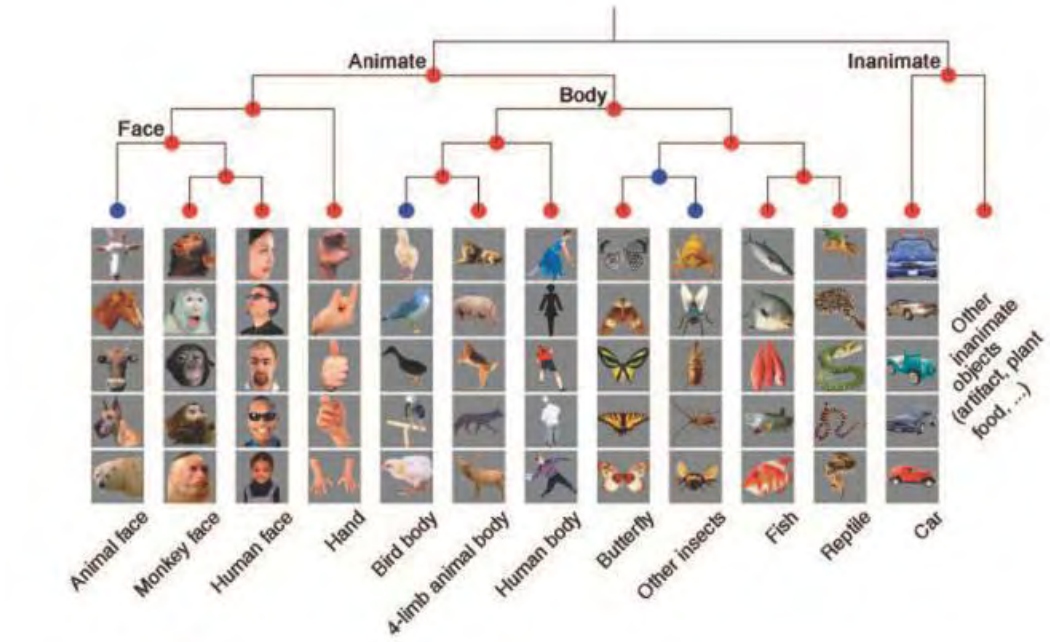


Figure I.3: Arbre des catégories reconstruit grâce à la similarité des patterns d'activation dans IT chez le singe (Kiani et al, 2007)

Les primates sont donc capables de catégoriser très rapidement les objets visuels auxquels ils sont confrontés. Même si leurs performances comportementales sont similaires à celles des hommes, il apparaît en plus que les catégories qu'ils élaborent sont de même nature.

III) La catégorisation auditive

Comme dans le domaine visuel, nous sommes capables de catégoriser des objets auditifs. Toutefois cette catégorisation est plus ambigüe dans le domaine de l'audition car elle peut se faire sur deux aspects du son, comme le définit Gaver (Gaver, 1993) :

- Soit nous catégorisons le son sur la base de ses propriétés accoustiques, le son peut par exemple être jugé grave ou aigu, rythmé et répétitif ou variable, faible ou fort. Gaver définit cette écoute comme musicale (*musical listening*). Ceci revient à trouver dans des sons différents certains invariants qui permettent de les identifier comme appartenant à la même catégorie.
- Nous pouvons également catégoriser le son en fonction de la source du-dit son: bruit de moteur, cri d'animal, son d'instrument de musique, bruit de pas, etc.

Dans cette seconde approche, Gaver parle d'écoute quotidienne (*everyday listening*).

Dans ce deuxième cas, et c'est celui que nous étudierons, les propriétés accoustiques peuvent varier drastiquement. La catégorisation requiert donc davantage d'abstraction: il faut associer le son avec l'objet, l'être vivant ou même l'action qui l'a produit, et catégoriser cet objet. Catégoriser un son repose donc sur la capacité du système auditif à détecter, extraire, distinguer et grouper des régularités spectrotemporelles d'un évènement accoustique en des unités perceptuellement stables. L'abstraction intervient donc à deux niveaux : la représentation mentale de la causalité du son et l'inclusion de l'objet dans une catégorie. C'est sans doute pour cette raison que les études de catégorisation de sons naturels chez l'animal sont rares, hormis en ce qui concerne la reconnaissance des vocalisations intraspécifiques. Nous nous concentrerons donc principalement sur la catégorisation de sons chez l'homme mais évoquerons quelques études réalisées chez le primate non-humain. Et n'oublions pas les observations de Seyfart : certes les animaux catégorisaient visuellement le prédateur, mais les autres individus du groupe répondaient à son alarme par un comportement adapé. Ils étaient donc capables de catégoriser les vocalisations.

Contrairement aux images, les sons ont en outre une valence émotionnelle importante qui repose notamment sur leurs propriétés accoustiques : un son peut-être jugé particulièrement aversif bien qu'il ne soit pas lié à une menace, comme par exemple le crissement des ongles sur un tableau noir, le grincement d'une fausse note sur un violon… La valence émotionnelle négative des sons est d'ailleurs largement utilisée notamment pour la création d'alarmes (Bergman et al., 2009). Lors des études de catégorisation auditive sur la base de la source du son, il faut donc s'assurer au préalable de la neutralité émotionnelle des sons utilisés.

A) La catégorisation des sons en psychophysique chez l'homme

Contrairement au domaine visuel, les études en catégorisation auditives sont rares. De plus, la plupart des études de catégorisation sonore sont davantage assimilables à de

l'identification libre de sons dans la mesure où les sujets ne doivent pas classer des sons naturels dans deux (ou plus) catégories prédeterminées par l'expérimentateur. En 2007, Gygi et collègues (Gygi et al., 2007) se sont penchés sur les différentes manières dont on peut classer des sons : soit en fonction de leurs propriétés accoustiques (voir la section sur les caractéristiques physiques des signaux sensoriels), soit selon la catégorie de la source. Dans la dernière expérience de leur étude, ils ont demandé à 16 sujets de grouper les sons qui, selon eux, "allaient ensemble" en 5 à 12 groupes. Bien que les sujets n'aient pas eu d'instruction sur le type de catégorisation qu'ils devaient effectuer (i.e. catégorisation accoustique ou par source), ils ont tous groupé les stimuli en fonction de la catégorie de la source. De plus, les catégories sont assez stables parmi les sujets (l'intégralité des sujets a créé une catégorie "animal" et 85% d'entre eux ont également créé la catégorie "véhicule"), ce qui indique que le processus de catégorisation se base toujours sur les mêmes concepts quelle que soit la modalité sensorielle impliquée. Giordano et collègues (Giordano et al., 2010) ont exploré le temps d'identification de sons en fonction de leur catégorie. Ils avaient sélectionné 71 sons "vivants", c'est à dire produits par des actions telles que la respiration, la marche, des vocalisations, etc. et 69 "non-vivants" produits par des objets mécaniques, des véhicules ou les éléments naturels tels que le vent ou la pluie. Les sujets devaient identifier chaque son par un verbe et un ou deux noms (par exemple : une porte qui claque). En comparant les deux sets de sons, ils ont noté que les sons vivants étaient identifiés plus rapidement et plus précisément que les sons non-vivants (temps médians de 18,5 et 22,5 secondes respectivement). Nous ne pouvons toutefois pas directement comparer ces temps d'identification de sons à des temps de réaction obtenus dans la catégorisation visuelle puisqu'il s'agissait de les décrire verbalement. Il semble tout de même que la catégorie à laquelle appartient un son présuppose de la précision avec laquelle il sera identifié, comme cela avait été montré dans le domaine visuel (Gaffan and Heywood, 1993).

Dans une autre étude (Lebrun et al., 2001) où les participants devaient évaluer la familiarité de sons, des temps de réaction beaucoup plus courts ont été obtenus. Dans cette étude, les sujets entendaient un set de 200 sons, un quart étant des sons environnementaux, un quart des bruits non sens, un quart des mots et pour le reste des pseudo-mots. Ils devaient alors presser un bouton dès que le son leur était familier. Les sons duraient une seconde et les temps de réaction observés étaient de l'ordre de 1,5

seconde à partir de l'onset du stimulus. Ces temps de réactions sont davantage similaires à ceux obtenus en vision, d'autant qu'il ne faut pas oublier que les signaux auditifs ont une composante temporelle intrinsèque dont on ne peut s'abstraire.

B) Corrélats neuraux de la reconnaissance des sons chez le primate et l'homme

Il était tentant de faire un parallèle entre le traitement des sons et le traitement des images sous-tendant leur catégorisation. Les travaux en neuro-accoustique ont donc naturellement cherché l'existence de voies auditives hiérarchiques similaires aux voies visuelles. Rapidement ont alors émergé des modèles de traitement parallèle du son, une voie ventrale traitant l'identité du son et une voie dorsale traitant de sa localisation spatiale.

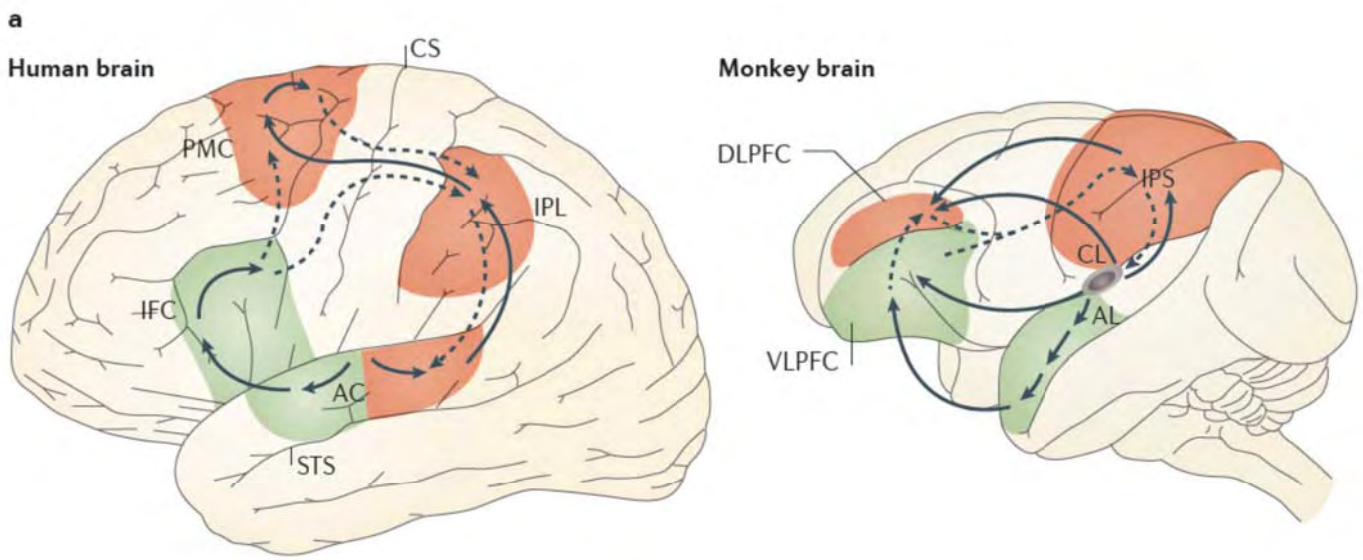


Figure I.4: En orange, voie auditive dorsale (localisation des sons) ; en vert, voie ventrale (identification des sons)(Bizley and Cohen, 2013). AC: Auditory cortex ; AL: région antérolatérale de la ceinture ; CL: région caudolatérale ; DLPFC: cortex préfrontal dorsolatéral, IFC: cortex frontal inférieur ; IPL: lobe intrapariétal ; IPS: sulcus intrapariétal ; PMC: cortex prémoteur ; STS: sulcus temporal supérieur ; VLPFC: cortex préfrontal ventrolatéral. (Bizley and Cohen, 2013)

Ainsi chez l'homme comme chez le macaque, les chercheurs ont pu démontrer une différenciation du traitement de la localisation et de l'identité des sons dans le cortex

(Alain et al., 2001, Tian et al., 2001). Chez l'homme, Alain et collègues ont montré que l'activité dans le cortex auditif et le gyrus préfrontal inférieur était davantage corrélée à l'identité du son tandis que sa localisation était liée à une activité dans les aire temporales inférieures et dans le cortex pariétal. La même année, une étude chez le macaque a mis en évidence une dissociation de traitement similaire. Tian et collègues ont ainsi implanté des électrodes dans la ceinture latérale du cortex auditif de deux singes, dans les régions antérolatérale, médiolatérale et caudolatérale. Ils ont enregistré l'activité des neurones de ces aires lorsqu'étaient jouées 7 vocalisations différentes, à 7 localisations différentes, les singes n'étant engagés que dans une tâche d'écoute passive. Leurs résultats ont mis en évidence que les neurones de la région antérolatérale déchargeaient pour des vocalisations spécifiques, quelle que soit la provenance du son, tandis que les neurones de la région caudolatérale déchargeaient quand le son provenait d'une localisation spécifique, quelle que soit la vocalisation émise. De plus, ces deux régions reçoivent des informations de deux régions différentes de l'aire auditive centrale et projettent dans des régions différentes, confirmant l'existence de deux circuits distincts "Où" et "Quoi", comme en vision.

Concentrons-nous désormais sur l'encodage des catégories dans la voie auditive ventrale. Des études récentes en IRM fonctionnelle chez l'homme tendent à montrer que comme pour les catégories visuelles, la catégorie d'un son serait encodée de manière distribuée dans les aires tardives de la voie auditive ventrale, notamment dans le gyrus temporal supérieur et le sulcus temporal supérieur (Doehrmann et al., 2008, Staeren et al., 2009, Tsunada and Cohen, 2014). Tsunada et collègues ont par ailleurs enregistré des populations de neurones du gyrus temporal supérieur chez le singe lors d'une tâche de catégorisation auditive "dad" versus "bad". Leur stimuli comprenaient les deux prototypes ainsi que des versions morphées de ces deux mots afin de créer un continuum (Tsunada et al., 2011). D'un point de vue comportemental, comme attendu, les singes ont répondu de manière catégorielle avec une frontière nette entre les deux catégories. Mais ce qui est plus surprenant, c'est qu'une large portion des neurones enregistrés répondaient également de manière catégorielle, en tout ou rien.

Si l'on s'intéresse maintenant à des sons naturels qui ne soient pas des mots, des études en EEG chez l'homme ont montré que les sons environnementaux étaient traités de manière similaire aux mots dans la voie ventrale, ce qui suggère un traitement sémantique du son. De plus, des différences importantes de traitement ont été observées pour des sons "non-sens", c'est à dire des bruits ne correspondant à aucun objet ou action (Lebrun et al., 2001).

IV) Les effets de contexte

A) Au sein d'une modalité

En conditions écologiques, nous ne percevons jamais une image ou un son de manière isolée. Notre environnement est complexe, il nous fournit des indices qui peuvent nous aider aussi bien que nous induire en erreur pour la reconnaissance d'un stimulus en particulier. Nous avons des connaissances a priori, acquises au cours de la vie, des co-occurrences les plus fréquentes d'objets dans une scène. Les premières expériences sur l'induction en erreur par la suggestion contextuelle ont été menées dans le domaine du langage. Si l'on demande à des sujets de retenir une liste de mots appartenant tous au même champ lexical, lors de la phase de rappel où l'expérimentateur cite des mots en demandant au sujet s'ils étaient dans la liste, on constate un fort taux de faux souvenirs pour des mots distracteurs qui appartiennent au même champ lexical que les cibles (Deese, 1959).

Le même phénomène a été montré quarante ans plus tard dans le domaine visuel par Miller et Gazzaniga(Miller and Gazzaniga, 1998). Dans leur expérience, les sujets devaient mémoriser une scène de laquelle quelques éléments avaient été retirés (figure I.5 B ou C), puis ils étaient soumis à une phase de rappel ou l'expérimentateur donnait à voix haute une liste d'objets, les sujets ayant pour consigne de répondre si oui ou non l'objet était dans l'image.

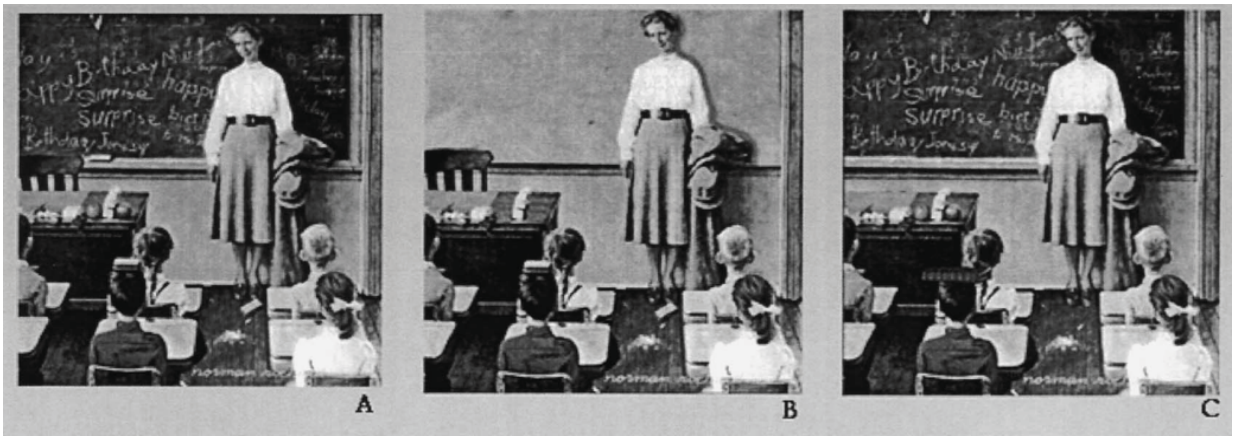


Figure I.5: Stimuli utilisés dans l'expérience de Miller et Gazzaniga. A: image originale, B et C images modifiées desquelles on a retiré quelques éléments

De même que dans le domaine du langage, les sujets étaient persuadés de se souvenir d'avoir vu certains objets, qui auraient été logiquement présents dans le contexte, mais qui n'y étaient pas.

1) L'influence du contexte dans la reconnaissance d'images

Le premier à s'intéresser au phénomène d'influence contextuelle dans la reconnaissance visuelle rapide d'objets a été Palmer (Palmer, 1975). Pour étudier l'influence du contexte, il avait imaginé un protocole de type amorçage. Ainsi il présentait le dessin d'une scène (dessin contouré en noir et blanc) pendant deux secondes, puis flashait très brièvement le dessin d'un objet. Les sujets avaient pour consigne de nommer l'objet. Il a alors montré que les sujets identifiaient l'objet flashé plus rapidement et plus précisément lorsque la scène présentée en amorce était congruente avec l'objet.

Mais que se passe-t-il quand l'objet à catégoriser est incorporé au sein d'une scène? Le traitement de la scène peut-il interférer avec le traitement de l'objet? Biederman et collègues ont étudié différents types de violation de la congruence entre l'objet et la scène: incongruence de taille, de position (incluant le support de l'objet, et l'interposition où l'on peut voir l'arrière plan à travers l'objet) et l'incongruence sémantique (Biederman et al., 1982). Ces chercheurs ont montré que tous les types d'incongruence altéraient les performances de détection d'objet des sujets. L'incongruence sémantique dégradait les performances des sujets autant que les violations d'ordre physique (taille, position) et

étaient tout autant détectables. Ces conclusions mettaient alors à malle les hypothèses sur un traitement sériel de l'information visuel où le physique serait traité avant l'accès au contenu sémantique des scènes et des objets. Ces résultats ont été remis en cause une quinzaine d'années plus tard par Hollingworth et Henderson (Hollingworth and Henderson, 1998). Ils ont répliqué l'expérience de Biederman de 1982, puis ont ajouté des conditions d'indiçage verbal différentes de celles utilisées par Biederman. Ils ont ainsi montré que si l'indice verbal désignant l'objet à détecter était congruent avec la scène présentée alors les participants avaient davanage tendance à dire qu'ils avaient vu l'objet, qu'il soit ou non représenté. L'effet de la congruence sémantique dans ce cas changeait le biais de réponse des participants, mais n'améliorait pas leurs performances (en augmentant le taux de fausses alarmes). Pourtant, par la suite, les chercheurs ont continué à s'intéresser aux effets de congruence sémantique entre l'objet et son contexte, aux interactions possibles notamment en explorant les latences de catégorisation des scènes et des objets. Dans deux études successives, Rousselet et collègues ont exploré le temps nécessaire à la catégorisation d'objet dans un premier temps et de la scène dans un second (Rousselet et al., 2003, Rousselet et al., 2005), et il est apparu que la catégorisation de l'objet était plus rapide que la catégorisation de la scène. Toutefois, la catégorisation de l'objet était réalisée au niveau super-ordonné (Animal versus Non-Animal ou Visage versus Non-visage) alors que la catégorisation de la scène était réalisée au niveau de base (mer, montage, intérieur et ville), or il a été montré par la suite que la catégorisation au niveau super-ordonné était plus rapide que la catégorisation au niveau de base, pour les objets du moins (Mace et al., 2009). Cette différence de temps de traitement n'était donc peut-être due qu'à la différence de niveau de catégorisation demandée. Cette hypothèse fut confirmée en 2007, par Joubert et collègues (JOUBERT et al., 2007) qui se sont intéressés à la catégorisation de scènes au niveau super-ordonné. Ils ont retrouvé des temps de réaction similaires, mais pas plus rapides, à ceux obtenus lors de tâches de catégorisation d'objets. Un deuxième résultat intéressant de cette étude est que si un objet saillant était présent au premier plan de l'image, alors la catégorisation était ralentie, et d'autant plus si l'objet était incongruent avec la scène (comme un arbre au premier plan d'une scène urbaine), ceci suggère que le traitement de la scène et de l'objet sont effectués en parallèle et en interrelation. Un an plus tard, les mêmes auteurs ont donc recherché quelle était l'influence du contexte sur

la catégorisation de l'objet au premier plan, au niveau super-ordonné "Animal" versus "Non-Animal" (JOUBERT et al., 2008). Ils ont manipulé des images en collant une vignette d'animal ou d'objet sur une scène naturelle ou artificielle, créant ainsi des paires congruentes (i.e. animal et scène naturelle/objet et scène artificielle) ou incongruentes (l'association inverse). En condition contrôle, ils ont collé les mêmes vignettes sur fond gris. Ils ont alors observé une altération des performances, tant en termes de temps de réaction que de réussite à la tâche pour des paires incongruentes en comparaison aux paires congruentes, mais n'ont obtenu ni altération ni amélioration des performances pour les vignettes collées sur fond gris, relativement aux paires congruentes. Au niveau du traitement cérébral, Rémy et collègues ont récemment exploré les zones du cerveau impliquées dans ces effets de déterioration des performances dus à l'incongruence entre l'objet et son contexte (Remy et al., 2014). Elles ont pour cela répliqué la tâche de catégorisation visuelle rapide Animal versus Non-Animal, et utilisé les mêmes stimuli que Joubert et collègues (2008). Les sujets ont réalisé la tâche dans le scanner IRM. Ces chercheuses ont ainsi mis en évidence une augmentation du signal en IRM pour les stimuli incongruents (i.e. Animal dans scène artificielle ou Objet dans scène naturelle) au niveau du cortex parahippocampique antérieur droit, et ce quelle que soit la catégorie de l'objet au premier plan. Dans le cas où l'objet à catégoriser est bien visible, le contexte n'aide donc pas à la catégorisation, par contre il peut perturber les sujets s'il est incongruent avec l'objet. Dans des conditions de visibilité réduite par contre, le contexte pourrait fournir des indices importants pour la reconnaissance de l'objet. Bar l'avait ainsi illustré dans une revue (BAR, 2004):

Figure I.6: Influence du contexte sur l'objet en noir, identique sur les deux images.

Les deux objets noirs sont identiques sur les deux images, mais ils sont flous et non identifiables individuellement. Pourtant, avec le contexte on identifie un sèche-cheveux à

gauche et une perceuse à droite. Toutefois, ces résultats ne peuvent s'appliquer qu'au sujet sain. En effet, chez des patients atteints de DMLA, il a été montré que la présentation d'un objet dans un contexte congruent, non seulement ne facilitait pas sa reconnaissance mais au contraire altérait les performances des sujets. Cet effet pouvait s'expliquer par l'augmentation d'informations visuelles à traiter chez des sujets ayant déjà une acuité visuelle très diminuée et donc des difficultés à reconnaître les objets (Boucart et al., 2008).

Lors de l'élaboration d'une tâche de catégorisation, il ne faut donc pas négliger les effets éventuels que peut avoir le contexte sur la reconnaissance de l'objet.

2) L'influence du contexte dans la reconnaissance des sons

Comme dans le domaine visuel, le contexte peut faciliter l'identification d'un son ou induire une identification erronée. La facilitation via le contexte auditif est bien connue en linguistique: nous comblons automatiquement les syllabes manquantes dans les mots que nous entendons, grâce au reste de la phrase (Warren, 1970). Mais il existe aussi pour l'identification des sons environnementaux ou naturels. En 1991, Ballas et Mullins ont mené une expérience sur la congruence contextuelle auditive. Ils ont sélectionné des paires de sons partageant des caractéristiques accoustiques similaires (un bruit de friture et le bruit d'une mèche d'explosif qu'on allume), mais étant parfaitement discriminables isolément (test d'identification des sons isolés par les sujets préalablement à l'expérience). Puis les sons étaient présentés dans des séquences contenant des sons créant un contexte auditif clairement reconnaissable (comme les bruits d'une cuisine d'un restaurant ou ceux d'une rue), et au milieu était intégré le son cible, entouré de deux bips. La tâche du sujet était alors d'identifier ce son cible. Le contexte ne facilitait pas l'identification des sons en condition congruente relativement à la condition isolée (bruit de friture dans une ambiance sonore de cuisine), mais quand les sons étaient présentés dans une séquence incongruente (bruit de mèche dans une ambiance de cuisine), la réponse du sujet se trouvait très fortement biaisée vers l'autre son de la paire, le contexte l'ayant alors induit en erreur.

D'autre part, certains sons partagent de très fortes similitudes acoustiques, au point qu'il peut être difficile de les discriminer, bien que la source qui les ait produits soit complètement différente. Dans ce cas, lorsqu'on écoute un tel son isolé, on a autant de chance de l'identifier comme provenant de la source A que de la source B. Le contexte peut en revanche biaiser la réponse vers l'une ou l'autre des sources. C'est ce qu'ont voulu tester Niessen et collègues dans leur expérience (Niessen et al., 2008). Ils ont sélectionné des paires de sons similaires acoustiquement et ont créé des paires homonymes chimériques en associant l'enveloppe temporelle de l'un avec la structure temporelle fine de l'autre, ils ont ensuite sélectionné le son qui apparaissait comme le plus naturel. Bien que cette construction soit biaisée (cf paragraphe sur les caractéristiques physiques des signaux sonores), ils ont tout de même pu observer un effet de contexte: dans un contexte donné, les sujets avaient tendance à identifier le son comme provenant de la source congruente avec le contexte, alors que de manière isolée, ils donnaient aussi bien la source de l'enveloppe du son que celle de la structure fine.

Enfin Gygi et Shafiro ont montré que des sons pouvaient bénéficier d'une incongruence contextuelle (Gygi and Shafiro, 2011). Dans leur expérience ils ont sélectionné 14 scènes sonores et 31 sons isolés. Ils ont ensuite créé 56 appariements, la moitié étant congruents et l'autre non. Les sujets devaient identifier les sons parmi une liste qui leur était fournie au préalable. Dans cette étude, Gygi et Shafiro ont alors montré que les sons incongruents étaient identifiés plus précisément que les sons congruents. Il faut noter tout de même que, contrairement aux études précédemment citées, les sons utilisés comme cibles dans ces derniers travaux n'étaient pas ambigüs, ce qui peut expliquer qu'ils aient trouvé un avantage en situation d'incongruence.

B) Multisensoriel : quand une modalité devient le contexte de l'autre

Dans les phénomènes d'intégration multisensorielle on peut considérer qu'une modalité est contextuelle pour l'autre, en réciprocité. Un effet bien connu de l'influence d'une modalité sur l'autre, est l'effet McGurk (McGurk and MacDonald, 1976): en présentant à des sujets un film où une femme répétait la syllabe "ga" et une bande son où elle répétait

la syllabe “ba”, 98% des sujets ont rapporté avoir perçu “da”, dans la condition inverse, les sujets ont répondu avoir perçu “ba” dans 30% des cas et une combinaison “ba-ga” ou “ga-ba” dans 50% des cas. Il y a donc interférence entre ce qui est perçu visuellement et auditivement, notre cerveau tente de combiner les deux percepts pour nous en donner une interprétation cohérente, mais l’effet n’est pas parfaitement symétrique.

En 2009, Özcan et Van Egmond (Özcan and Egmond, 2009) se sont penchés sur l’influence du contexte visuel sur l’identification de sons. Dans une phase préliminaire, ils ont demandé à 56 sujets d’identifier des images et des sons, isolément, pour évaluer si les stimuli étaient reconnaissables. Puis ils ont demandé à d’autres sujets d’identifier des sons, en les précédant par la présentation d’une scène ou d’un objet donnant un indice sur le son à reconnaître (ex: scène de salle de bain, ou image d’un tube de dentifrice, son d’eau qui coule d’un robinet). Le taux d’identification correcte des sons s’en est trouvé fortement augmenté, alors que les temps de réaction étaient, eux, amoindris. Cette étude n’explorait que l’avantage éventuel du contexte visuel sur l’identification des sons et ne s’intéressait pas aux potentielles perturbations en cas d’incongruence entre les deux modalités. Or, selon Laurienti et collègues (Laurienti et al., 2004) la congruence sémantique des stimuli serait cruciale pour obtenir une intégration multisensorielle efficace. Leur paradigme expérimental était très simple: ils présentaient des cercles colorés au centre d’un écran, rouge, vert ou bleu et le mot rouge, vert ou bleu était prononcé dans des hauts-parleurs. Les sujets devaient presser un bouton lorsqu’ils voyaient la couleur bleue ou entendaient le mot “bleu”, ils devaient en presser un autre pour de la couleur rouge, et ignorer le vert. Les essais pouvaient être unimodaux, bimodaux congruents (la couleur présentée correspondant au mot) ou bimodaux incongruents. Dans les essais bimodaux congruents les sujets obtenaient de meilleurs temps de réaction que dans les conditions unimodales, et étaient même plus rapides que ce que prévoit le race-model (cf étude 3), suggérant qu’il y avait une intégration multisensorielle. En condition bimodale incongruente les temps de réaction des sujets étaient par contre ralentis relativement aux conditions unimodales, suggérant une interférence d’une modalité sur l’autre.

Doehrmann et Naumer ont passé en revue des études portant sur l’importance de la congruence sémantique dans l’intégration audio-visuelle(Doehrmann and Naumer, 2008). Ils reportent ainsi un effet assymétrique de la congruence sémantique dans les

expériences en audio-visuel: si on doit identifier un son avec une image incongruente, les performances sont davantage affectées que si l'on doit identifier une image alors qu'un son incongruent est joué. Cela s'expliquerait par la dominance visuelle de l'Homme. En 2008, Schneider et collègues ont également mis en évidence la dominance visuelle chez l'homme. D'une part les sujets sont plus précis et plus rapides lorsqu'il s'agit de reconnaître des images plutôt que des sons. D'autre part, les sujets devaient réaliser une tâche d'évaluation de taille ("l'objet que vous voyez/entendez rentre-t-il dans une boîte à chaussures?") avec une amorce auditive ou visuelle avant la cible. Les temps de réaction se trouvaient alors rallongés lorsque l'amorce et la cible n'avaient pas le même ordre de grandeur de gabarit, et d'autant plus que l'amorce était dans la modalité visuelle (Schneider et al., 2008).

V) Les traits diagnostiques des catégories : haut niveau versus bas niveau ?

Dans les stimuli utilisés lors des tâches de catégorisation, les paramètres physiques sont étroitement liés à la signification du stimulus, rendant difficile l'identification des paramètres indispensables ou ayant une influence lors de ces tâches. Toutefois des chercheurs ont tenté de dissocier les caractéristiques de haut et bas niveau des stimuli pour évaluer leur contribution relative à leur identification.

A) Définition haut niveau et bas niveau en audition et vision

Il est délicat de trouver une définition consensuelle du bas niveau et du haut niveau des stimuli, qu'ils soient auditifs ou visuels. Ceci provient notamment du fait que nous ne comprenons que très peu d'éléments concernant le fonctionnement des aires cérébrales de haut niveau. Dans cette thèse nous parlerons donc de bas niveau pour les propriétés physiques des stimuli traitées par la rétine jusqu'à l'aire V1 dans le domaine visuel ou par la cochlée dans le domaine auditif.

Les caractéristiques de haut niveau comprennent donc le contenu sémantique des stimuli mais également des propriétés intermédiaires sur lesquelles nous avons peu de contrôle.
Dans les études que nous présenterons dans les trois chapitres de ce manuscrit, nous avons tenté d'égaliser les paramètres de bas niveau des stimuli, pour avoir une meilleure compréhension des phénomènes de la catégorisation. Nous sommes bien conscients que, dans l'état actuel des connaissances en neurosciences, l'égalisation que nous avons appliquée n'est pas optimale, mais elle offre l'avantage d'éliminer déjà certains biais.
Avant de s'aventurer plus loin, nous allons faire une brève revue des travaux déjà réalisés pour tenter de discriminer le bas niveau du haut niveau des stimuli.

B) Contribution des caractéristiques de bas niveau en catégorisation visuelle

1) Les caractéristiques de haut et bas niveau dans le domaine visuel

Bien que nous ne connaissions pas encore l'étendue des traitements effectués par les neurones de la rétine à ceux de l'aire V1, des travaux réalisés ces cinquante dernières années ont mis en évidence un certain nombre d'encodages de caractéristiques des stimuli visuels.
De manière très simplifiée, nous pourrions dire que les cellules photoréceptrices de la rétine (cônes pour la vision diurne et colorée, et les batônnets pour le constraste et la luminosité, en vision crépusculaire) transforment le signal lumineux en signal électrique, puis le transmettent aux cellules bipolaires et enfin aux cellules ganglionnaires.

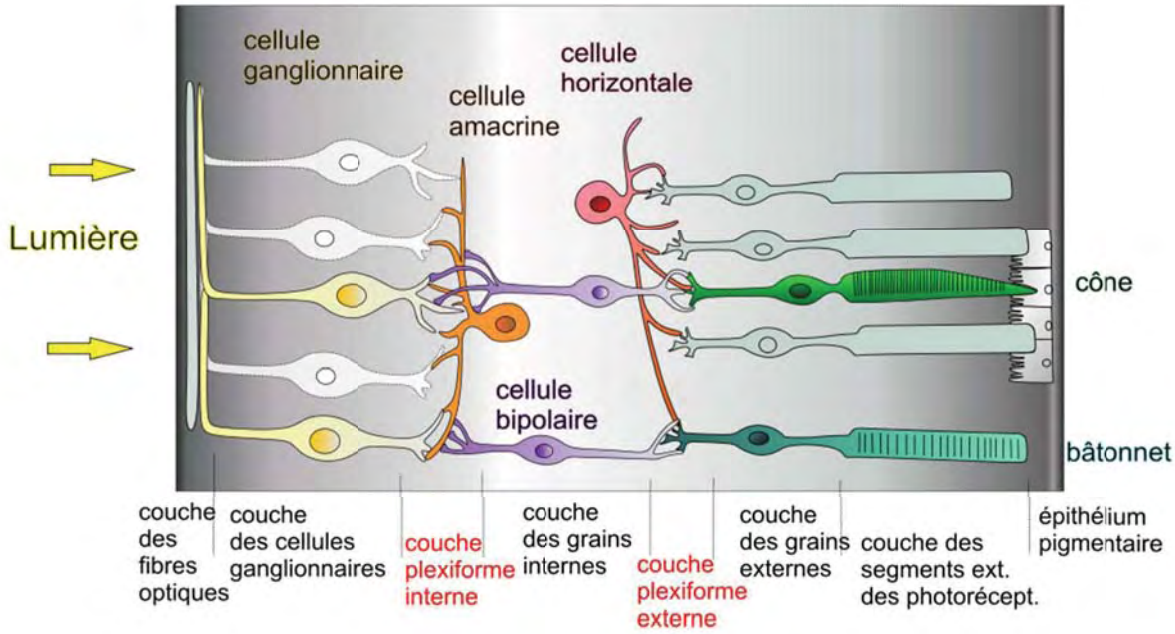


Figure I.7 : Schéma de la structure rétinienne (dessin modifié, tiré de Neuroscience, 2nde édition, (2001))

La rétine encode donc les informations de couleurs (que nous ne traiterons pas dans cette thèse) et de luminance. Toutefois, plus que la luminance absolue des images, la rétine encode les changements locaux de luminance, et donc le contraste (Kuffler, 1953).

Les cellules ganglionnaires de la rétine projettent ensuite dans le noyau géniculé latéral (LGN) qui relaie l'information au cortex visuel primaire.

Les neurones de V1 présentent quant à eux une sélectivité d'orientation et de direction, et encodent les fréquences spatiales ainsi que leur orientation (Hubel and Wiesel, 1968).

2) Le contrôle des caractéristiques de bas niveau en psychophysique

Nombre d'études ont exploré la contribution de deux paramètres communément dissociés à la catégorisation des scènes naturelles: le spectre d'amplitude et la phase. Le spectre d'amplitude d'une transformée de Fourier correspond aux variations de luminance aux fréquences spatiales retrouvées dans l'image ainsi que leur orientation, tandis que le spectre de phase porte la localisation de ces fréquences spatiales, c'est donc la phase qui supporte les contours des formes.

Si on égalise l'amplitude, les performances en terme de taux de réussite ne varient que très peu mais les temps de réaction sont augmentés. Si par contre on randomise la phase des fréquences spatiales, on détruit les contours des formes présentes dans la scène, et au delà d'un certain taux de randomisation, les images ne sont plus reconnaissables et les performances chutent drastiquement (Joubert et al., 2009). Pourtant, si on se fie à des expériences en priming, il semblerait que c'est l'information d'amplitude qui prédispose le mieux à reconnaître une scène: avec un prime constitué du même spectre d'amplitude que l'image cible, mais avec une phase randomisée, on accélère la reconnaissance de la cible. Si par contre le prime est constitué de l'information de phase intacte mais d'une amplitude plate, il n'y a pas d'accélération de la reconnaissance de la cible (Guyader et al., 2004). En contradiction avec ces résultats, Loschky et collègues ont montré que si les sujets sont engagés dans une tâche de catégorisation de scènes naturelles, l'information d'amplitude non localisée, c'est à dire après randomisation de la phase, ne permet pas de discriminer au delà du niveau de chance. (Loschky et al., 2007).

Enfin, peu d'études existent chez le singe sur la contribution des caractéristiques de bas niveau dans la catégorisation d'images. Toutefois nous pouvons citer une expérience de Girard et Koenig (Girard and Koenig-Robert, 2011) en saccades oculaires. Une guenon macaque et neuf sujets humains ont été soumis à une tâche de catégorisation visuelle ultra rapide "Animal" versus "Non Animal"en choix forcé, où les stimuli étaient présentés par paire (une cible, image contenant un animal, et un distracteur). Tous les stimuli étaient des photographies en noir et blanc égalisées en spectre d'amplitude. Les résultats ont montré que la guenon et les humains avaient des performances similaires, légèrement altérées par rapport à celles obtenues pour des images intactes mais bien au delà du niveau de chance. Les auteurs ont ensuite cherché à lier les performances obtenues pour chaque item aux basses fréquences spatiales contenues dans les images. Pour ce faire, ils ont appliqué un filtre passe-bas sur les images, et ont soumis cinq nouveaux sujets à une tâche de catégorisation, sans contrainte temporelle. Les sujets devaient de surcroît évaluer leur certitude quant à la présence d'un animal dans l'image. Cette transformation affectait beaucoup les contours des formes, et les auteurs ont montré que les stimuli les plus reconnaissables dans leur version filtrée étaient ceux qui conduisaient au plus fort taux de réussite dans la tâche de catégorisation rapide. D'après

ces résultats, le contenu en basses fréquences spatiales serait indispensable pour la catégorisation.

Les études sur la contribution des traits de bas niveau des images (i.e. luminance, contraste et fréquences spatiales) n'ont pour l'instant pas conduit à un consensus, notamment à cause de la diversité des paradigmes utilisés. De plus, différents mécanismes semblent à l'oeuvre et il reste très difficile de les dissocier.

C) Caractéristiques acoustiques et catégorisation auditive

1) Les caractéristiques du son traitées par la cochlée

L'homme perçoit les sons situés dans une gamme de fréquences allant de 20 à 20000Hz et dans une gamme d'intensité couvrant 120 décibels. Ces facultés reposent majoritairement sur les propriétés mécaniques et biophysiques de la cochlée, l'organe périphérique dédié à l'audition chez les mammifères. La cochlée consiste en trois tubes membraneux adjacents enroulés en spirale : la rampe vestibulaire, la rampe tympanique et le canal cochléaire situé entre les deux. Dans le canal cochléaire, se trouve l'organe de Corti. Il repose sur la membrane basilaire, séparant la rampe tympanique du canal cochléaire, et est recouvert par la membrane tectoriale. L'organe de Corti possède deux types de mécanorécepteurs, les cellules ciliées internes qui sont les récepteurs sensoriels et transforment le signal sonore en signal électrique, et les cellules ciliées externes qui peuvent amplifier une vibration locale. Lorsqu'une vibration sonore entre dans le canal auditif, elle fait vibrer le tympan qui transmet mécaniquement la vibration via les osselets à l'entrée de la rampe tympanique. Il y a donc une différence de pression entre la rampe tympanique et les deux autres canaux, ce qui provoque un déplacement transversal de la membrane basilaire. Enfin ces vibrations sont transmises aux cellules ciliées externes puis internes de l'organe de Corti, ces dernières effectuant l'opération de transduction du signal avant d'envoyer le signal nerveux au nerf auditif.

La cochlée est organisée de manière tonotopique de la base vers l'apex: les fréquences les plus hautes sont traitées à la base de la cochlée alors que les basses fréquences sont

traitées à son apex. Cette organisation tonotopique se retrouve également au niveau de l'aire auditive A1.

2) De la cochlée à l'aire A1

Le traitement du son commence par la décomposition en fréquences réalisée par la cochlée. Les neurones de la cochlée projettent ensuite leur axone vers le noyau cochléaire qui à son tour projette des connexions aux complexes olivaires supérieurs des deux côtés et au colliculus inférieur controlatéral (cf figure I.8). Enfin les signaux sont envoyés au cortex auditif via le corps géniculé médian.

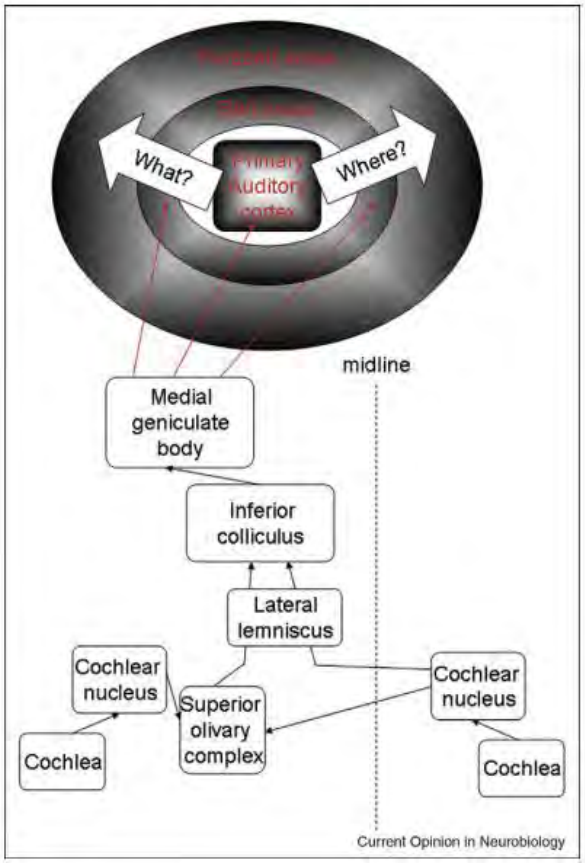


Figure I.8 : Voie auditive ascendante (Nelken, 2008)

L'aire A1, bien que portant le nom de cortex auditif primaire, ne peut être comparée à l'aire visuelle V1. En effet les neurones de l'aire A1 répondent à des objets auditifs plutôt qu'à de simples combinaisons de caractéristiques accoustiques. Ainsi Bar-Yosef et Nelken ont montré que la réponse de neurones de A1 n'était que peu affectée par la soustraction de différentes composantes du son (Las et al., 2005, Bar-Yosef and Nelken, 2007). De plus en 2005, il avait été montré que de faibles composantes acoustiques présentées dans du

bruit pouvaient être surreprésentées dans A1 dans la mesure où ils étaient pertinents pour la tâche, et activer les neurones de cette aire au même niveau que si elles avaient été présentées seules (Las et al., 2005). L'aire A1 effectue donc des traitements de l'information auditive allant bien au-delà de la simple analyse des caractéristiques purement physiques du son. C'est pourquoi nous avons choisi de considérer comme « bas niveau » les paramètres acoustiques traités par la cochlée.

3) Les études en psycho-acoustique

Bien que l'homme soit capable de juger de la similarité acoustique des sons, lorsqu'il est engagé dans une tâche de catégorisation auditive, il se base davantage sur la catégorie de la source du son (Gygi et al., 2007).

Une étude récente a exploré l'importance du timbre pour la catégorisation de sons très brefs (Suied et al., 2014). Le timbre est ce qui donne une identité au son : deux instruments peuvent jouer la même note à la même hauteur, et pourtant les deux sons produits seront discriminables, les différences de timbre résultent donc principalement de la combinaison des différentes harmoniques d'un son avec leurs intensités propres, et notamment des phases d'attaque. Pour étudier l'importance du timbre, ils ont utilisé trois catégories de sons : les voix, les instruments à cordes et les percussions. Les stimuli avaient été égalisés en termes de hauteur du son, et sous échantillonnés avec un filtre sinusoïdal. Les participants ne pouvaient se baser que sur le timbre pour catégoriser le son, c'est-à-dire les composantes fréquentielles et leurs amplitudes. Les voix peuvent ainsi être correctement discriminées à partir d'extraits de 8ms, pour les instruments de musique, des durées de 16ms seraient nécessaires pour pouvoir les discriminer avec fiabilité.

Dans leur étude en 2008 Niessen et collègues avaient formé des sons chimériques à partir de l'enveloppe temporelle d'un son, et la structure fine d'un autre (Niessen et al., 2008). Ce faisant, ils pensaient avoir créé des sons homonymes. En réalité ils se sont aperçus a posteriori que l'enveloppe temporelle semblait être un prédicteur plus fort de l'identité du son que sa structure fine. En effet l'enveloppe du son porte une information de timbre, puisqu'elle supporte les phases d'attaque. Leur objectif initial était de tester l'influence du contexte auditif sur l'identification d'objets sonores, mais ils ont dû nuancer

leurs résultats : il était bien plus facile de biaiser l'identification du son vers celui dont ils avaient utilisé l'enveloppe temporelle. Cette étude a donc permis de mettre en évidence les caractéristiques acoustiques sur lesquelles on se base lors de la reconnaissance d'un objet sonore.

D) Le contrôle des paramètres de bas niveau dans l'exploration du rôle du contexte

Comme nous l'avons noté dans le paragraphe IV, les effets de contexte peuvent s'observer au sein d'une modalité (auditive ou visuelle) mais également une modalité peut avoir un rôle contextuel pour l'autre. Toutefois, les études tentant de distinguer les effets de facilitation dus aux caractéristiques de haut niveau ou de bas niveau des stimuli en psychophysique sont rarissimes voire inexistantes (nous ne traitons pas ici des études de classification d'images par ordinateur), que l'on se situe en uni- ou multisensoriel. En effet dans les études dans le domaine visuel ((JOUBERT et al., 2008, Fize et al., 2011) comme auditif (Niessen et al., 2008, Gygi and Shafiro, 2011), les auteurs se sont surtout concentrés sur les caractéristiques de l'objet à catégoriser, et au mieux ont égalisé certains paramètres du contexte : luminance et contraste globaux en visuel, amplitude et durée dans le domaine auditif. En multisensoriel, nous nous devons toutefois de citer l'étude de Werner et Noppeney (Werner and Noppeney, 2010). Dans leur étude, ils ont comparé les performances de catégorisation d'objets (outils versus instrument de musique) dans différentes conditions : film seul, son seul, film + bande son (toujours congruents), film+bruit blanc, son+stimulation visuelle sans contenu. Bien que les stimuli ayant un contenu sémantique et ceux n'en ayant pas ne soient pas équivalents d'un point de vue des caractéristiques physiques, ils ont montré que le gain multisensoriel n'était pas dû seulement à la stimulation conjointe de deux modalités, mais qu'il était nécessaire que le stimulus dans une modalité contienne une information pertinente afin de favoriser la catégorisation du stimulus dans l'autre modalité.

L'importance relative des caractéristiques de bas niveau des stimuli dans l'effet de facilitation contextuelle reste donc un domaine inexploré en psychophysique, et nous proposons dans cette thèse deux expériences apportant des éléments de réflexion sur cette question.

VI) Problématiques de la thèse

Dans cette thèse nous nous proposons donc d'étudier les contributions relatives des caractéristiques de haut et bas niveau des stimuli dans la catégorisation chez l'homme et le singe.

La catégorisation est un processus cognitif d'assez haut niveau, mais dans les stimuli les informations de bas niveau sont toujours présentes, comment alors s'en abstraire, au moins partiellement, pour caractériser les corrélats neuraux spécifiques de la catégorisation visuelle ? C'est à cette première question que nous répondrons dans la première partie de cette thèse, en utilisant une technique de stimulation visuelle (SWIFT) mise au point par mon prédécesseur. Cette technique a prouvé son efficacité chez l'homme, et nous l'avons appliquée au singe macaque. SWIFT détruit les contours des objets de manière cyclique mais conserve les contrastes, les fréquences spatiales et la luminance, les images sont alors détruites et reconstruites cycliquement dans le temps. Dans cette première étude, nous avons enregistré, via l'électrocorticographie, l'activité des aires visuelles évoquée par les séquences SWIFT lors d'une tâche de catégorisation chez le singe. Nous avons alors pu comparer l'activité évoquée par l'image détruite à celle évoquée par l'image d'origine, à une précision spatio-temporelle élevée.

L'homme et le singe peuvent catégoriser efficacement des images, et des études ont montré que le contexte entourant la cible de la catégorisation pouvait influencer les performances. Cette interférence a lieu précocement (l'effet est d'autant plus marqué que la réponse du sujet a été rapide) et pourrait donc être dû aux traits bas niveau du contexte dont l'analyse concomitante viendrait perturber celle de l'objet. De plus les études portant sur l'importance diagnostique de différentes caractéristiques de bas niveau des images n'ont pas mené à un consensus clair. Nous avons donc décidé d'aborder la question de l'importance du spectre d'amplitude des scènes naturelles dans la catégorisation d'objets chez le macaque et l'homme en psychophysique.

Enfin, dans le troisième chapitre de cette thèse, nous nous sommes intéressés à l'influence d'une modalité sur l'autre en construisant un protocole de catégorisation audiovisuelle. L'importance de la congruence sémantique entre les différentes entrées

sensorielles a été clairement démontrée par le passé, mais jamais dans des études portant sur la catégorisation « forcée » en multimodal (c'est-à-dire où les catégories sont prédéfinies par l'expérimentateur, en opposition à la catégorisation libre où les sujets forment leurs propres catégories), qui sont en outre extrêmement rares. Encore plus rares sont celles qui ont tenté de contrôler les caractéristiques physiques des stimuli. Dans ce dernier chapitre nous proposons d'étudier et de quantifier l'effet de congruence sémantique multisensorielle lors d'une tâche de catégorisation chez l'homme. Pour s'abstraire d'effets purement dus à la présence de deux entrées sensorielles, nous avons choisi de toujours présenter simultanément une séquence visuelle et une séquence sonore. Dans le domaine visuel, afin d'égaliser au mieux les caractéristiques physiques des images, nous réutilisons la technique SWIFT. Dans le domaine auditif, nous avons pris soin de conserver au maximum le contenu fréquentiel des sons, en créant un bruit « non-sens » à partir des sons utilisés dans l'expérience. Enfin, afin de pouvoir au mieux comparer les conditions audiovisuelles congruentes et incongruentes, nous avons toujours présenté des séquences des deux catégories superposées dans les deux modalités sensorielles. Ce protocole quelque peu complexe avait pour but d'isoler de la manière la plus fiable l'effet de la congruence sémantique sur le gain multisensoriel.

Chapitre 1 : Les corrélats neuraux de la reconnaissance d'objets dans une tâche de catégorisation.

Grâce à des travaux antérieurs, la faculté de catégorisation visuelle a été démontrée chez de nombreuses espèces animales (Herrnstein and Loveland, 1964, Fabre-Thorpe et al., 1998) dont le primate. Mais que se passe-t-il dans les aires visuelles lorsqu'il catégorise des images ? Nous avons voulu dans cette étude caractériser la signature neurale de la reconnaissance d'images en tant que cibles de la tâche de catégorisation. Toutefois, chaque fois que notre système visuel reçoit une stimulation, il va traiter les différents paramètres physiques de l'image (contraste, luminance, fréquences spatiales…) pour fournir aux aires supérieures une représentation cohérente du stimulus perçus. La difficulté d'étudier les corrélats neuraux de la catégorisation visuelle réside donc dans la possibilité de s'affranchir de l'activité liée aux traitements de ces informations physiques pour se concentrer sur l'activité liée au traitement du contenu sémantique. Pour résoudre ce problème nous avons utilisé une technique de stimulation visuelle mise au point par mon prédécesseur (Koenig-Robert and VanRullen, 2013) et expérimentée chez l'homme dans une tâche de reconnaissance d'objets associée à des enregistrements EEG. Cette technique module les contours des images de manière cyclique au cours du temps, détruisant ainsi périodiquement leur forme et donc leur contenu sémantique, tout en conservant stables les caractéristiques de bas niveau. En enregistrant l'activité corticale du singe (électrocorticographie), nous avons pu mettre en évidence des signaux électrophysiologiques directement liés à la reconnaissance de l'image en tant que cible de la tâche. Grâce à cette étude nous avons pu mettre en évidence un parallèle fort entre l'homme et le singe concernant les corrélats neuraux de la reconnaissance d'objets.

I) Les corrélats neuraux de la reconnaissance d'objets chez l'homme

En 2013 Roger Koenig-Robert et Rufin VanRullen ont mis au point une nouvelle technique de stimulation visuelle pour tenter d'isoler les corrélats neuraux de la reconnaissance d'objets. Leur constat de départ était que de nombreuses études, principalement chez le singe, ont proposé des modèles de l'émergence des représentations d'objets dans le système visuel le long de la voie ventrale, mais aucune étude n'a pu explorer la valeur sémantique de ces représentations. Cette impossibilité relevait surtout des techniques de stimulation visuelles utilisées qui ne permettaient pas de distinguer les traitements neuraux des caractéristiques physiques des stimuli de ceux des attributs sémantiques.

Pour répondre à cette problématique, ils ont créé des séquences d'images (présentées ensuite comme un film) où une image d'origine était *scramblée* via des transformées en ondelettes puis reconstruite de manière cyclique. La décomposition en ondelettes offre l'avantage, contrairement à la transformée de Fourier, d'être localisée dans l'espace. Elle leur a donc permis de détruire les contours tout en conservant les attributs locaux de bas niveau tels que la luminance, le contraste et les fréquences spatiales. Ils ont nommé cette technique : SWIFT (Semantic Wavelet Induced Frequency Tagging).

Le but de cette technique était donc de pouvoir isoler les corrélats neuraux de la reconnaissance sémantique de ceux du traitement bas niveau de l'image, de plus en choisissant des images non canoniques des objets représentés, ils s'assuraient de favoriser la non-reconnaissance des objets, et étaient alors en mesure de comparer le signal obtenu pour des stimuli reconnus et non reconnus.

A) La création des séquences SWIFT

L'algorithme de Koenig et VanRullen utilise la méthode des ondelettes pour extraire les composantes de l'image dans 3 directions du plan (horizontale, verticale, diagonale) à différentes fréquences spatiales (Koenig & VanRullen (2012). Cette technique permet de moduler les contours locaux de manière cyclique tout en conservant en permanence les caractéristiques de bas niveau des images, contrairement à la SSVEP classique présentée en figure 1.1 (pour plus de détails sur la technique de création des séquences SWIFT, se référer au paragraphe II-B-8 de ce chapitre).

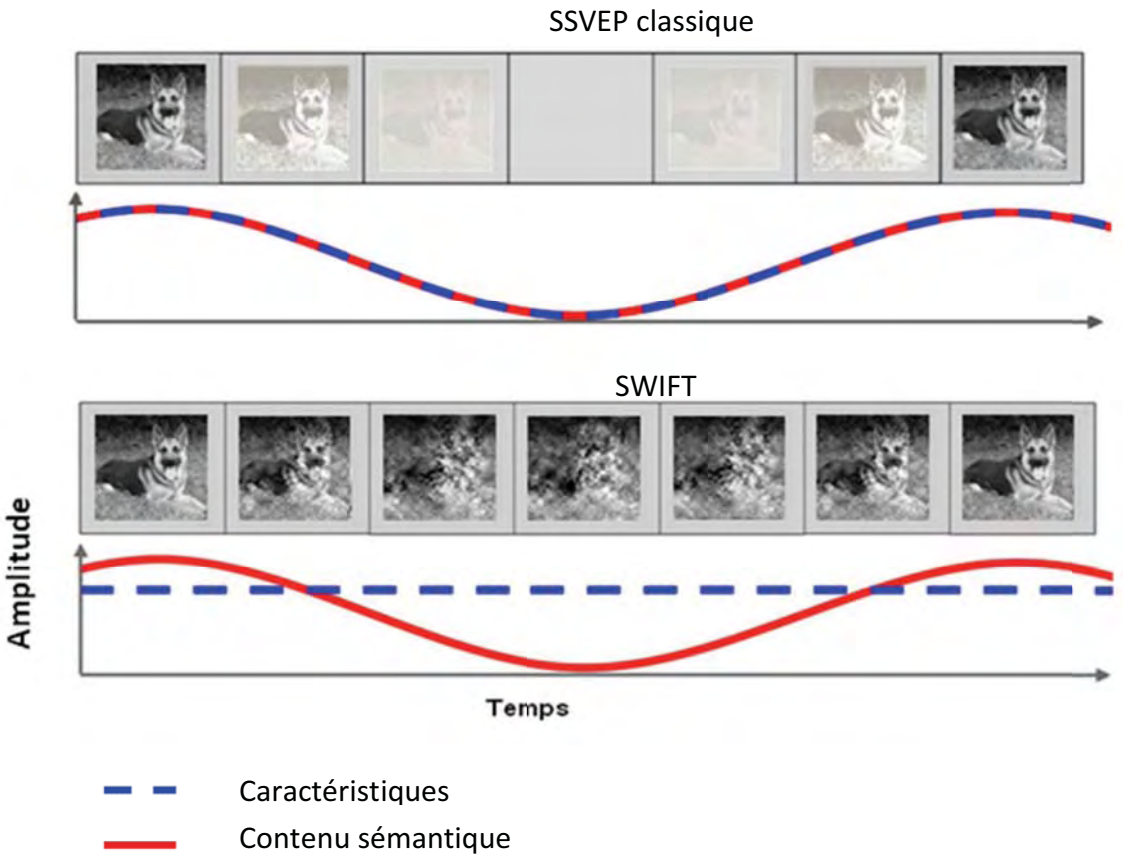


Figure 1.1 : Comparaison des séquences SSVEP et SWIFT

B) Protocole

Dans cet article, Koenig et VanRullen ont mis au point trois expériences différentes, mais nous ne discuterons ici que la première.

Cette expérience consistait en une tâche de reconnaissance d'objets utilisant des séquences SWIFT, 17 sujets y ont pris part. Ils ont présenté aux sujets 100 séquences différentes, séparées en 4 blocs de 25 essais. Les sujets devaient presser un bouton dés qu'ils reconnaissaient un objet réel dans l'image, puis un autre s'ils étaient capables de le nommer. Les images avaient été choisies de manière à être difficilement identifiables.

Les essais se déroulaient comme suit : tout d'abord une séquence SWIFT de 30 secondes (*naive period*, figure 1.2), contenant 44 cycles, donc 44 apparitions de l'image non bruitée (i.e. onsets sémantiques), séquence durant laquelle les sujets devaient donner leur réponse, puis l'image d'origine apparaissait pendant 2 secondes (*steady image*, figure 1.2), et une séquence SWIFT identique mais de 10 secondes seulement était présentée à

nouveau (*cognizant period*, figure 1.2), les sujets devant alors simplement fixer l'écran. Les stimuli étaient des photographies en noir et blanc d'objets réels dans 80% des cas, et des textures créées via l'algorithme de Simoncelli dans 20% des cas (Portilla and Simoncelli, 2000b). Ces textures ne contenaient aucune information sémantique et servaient ainsi de contrôle pour les images réelles.

Durant toute l'expérience le signal EEG était enregistré par le système Biosemi à 64 canaux, échantillonné à 1024Hz. Ils ont ensuite analysé les données en termes de potentiels évoqués via EEGlab. Le signal a été sous-échantillonné à 128Hz et filtré par bande passante entre 3.5 et 30Hz. Ils ont analysé les tracés de quatre électrodes centro-pariétales (Cz, CP1, CPz et CP2).

C) Résultats

Pour isoler le signal dû à la reconnaissance sémantique des objets, Koenig et VanRullen ont comparé les tracés obtenus pour un seul cycle swift (moyenné) dans 3 conditions distinctes : quand le contenu sémantique était présent et identifié (condition QR sur figure 1.2), quand le contenu était présent mais non reconnu (condition TR) et quand le contenu sémantique était absent, c'est-à-dire quand le stimulus était une texture (condition NO).

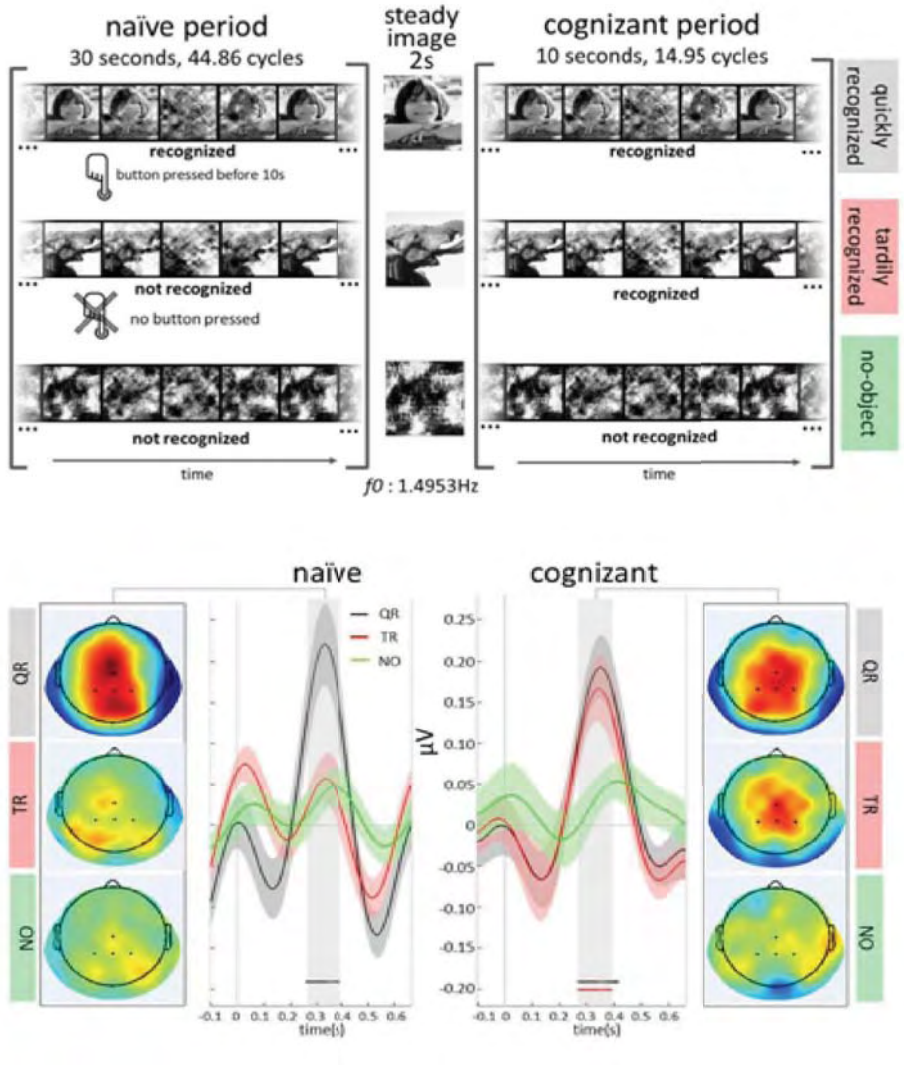


Figure 1.2 : En haut : 3 exemples d'essais représentant les 3 conditions objet réel identifié, objet réel non identifié et absence d'objet. En bas, les potentiels évoqués (pour un cycle, moyenné)dans ces 3 conditions, au cours des 2 périodes de présentation des séquences SWIFT

La figure 1.2 nous montre donc les potentiels évoqués par les 3 conditions lors des deux périodes. Lorsque l'objet avait été rapidement identifié, ils ont observé un pic positif autour de 300ms après l'onset sémantique à la fois dans la première et la seconde période. En revanche, lorsque l'objet n'avait pas été identifié dans la première période, aucun pic n'a été observé dans la première période, alors qu'après présentation de l'image d'origine, un pic était présent lors de la seconde période, identique à celui obtenu pour des images identifiées rapidement. Enfin aucune activité n'était liée, au niveau de ces électrodes, à la présentation de textures, ni lors de la première période, ni lors de la seconde.

Ils ont donc conclu que cette activité électrophysiologique après l'onset sémantique était représentative de la reconnaissance de l'objet. Elle ne pouvait en tout cas pas être liée aux paramètres physiques de bas niveau des images puisqu'une même image évoquait des activités différentes en fonction de sa reconnaissance par le sujet.

Dans l'étude à suivre, nous avons donc voulu tester cette technique SWIFT chez le singe, pour vérifier si on obtenait une activité électrophysiologique associée à la reconnaissance d'une image en tant que cible de la tâche de catégorisation.

II) Les corrélats de la reconnaissance d'objets chez le singe Macaque

A) Introduction

Au 16ème siècle René Descartes décrivait le comportement animal sous une approche purement mécaniste dans sa théorie des « animaux-machines », niant de ce fait l'existence d'une conscience et de toute forme de pensée animale. Selon cette théorie le comportement animal obéirait au seul principe de causalité : un même stimulus de l'environnement entraînerait toujours la même réaction. Cette théorie, déjà largement remise en question par les contemporains de Descartes, est aujourd'hui réfutée et l'on s'accorde généralement pour attribuer une forme de pensée aux animaux, ou du moins aux vertébrés supérieurs. La question de la conscience a ainsi longtemps été réservée aux philosophes, mais elle est devenue un sujet central des recherches en neurosciences au cours des 20 dernières années. Parmi les pionniers, Crick et Koch (Crick and Koch, 1990, 1998) font l'hypothèse que la conscience émerge d'un ensemble d'interactions électriques et moléculaires entre neurones et que la principale fonction de la conscience visuelle serait de générer la meilleure interprétation possible d'une scène et de rendre cette interprétation disponible afin de produire une réponse comportementale adaptée. Dans cette étude, nous postulons que l'animal, et plus particulièrement le singe, possède une conscience, et nous nous proposons d'en étudier les corrélats neuraux.

1) La mesure de la conscience

Pour étudier les réponses électro-physiologiques liées à la perception consciente il faut qu'un même stimulus soit parfois visible, parfois invisible, tous paramètres égaux par ailleurs (contraste, luminance, temps d'exposition) afin de comparer les réponses obtenues dans les deux conditions. Les techniques habituellement utilisées pour répondre à cette problématique reposent sur la bistabilité perceptuelle avec des stimuli ambigus (une même image pouvant être perçue de deux manières différentes, comme par exemple le cube de Necker) ou bien des protocoles en rivalité binoculaire (chaque œil reçoit une image différente, mais le sujet n'en perçoit qu'une seule à la fois, et sa perception alterne entre les deux stimuli). Ces méthodes ont été appliquées chez le macaque et ont montré que leur perception variait à la même fréquence que chez

l'homme (Logothetis, 1998). Toutefois ces techniques supposent de faire confiance au sujet qui reporte sa perception et ne permettent pas de déterminer, chez le singe, si l'animal reconnait le contenu sémantique du stimulus (c'est-à-dire la signification de l'image, si elle existe). Selon les modèles actuels, les représentations des objets émergeraient progressivement au long des différentes aires visuelles constituant la voie ventrale. Ainsi, dans les premières étapes du traitement visuel, sont extraits les traits visuels simples, mais plus on progresse vers des aires visuelles supérieures, plus les neurones sont sélectifs de formes complexes, jusqu'à l'aire IT (cortex Inféro-Temporal) où les assemblées de neurones semblent sélectives de catégories d'objets(Tanaka, 1996, KIANI et al., 2007a). Jusqu'à présent la technique couramment utilisée pour étudier l'émergence de représentations d'objets dans les aires visuelles est la technique SSVEP (Steady State Visual Evoked Potential). Par l'enregistrement de l'électroencéphalogramme de surface, cette méthode permet de suivre et de caractériser l'activité électrique générée par un stimulus qui apparait et disparait à fréquence constante, et grâce à cela Kaspar et al (Kaspar et al., 2010) ont mis en évidence, chez l'humain, des différences en amplitude des potentiels évoqués par des images avec un contenu sémantique par rapport à des images sans signification. Toutefois au cours d'une séquence les caractéristiques physiques de l'image sont modulées parallèlement au contenu sémantique. On ne peut donc pas dissocier les composantes liées aux caractéristiques de bas niveau de celles liées au contenu sémantique dans les potentiels évoqués enregistrés.

Dans cette étude, nous avons choisi d'aborder la question de la perception visuelle consciente chez le macaque grâce à la technique SWIFT (Semantic Wavelet Induced Frequency Tagging) créée par Koenig & VanRullen (Koenig-Robert and VanRullen, 2012). Comme en SSVEP classique, la méthode SWIFT consiste à faire apparaître et disparaître un stimulus à fréquence constante au cours d'une séquence, et donc de caractériser l'activité électrique liée à sa perception. Mais contrairement à la SSVEP, les traits visuels de bas niveau (luminance globale de l'image, contraste, spectre fréquentiel) sont conservés tout au long de la séquence. Ainsi en s'affranchissant des modulations des traits visuels de bas niveau, il est possible d'isoler spécifiquement l'activité cérébrale liée au contenu sémantique de l'image. Cette technique, appliquée chez l'homme lors d'une

tâche de reconnaissance d'objets, a mis en évidence, grâce à l'analyse des potentiels évoqués, une activité électrique différentielle en fonction de la capacité du sujet à identifier l'objet présenté.

2) L'électroencéphalographie chez le singe vigile

Bien que l'EEG, et plus particulièrement l'analyse des potentiels évoqués, soit communément employée chez l'homme, son utilisation chez le singe vigile est récente et encore rare. Gould et al en 1974 (Gould et al., 1974) font partie des premiers à utiliser cette technologie chez le singe afin de déterminer l'implication de certaines aires cérébrales dans la discrimination visuelle. Trente- ans plus tard, Woodman et al (Woodman et al., 2007) comparent pour la première fois les potentiels évoqués obtenus chez le singe et chez l'homme durant la même tâche cognitive. Leurs conclusions sont édifiantes car montrent de fortes similitudes dans la forme des potentiels évoqués. Cette étude ouvre la voie à des explorations plus poussées de l'activité électrique du cortex chez le primate durant des tâches cognitives, et autorise une extrapolation prudente à l'homme.

Le singe utilisé dans cette étude est pourvu, depuis plusieurs années, d'électrodes intracorticales implantées dans les aires visuelles. Ces électrodes enregistrent l'activité électrique résultante d'une population neuronale au sein d'une aire visuelle. Elles fournissent donc une meilleure résolution spatiale que l'EEG classique qui ne permet pas de distinguer différentes sources situées dans des aires corticales proches. Toutefois les données collectées chez le singe en électrocorticographie (ECoG) étant de même nature que celles obtenues chez l'homme en EEG, elles peuvent être comparées.

3) La catégorisation visuelle : tâche cognitive pour accéder à la perception consciente chez le singe

Lors de l'application de la technique SWIFT chez l'homme, le sujet recevait comme consigne d'appuyer sur une touche dés lors qu'il reconnaissait un objet et qu'il pouvait le nommer. Les singes ne possédant pas le langage, nous avons choisi d'aborder la reconnaissance des objets via une tâche de catégorisation « Animal » versus « Non-Animal » à laquelle notre macaque a été très largement entraîné dans les années précédentes.

Reconnaitre un objet, et l'inclure dans une catégorie, suppose d'en avoir une représentation mentale. Au sein d'une catégorie, les objets partagent des caractéristiques communes, mais ne sont pas identiques. Pour déterminer l'appartenance d'un objet à une catégorie, il faut donc extraire les propriétés pertinentes et ignorer les autres. Les premières observations démontrant des capacités de catégorisation chez le primate ont été réalisées sur des singes vervets en milieu naturel. Cette espèce possède un large éventail de vocalisations, et produit des cris d'alarmes spécifiques du prédateur (reptile, félin, oiseau de proie), permettant un comportement de fuite adapté de l'ensemble du groupe(Seyfarth et al., 1980). A partir de ces observations, les neurobiologistes ont soumis les singes à des tests de catégorisation en laboratoire. Après avoir appris la règle de catégorisation sur un nombre réduit de stimuli, les singes sont capables de généraliser cette règle à de nouveaux stimuli (Yoshikubo, 1985, Fabre-Thorpe et al., 1998). Les chercheurs ont ainsi démontré que les singes étaient capables de catégoriser des images, en fonction de la présence ou de l'absence d'un objet cible, avec une précision équivalente à celle de l'homme, et une vitesse accrue (Fabre-Thorpe, 2003b). De plus, en 2011, Fize et al (Fize et al., 2011)ont mis en évidence une modulation identique du temps de réaction chez l'homme et le singe, lors d'une tâche de catégorisation « Animal » versus « Non-Animal » en fonction de la taille de la cible. Ainsi les deux espèces présentent une augmentation de 30ms de leur temps de réaction pour la catégorisation de petits animaux (en nombre de pixels). Cela semble indiquer que les processus impliqués dans la catégorisation visuelle sont les mêmes chez l'homme et le singe.

4) Les bases neurales de la catégorisation visuelle

Partant de ces données comportementales, se pose alors la question des bases cérébrales et des mécanismes sous-tendant de telles aptitudes cognitives chez les primates et les hommes. Les études anatomiques comparatives ont mis en évidence une forte similarité organisationnelle du cortex visuel entre le singe macaque et l'homme (Hubel and Wiesel, 1977, Imbert, 1999). L'analogie anatomique a par la suite été complétée par des travaux fonctionnels. Durant des années, l'exploration des voies visuelles chez le singe se faisait sous anesthésie, mais avec l'évolution des techniques d'imagerie et d'enregistrement, les études fonctionnelles sont désormais possibles chez le singe vigile, autorisant alors l'accès directs aux supports neuraux des tâches cognitives. Ainsi Kiani et al en 2007 (Kiani et al.,

2007b)ont soumis des macaques à des tâches de fixation passive et de *serial delayed matching to sample* tout en procédant à des enregistrements unitaires de neurones dans IT. Ils ont mis en évidence de cette manière que les catégories visuelles étaient encodées dans IT par un pattern d'activation distribué. En 2008, Kriegeskorte (Kriegeskorte et al., 2008b) a été plus loin et a démontré via une étude en IRM fonctionnelle chez l'homme, et des enregistrements unitaires chez le singe, que les catégories visuelles étaient codées de manière similaire chez ces deux espèces. Pour ce faire, il a comparé au sein de chaque espèce, les profils d'activation générés par la présentation d'un stimulus, et les a regroupés par similarité. Enfin il a examiné les groupes de stimuli ainsi formés et a constaté que ces ensembles étaient constitués des mêmes items chez l'homme et le singe (ex : les visages forment une catégorie, les animaux à 4 pattes en forment une autre etc…). Toutefois l'analyse par enregistrements unitaires présente des limitations. Tout d'abord on dispose de très peu de résultats obtenus par cette méthode chez l'homme, et ceux que l'on possède on été acquis chez des sujets malades, il est donc délicat d'extrapoler ces données au sujet humain sain. De plus ces résultats rendent compte de l'activité d'un nombre limité de neurones mais ne reflètent pas obligatoirement l'activité globale de la zone étudiée. Whittingstall & Logothetis en 2009 (Whittingstall and Logothetis, 2009), ont mené des expériences chez le macaque éveillé durant lesquelles ils enregistraient en parallèle des signaux EEG et des neurones de l'aire V1. Ces données étaient collectées alors que les macaques regardaient un film. Ils en ont conclu que la puissance du signal EEG dans les hautes fréquences et la phase dans les basses fréquences étaient corrélées au taux de décharge des neurones. Cette étude fait donc le lien entre l'enregistrement unitaire et l'EEG, et montre surtout qu'il n'y a pas de relation linéaire simple entre le taux de décharge d'une population finie de neurones et l'activité globale d'une aire cérébrale.

5) Problématique et protocole

Dans notre étude nous proposons donc une méthode innovante pour accéder à la conscience de la perception visuelle chez le macaque. Pour ce faire nous soumettons le singe à une tâche cognitive coûteuse de catégorisation « Animal » versus « Non-Animal ». Les séquences SWIFT utilisées comme stimuli rendent la cible (i.e. l'animal) difficilement détectable, permettant d'obtenir les deux conditions perceptuelles

souhaitées pour le même stimulus, à savoir si l'animal a été reconnu ou non. Durant cette tâche de catégorisation nous avons enregistré les signaux ECoG via les électrodes implantées dans le cortex visuel de l'animal. En analysant ces signaux en ERP (*Event Related Potential* = Potentiel évoqué) conjointement avec les performances comportementales, nous tentons donc de déterminer les corrélats électrophysiologiques de la reconnaissance consciente du contenu sémantique de l'image. Nous avons également vérifié les données psychophysiques de notre singe durant toutes les sessions d'expérimentations, afin de nous assurer que les réponses n'étaient pas données au hasard, aussi bien en termes d'identité du stimulus que de latence relativement à l'apparition de l'image non bruitée dans la séquence.

B) Matériel et méthodes

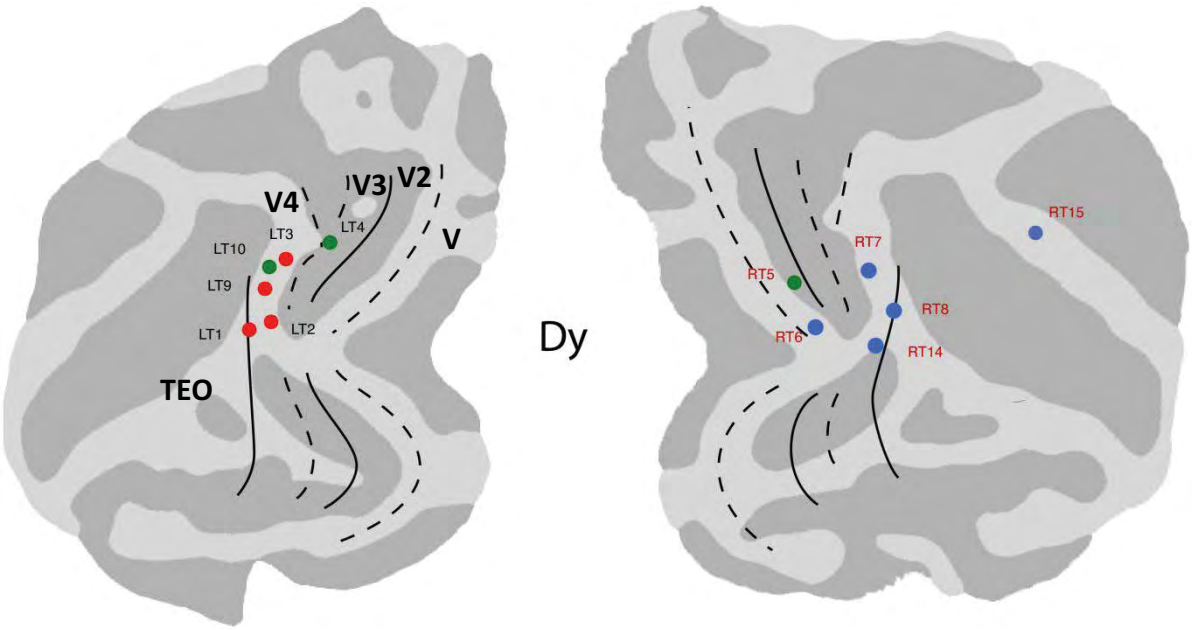
6) Le sujet

Un mâle adulte âgé de 19 ans a participé à cette étude, le seul doté d'électrodes implantées dans le cortex visuel. Ce singe était déjà expert dans une tâche de go/no-go en catégorisation visuelle rapide Animal vs Non-animal.

L'animal est décédé de cause naturelle en décembre 2013.

7) L'implantation des électrodes

Le singe porte 13 électrodes de mesure sous-durales dans les aires visuelles (voir figure ci-dessous), une électrode de référence et une autre de masse placées toutes deux dans le cortex préfrontal. Ce sont des électrodes filaires 200µm de diamètre en acier chirurgical, isolées entièrement d'un enrobage de Téflon excepté aux extrémités. Les électrodes se rejoignent au niveau d'un connecteur fixé sur la tête de l'animal par du ciment dentaire.



Gauche *Droite*

Figure 1.3 : sites d'implantation des électrodes dans les 2 hémisphères cérébraux du singe. Cortex étalé (zones sombres= sulci, zones claires= giri). La plupart des électrodes se trouvent sur la convexité du gyrus préluné correspondant à l'aire fonctionnelle V4.

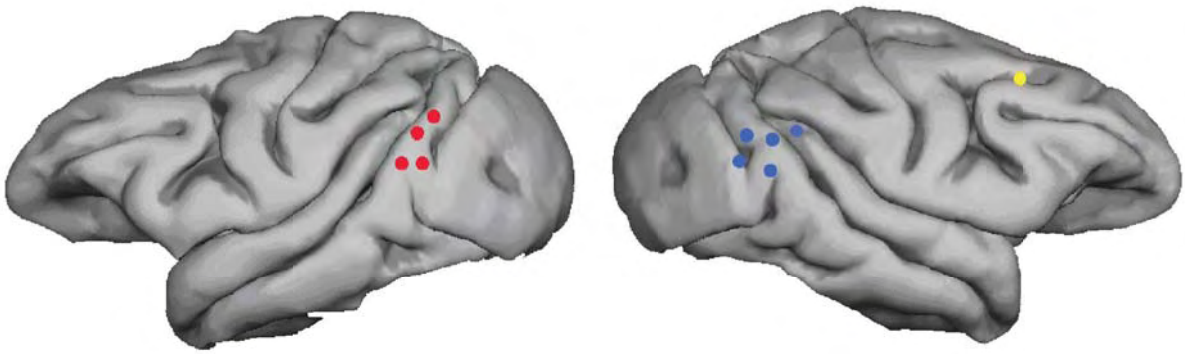


Figure 1.4 : sites d'implantation des électrodes dans les 2 hémisphères cérébraux du singe. Cortex 3D. En jaune : une électrode frontale placée dans le sillon arqué (proche de la région du FEF – Frontal Eye Field) Les électrodes apparaissant en vert sur la 1ère figure, ne peuvent être représentées sur la 2nde.

8) Les stimuli

Les images utilisées sont des photographies (non soumises à droit d'auteur) d'animaux pour les cibles (50% des stimuli) ou bien de paysages/ d'objets/d'intérieurs pour les distracteurs (25% des stimuli). Ces images ont préalablement été égalisées en contraste (12%), en luminance (50%) et en taille (365 x 237 pixels).

Figure 1.5 : Exemples de stimuli cibles

Figure 1.6 : exemple de stimuli distracteurs « objets »

La seconde étape a consisté à générer des textures (25% des stimuli) à partir de ces images réelles, grâce à l'algorithme de Simoncelli (Portilla and Simoncelli, 2000a). Ces textures ont les mêmes caractéristiques que les images d'origine (taille, luminance, contraste) mais ne contiennent pas d'information sémantique.

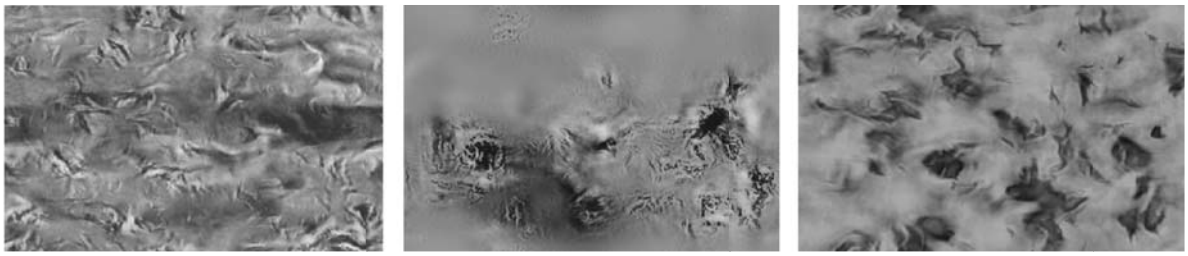


Figure 1.7 : Exemples de stimuli distracteurs « textures »

Enfin ces images sont transformées en séquences SWIFT grâce à l'algorithme de Koenig et VanRullen. La première étape consistait à appliquer une transformée en ondelettes, à 6 niveaux de décomposition (figures 1.8), c'est-à-dire que l'image a été convertie en une pyramide multi-échelles des orientations spatiales. A chaque échelle et à chaque position dans l'image, le contour local est représenté par un vecteur 3D représentant les poids des 3 orientations : horizontale, verticale et diagonale. La longueur du vecteur représente l'énergie du contour local.

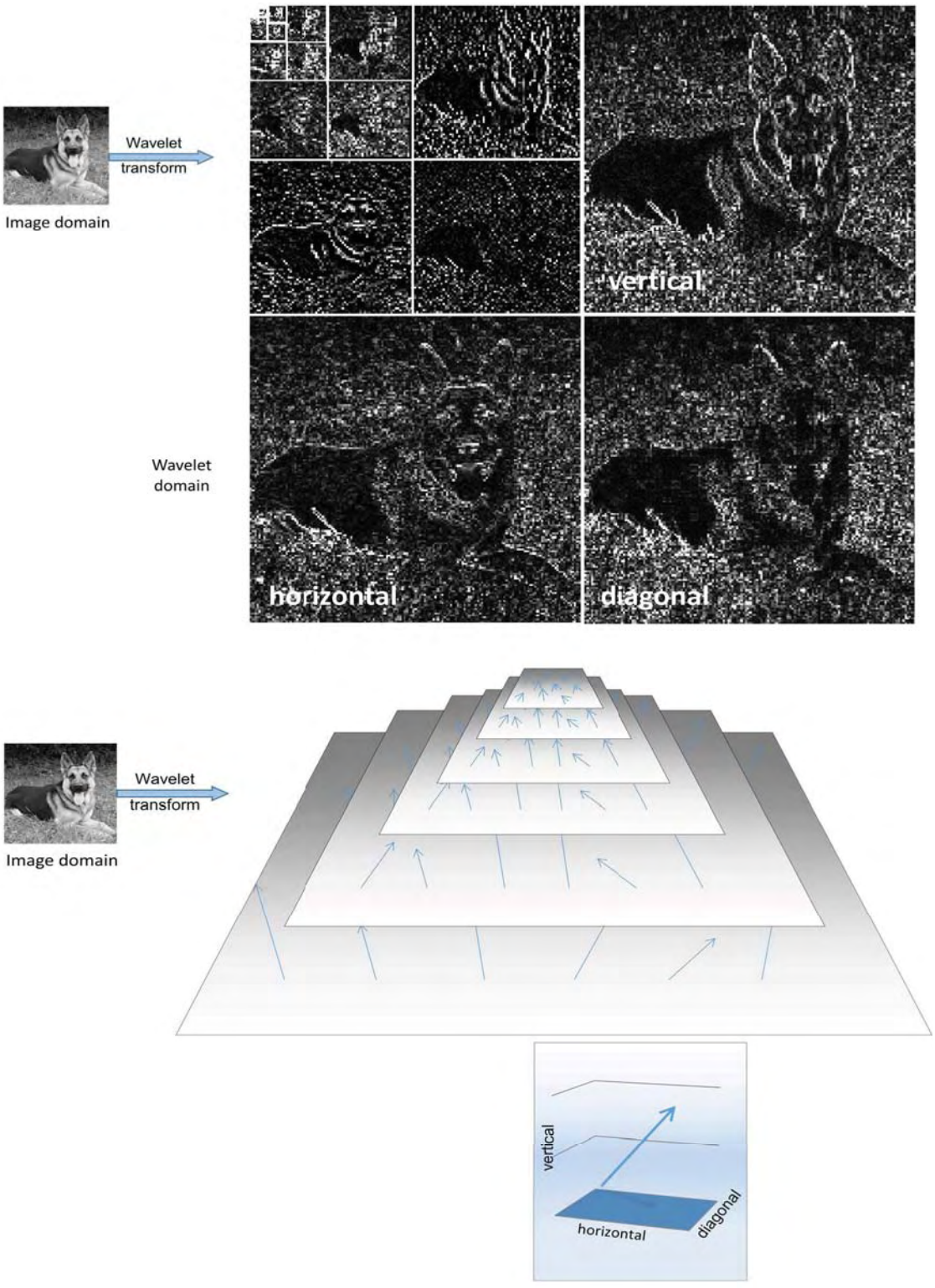


Figure 1.8 : Décomposée des images en ondelettes, figure fournie par Roger Koenig-Robert

Dans une seconde étape, à chaque échelle et chaque position 2 nouveaux vecteurs ont été créés aléatoirement, de même longueur que le premier (l'énergie du contour local est donc conservée). Ces 3 vecteurs définissent donc un cercle unique sur une sphère isoénergétique (figure 1.9). Le scrambling en ondelette était alors réalisé en faisant tourner le vecteur d'origine sur ce cercle ainsi décrit. Certains points de l'image ne faisaient le tour du cercle qu'une seule fois par cycle (fréquence fondamentale f0), tandis que d'autres pouvaient le faire jusqu'à 5 fois (2nd harmonique, 3ème, 4ème et 5ème). Les différentes fréquences étaient attribuées de manière aléatoire et équiprobable en chaque point de l'image. Par construction, l'image apparaissait complètement reconstruite une fois par cycle, quand tous les pixels revenaient au point de départ.

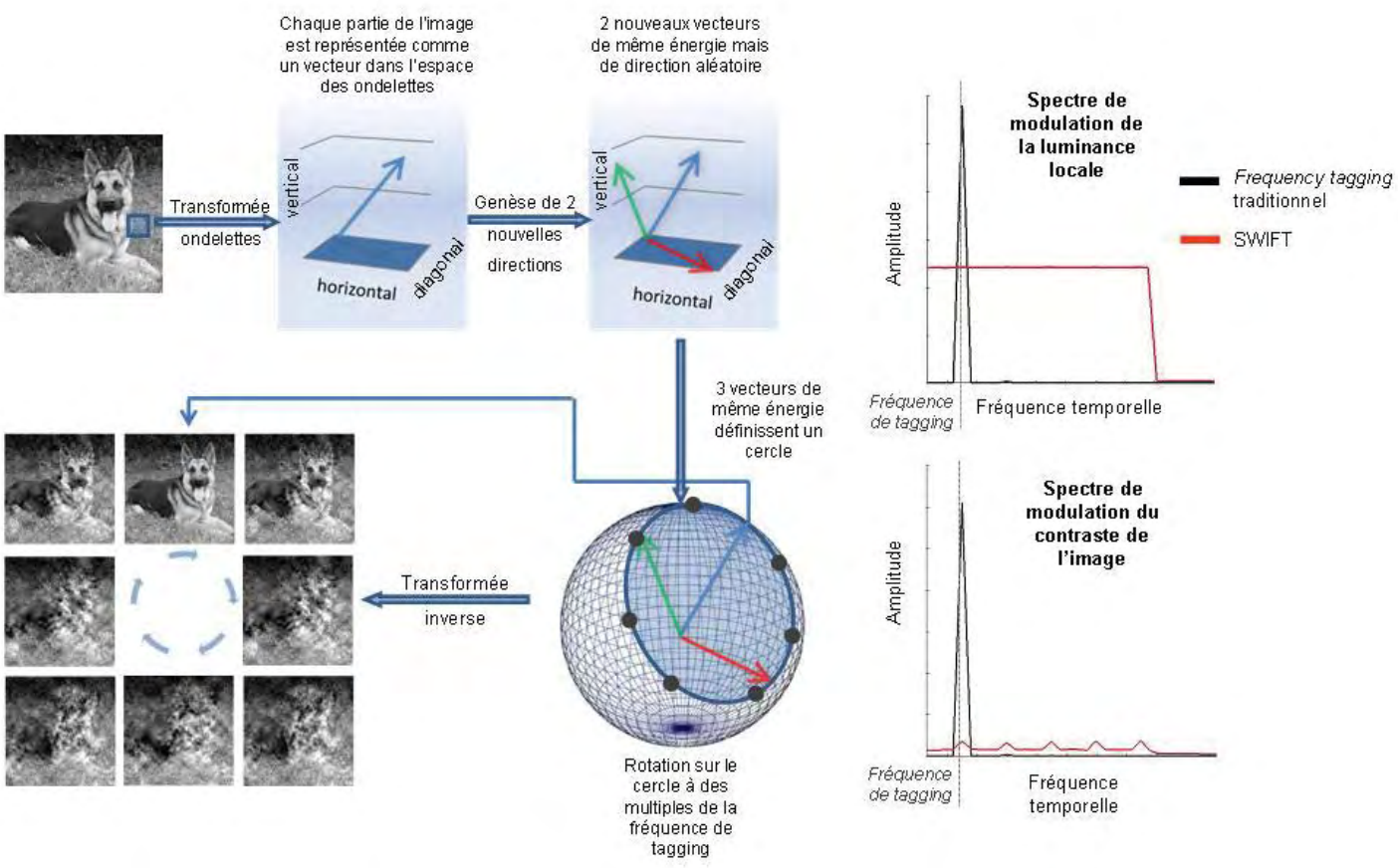


Figure 1.9 : Création d'une séquence SWIFT

A partir de chaque image, nous avons généré trois cycles différents, constitués de 47 images. Au cours d'un cycle l'image d'origine n'apparait qu'une seule fois. Une séquence est constituée de la combinaison aléatoire de trois cycles et demi (un même cycle peut être répété au cours d'une séquence).

Un cycle dure 671ms, une séquence complète dure donc 2,4 secondes au cours de laquelles le contenu sémantique sera visible à trois reprises (trois onsets sémantiques). Le premier onset sémantique apparait 471ms après le début de la séquence.

9) La tâche

Le singe devait effectuer une tâche de catégorisation visuelle Animal vs Non-Animal (en go/no-go).

Il était assis dans une chaise de contention (Chris Instrument), tête et mains libres. Pour déclencher un essai, il lui fallait poser sa main sur le boitier réponse pendant une seconde, et la laisser ainsi. Si l'image contenait un animal, il devait relâcher le bouton réponse et toucher l'écran (réponse go). Si l'image était un objet, une scène ou une texture, il devait garder sa main posée sur le boitier (réponse no-go). Les réponses motrices étaient enregistrées durant toute la durée de la séquence et jusqu'à 800ms après le dernier onset sémantique.

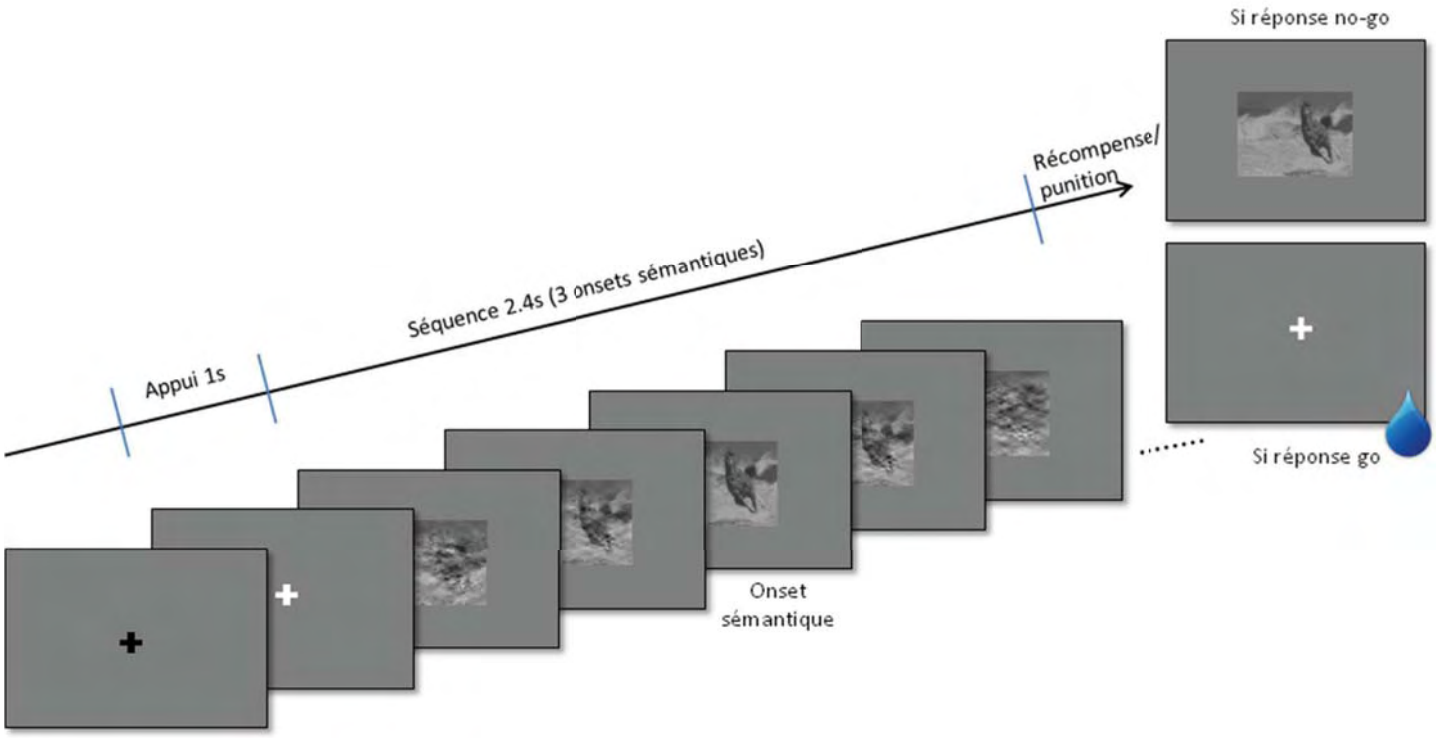


Figure 1.10 : Déroulement d'un essai cible. Lorsque l'image est un distracteur, c'est la réponse no-go qui est récompensée

Pour chaque réponse correcte, il était récompensé à la fin de l'essai par quelques gouttes d'eau grâce à un dispositif de récompense fixé à la chaise (Chris Instrument). Pour chaque réponse incorrecte, l'image d'origine réapparaissait pendant deux secondes et il n'obtenait pas de récompense. Afin de garder un niveau de motivation suffisant pour l'expérience, l'animal a été placé en restriction hydrique pendant toute la durée de l'expérimentation. Durant les sessions quotidiennes, il pouvait travailler et donc boire ad libitum. Au cours d'une session, la récompense était augmentée graduellement afin de conserver un bon niveau de motivation. Durant un essai contenant une cible, seule la première réponse motrice était enregistrée et nécessaire. L'animal a ainsi peu à peu appris à ne répondre qu'une seule fois par essai et à obtenir sa récompense de manière différée (à la fin de l'essai).

Ces procédures ont été approuvées par le comité d'éthique régional dont dépend le laboratoire (agrément ref. MP/05/05/01/05, C2EA-14 comité d'éthique de Marseille).

Chaque séance de travail contenait 20 images très familières (10 animaux/10 objets ou scènes), 20 images nouvelles (10 animaux, 5 objets ou scènes et 5 textures) ainsi que 20

images « récentes » c'est-à-dire déjà présentées en tant que nouvelles lors de séances précédentes (10 animaux, 5 objets ou scènes et 5 textures). Chaque nouvelle image était présentée à 10 reprises sous la condition que l'animal fasse au minimum 600 essais. Il pouvait ensuite travailler ad libitum, les images récentes et familières étant alors présentées 10 fois ou plus. En moyenne, une séance de travail durait 45 minutes.

Avant de mener cette expérience, nous avons progressivement entraîné le singe à la tâche. En effet, bien qu'il soit déjà expert en catégorisation visuelle rapide Animal/Non-Animal, il n'avait été jusque là confronté qu'à des images fixes flashées. Dans cette étude, les stimulations visuelles étaient beaucoup plus longues. Nous avons donc introduit graduellement des essais contenant des séquences SWIFT courtes au milieu d'essais en images fixes flashées. Nous avons ensuite allongé peu à peu la durée des séquences, afin que le singe s'habitue à recevoir sa récompense de manière différée par rapport à sa réponse.

10) Enregistrement des données

Durant les sessions de test, les signaux ECoG du singe sont enregistrés grâce à une chaine d'enregistrement EEG standard (logiciel NeuroScan 4.2. et système d'acquisition Synamps, Neuroscan Inc.). Le signal était échantillonné à 1000Hz et filtré par bande passante entre 1 et 200Hz. A cet enregistrement ont été associés différents *triggers* correspondant au numéro d'essai, le type de séquence (i.e. animal, texture ou objet), les latences d'apparition des contenus sémantiques de chaque séquence, le temps de réaction du singe et la fin de la séquence. Le signal était ensuite sous-échantillonné *offline* à 500Hz et filtré passe-bas à 30Hz.

Les données comportementales étaient collectées par Matlab grâce à la psychtoolbox (Brainard, 1997), et comportaient toutes les caractéristiques de la séquence, les réponses et les temps de réaction.

11) Analyse des données

Les données comportementales ont été analysées par Matlab. Pour chaque session ont été observés les taux de réussite par catégorie d'image et également par item unique pour les cibles. Nous avons ainsi cherché à isoler les résultats pour des cibles qui auraient été ratées durant les n (minimum n=1) premières présentations puis réussies au minimum trois fois consécutives. Les temps de réaction ont été également analysés afin de déterminer le cycle au cours duquel l'animal donnait sa réponse.

Nous avons analysé les enregistrements EEG grâce à EEGlab. Nous avons séparé les potentiels évoqués (moyenne des tracés ECoG après soustraction d'une ligne de base) pour chacune des 3 apparitions de l'image contenant l'information sémantique. Les données ont été traitées en fonction de la catégorie de l'image et de la réponse du singe. Enfin les données ont été nettoyées grâce à un filtre passe-bande afin d'éliminer d'éventuels artéfacts. Par ailleurs nous avons également mené une analyse par item unique en comparant les réponses électrophysiologiques pour une même cible, dans les deux conditions : reconnue ou non-reconnue.

C) Résultats

1) Résultats comportementaux

a) Performances globales

Malgré l'utilisation de séquences longues et la présence de bruit dans la stimulation visuelle, le singe atteint une réussite globale de 73 %. Il répond significativement d'avantage sur les cibles que sur les distracteurs (χ^2=4595, $p<10^{-5}$, ddl=1), et réalise donc la tâche « Animal vs Non-Animal » dans des conditions très différentes de sa tâche usuelle.

b) Performances sur les images nouvelles

Sauf indication contraire, nous désignerons par « images nouvelles », les 20 séquences introduites à chaque session et présentées jusqu'à dix fois lors de leurs premières sessions. Le singe effectue véritablement une catégorisation puisque il réalise correctement la tâche et se montre capable de la généraliser à de nouvelles images.

Le singe effectuait environ 45% de réponses go sur les nouvelles cibles, contre seulement 15% sur de nouveaux distracteurs (χ^2=664, $p<10^{-5}$, ddl=1). Ceci démontre qu'il est capable de généraliser la tâche à de nouvelles images, même si ses performances chutent par rapport aux images familières (82% de réponses sur les cibles familières)

c) Apprentissage

Les performances du singe étaient globalement stables au cours d'une session, avec une légère chute pour les dernières présentations qui pouvait s'expliquer par sa moindre motivation (satiété atteinte en fin de session). Toutefois dès les premières présentations d'une nouvelle séquence, le singe était capable de catégoriser correctement 50% des cibles et 80% des distracteurs, ce qui est significativement au dessus de la chance (χ^2=67, $p<10^{-5}$, ddl=1).

Nous avons également vérifié que l'animal était capable d'apprendre de ses erreurs et ne persistait pas dans des réponses incorrectes. Ainsi, nous avons constaté que lorsqu'il s'était trompé à la première présentation d'une cible, il se corrigeait dans près de 30% des cas et donnait une réponse correcte dès la seconde présentation (qui pouvait intervenir une centaine d'essais plus tard). De même, une erreur sur la première présentation de distracteurs (objets ou textures) est corrigée à 60% à la seconde

présentation. Toutefois cette probabilité de correction diminue avec le nombre de présentations, montrant une tendance générale du singe à persister dans ses erreurs au delà de trois réponses incorrectes.

Cette capacité d'apprentissage (même imparfaite) justifie l'analyse des résultats électrophysiologiques par item unique, c'est à dire en sélectionnant les séquences ratées les n premières présentations puis réussies au minimum trois fois consécutivement (par analogie avec l'homme, Koenig et al).

Les réponses aux images familières sont bien plus rapides qu'aux images nouvelles (test de Student : $t=-11.1$, $p<10^{-3}$). Etant donné le nombre important de réponses correctes avant même le premier onset sur les images familières, nous pouvons supposer qu'il connait les séquences par cœur. De fait les réponses EEG obtenues pour ces images ne seront pas prises en compte dans l'analyse des signaux électrophysiologiques.

2) Résultats électrophysiologiques

Nous ne présenterons ici que les potentiels évoqués par les images nouvelles et récentes. D'autre part, les signaux électrophysiologiques obtenus pour des textures et des objets étaient similaires (non représentés), ils sont regroupés dans les analyses suivantes dans une même catégorie « distracteur ».

a) Résultats généraux

Sont représentés dans la Figure 1.11 les potentiels évoqués par des cibles reconnues (H), des cibles non reconnues (M), des distracteurs correctement catégorisé (CR) et des distracteurs ayant généré des fausses alarmes (FA). La figure suivante illustre les activités des aires visuelles précoces V2 et V3, et des aires de complexité intermédiaires V4 et TEO, pour lesquelles les activités de quatre électrodes représentatives ont été moyennées sur les deux premiers cycles afin de disposer de comparaisons statistiques robustes.

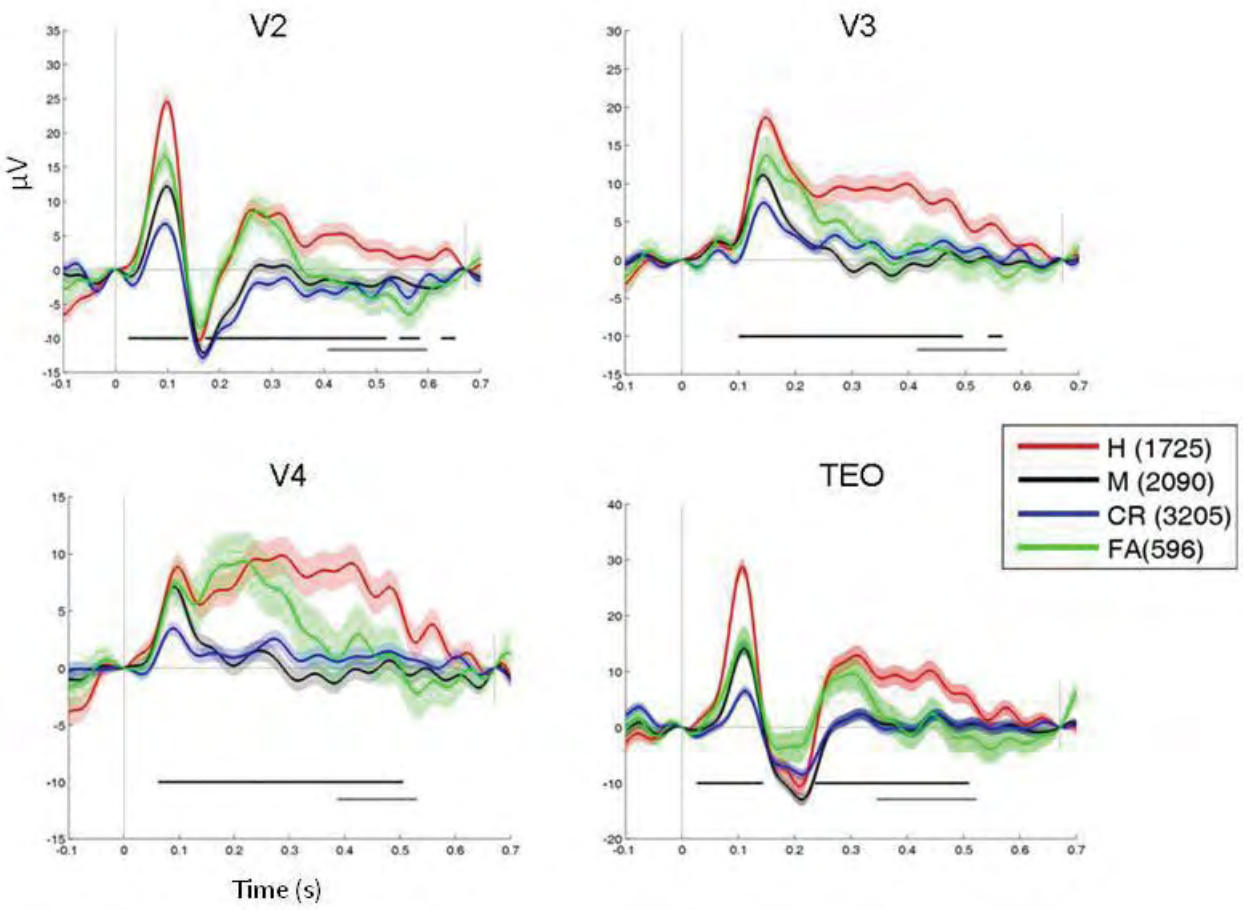


Figure 1.11: Potentiels évoqués par les nouvelles images, moyennés pour les 2 premiers onsets sémantiques. V2 : électrode RT6 ; V3 : électrode LT4 ; V4 : électrode LT10 ; TEO : électrode LT1. Ligne noire : différence significative entre les cibles reconnues et les distracteurs no-go (test de Student, $p<0.01$) Ligne grise : différence significative entre les cibles reconnues et les fausses alarmes (test de Student, $p<0.05$)

Le 1er pic positif vers 100 ms (onde P1), suivi par une déflection négative sur la plupart des électrodes (électrodes V2 : 170ms et TEO : 220ms, figure 1.11), a déjà été décrit dans les études antérieures du groupe. Mais ici le nombre d'essais associés à des erreurs permet pour la première fois d'observer la modulation de cette onde avec les différents types d'essais (H, FA, M, CR). Dans notre étude cette onde présente la particularité de diminuer au cours des trois cycles, et d'être d'autant plus ample que la séquence est associée à une réponse « go » (H et FA). Ceci est en accord avec Cauchoix et al (Cauchoix et al., 2015), qui ont montré que l'amplitude de cette onde P1 était corrélée dès 100ms à la réponse de

l'animal, dans le cadre de la catégorisation visuelle ultra-rapide à laquelle il a été longuement entrainé.

Le deuxième pic positif vers 250 ms (onde P2) semble être d'avantage spécifique des particularités du protocole expérimental. Cette onde n'est présente que lors des deux premiers cycles, et uniquement pour les cibles reconnues et les distracteurs ayant généré une fausse alarme. Dans la phase précoce de l'onde P2, la courbe des réponses go sur les cibles et celle des fausses alarmes se superposent. Dans une phase plus tardive (vers environ 350 ms) ces deux courbes se dissocient. Cette onde P2 pourrait être l'équivalent de l'onde associée à la perception consciente d'un stimulus déjà observée chez l'homme (Lamy et al., 2009), d'autant plus ample que le stimulus est perçu consciemment. Nos résultats paraissent donc confirmer l'hypothèse de l'existence de cette onde chez le singe.

Afin d'éliminer la possibilité que l'onde P2 soit directement liée à la réponse motrice, nous avons analysé les potentiels évoqués par chaque onset sémantique au cours d'une séquence, et mis en parallèle ces résultats électrophysiologiques avec les temps de réaction de l'animal (pour plus de lisibilité, les potentiels évoqués par les distracteurs correctement rejetés ne sont pas représentés sur la figure suivante).

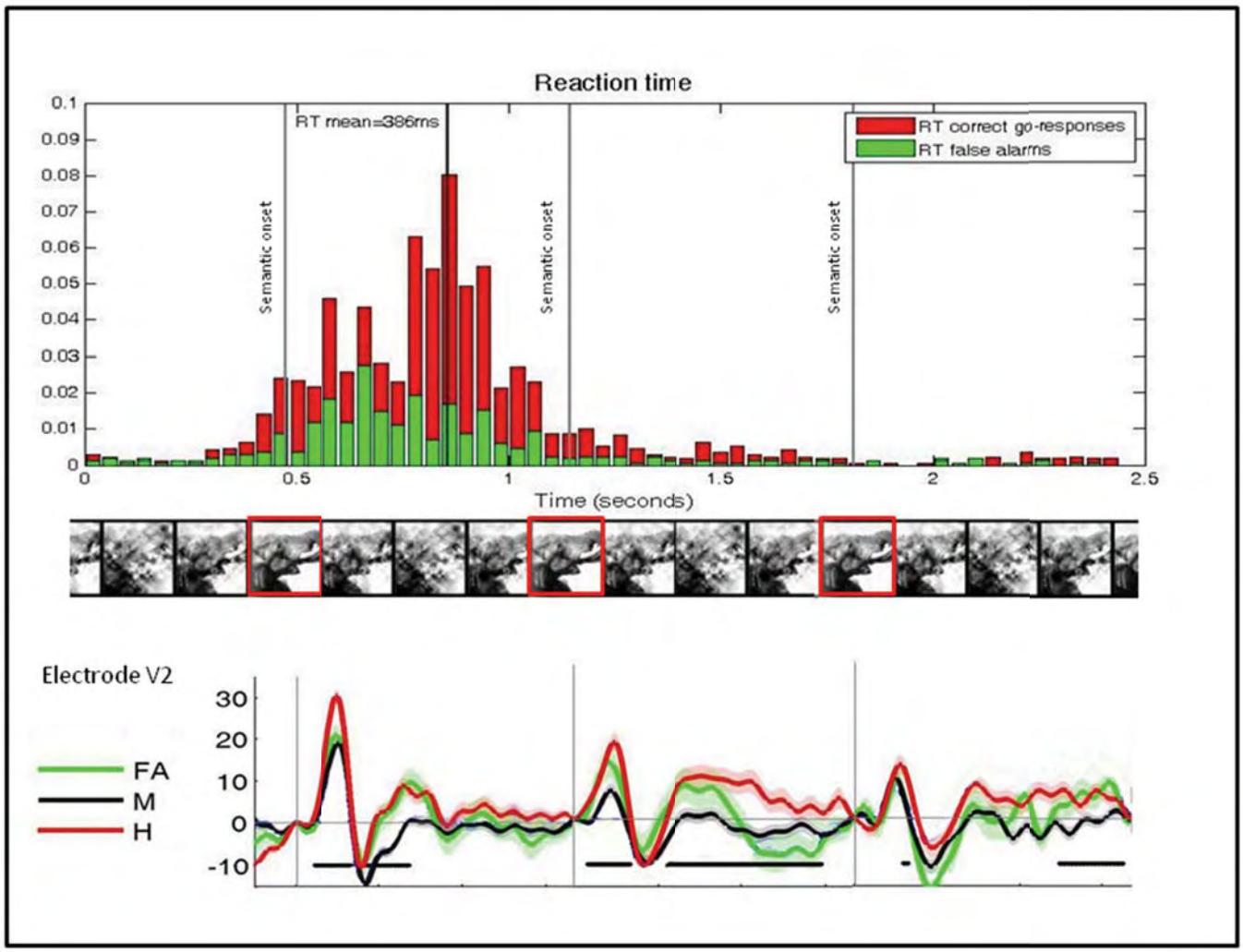


Figure 1.12 : Décours temporel du signal ECoG et répartition des temps de réaction au cours d'un essai.

Nous observons alors que le singe répond dans plus de 90% des essais au premier onset sémantique. Si nous observons des artéfacts dus à la réponse motrice, ils ne peuvent être qu'au niveau du premier potentiel évoqué dans nos tracés électrophysiologiques. Or nous observons que l'onde P2 est la plus large après le second onset sémantique, alors que le taux de réponses motrices est très faible. Cette onde P2 ne peut donc pas s'expliquer par la réponse motrice de l'animal.

De plus, de manière remarquable, l'onde P2 associée aux fausses alarmes est parfaitement superposée à celle associée aux hits lors du premier cycle mais redescend au niveau de base plus rapidement que celle associée aux cibles réussies (vers 400 ms, ligne de significativité grise sur figure 1.11). Tout se passe comme si cette activité était

liée à la « prise de conscience » qu'il n'y avait pas de cible dans l'image à laquelle le singe avait dans un premier temps répondu.

Enfin dans le 3ème cycle, les signaux sont très bruités. Nous pouvons raisonnablement supposer que l'animal ne regardait plus l'écran au moment où l'image apparaissait pour la troisième fois, étant donné qu'il avait déjà produit sa réponse.

Ces résultats montrent donc un ensemble d'activités différentielles en fonction de la catégorie du stimulus et de la réponse faite par le singe. Toutefois ils ne montrent pas directement la différence de réponse obtenue pour un même stimulus en fonction de sa perception consciente ou de l'absence de sa perception en tant que cible. C'est ce que l'on va montrer dans l'analyse suivante.

b) Analyse par item unique

Par analogie directe à l'analyse effectuée par Koenig et al chez l'homme, nous cherchons désormais à mettre en évidence une différence éventuelle de potentiels évoqués par un même stimulus, dans le cas où il aurait été correctement reconnu et dans le cas où il aurait généré une réponse incorrecte. Pour cela nous avons sélectionné les essais pour des cibles nouvelles ou récentes, qui auraient été ratées les n premières présentations (au minimum n=1) au sein d'une session puis réussies au minimum trois fois consécutives. Nous avons ensuite analysé les potentiels évoqués de la dernière erreur et de la première réussite pour ces cibles. Ainsi on peut affirmer que les réponses observées ne dépendent pas des caractéristiques intrinsèques de l'image (animal plus ou moins facilement reconnaissable, de taille importante, etc…) mais bel et bien de la perception qu'en a l'animal. Nous avons ainsi sélectionné 69 images qui répondaient à ces critères, mais ne présentons les tracés dans la condition « reconnue » que de 67 d'entre elles à cause d'importants artéfacts dans le signal. Nous avons ensuite moyenné comme précédemment les deux premiers cycles.

Les résultats sont présentés pour 4 électrodes dans 4 aires visuelles différentes, et sont visibles sur la page suivante.

Nous pouvons observer sur la figure 1.13 que l'onde P1 ne porte plus de différence significative à $p<0.05$ entre la condition reconnue ou non-reconnue, à l'exception de l'électrode RT6 sur un temps limité. Cette exception est relative, étant donné le niveau

élevé de bruit causé par le faible nombre d'essais sélectionnés. Il est donc raisonnable de penser que l'activité différentielle observée précédemment sur l'onde P1 ne correspondait qu'à une différence d'entrée visuelle et non de traitement de plus haut niveau (attention, perception), puisque l'entrée visuelle est ici identique entre les essais de type H et M.

Par contre nous observons des différences importantes pour l'onde P2, positive lorsque la cible a été reconnue. Les différences observées entre les deux tracés ne sont pas continûment significatives sur la durée de l'onde P2, mais l'amplitude de l'effet et sa durée, et l'écart constaté entre les erreurs associées (SEM), laissent penser que cette significativité se renforcerait avec un nombre d'essais plus important. Il est certain cependant que l'onde P2 n'est donc pas associée aux caractéristiques intrinsèques de l'image, mais bien plutôt à sa perception en tant que cible comme proposé plus haut.

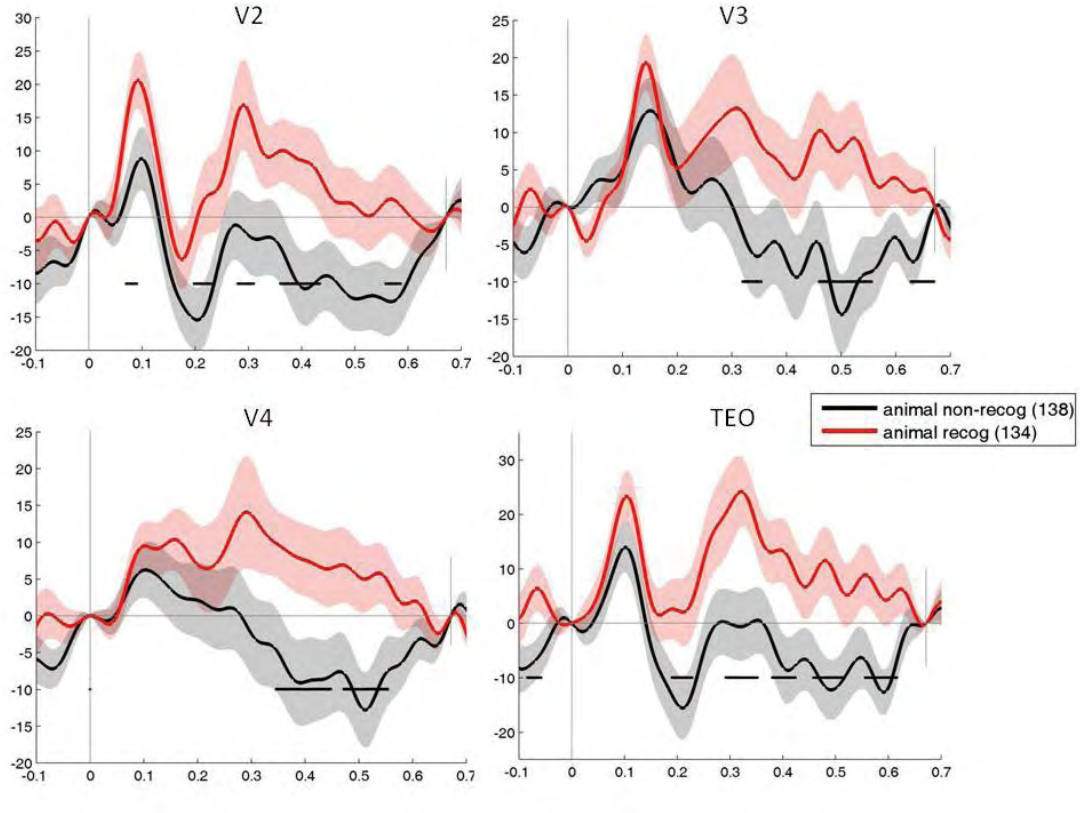


Figure 1.13 : Potentiels évoqués par des cibles nouvelles ou récentes, dans 2 conditions : reconnue ou non reconnue. Les lignes noires représentent la significativité de la différence entre les deux courbes (test de Student, $p<0.05$). V2 : électrode RT6, V3 : électrode LT4 ; V4 : électrode LT10, TEO : électrode LT1.

D) Discussion

1) Résultats généraux

Nos résultats ont montré que le singe réalisait la tâche de catégorisation avec des performances significativement au dessus de la chance. Nous avons également pu démontrer qu'il était capable de changer sa réponse après avoir fait plusieurs erreurs sur un même stimulus, ce qui nous a permis de mener une analyse du signal électrophysiologique par item unique en analogie avec ce qui a été fait chez l'homme (Koenig-Robert and VanRullen, 2013). Grâce à l'étude des temps de réaction nous avons mis en évidence qu'il répondait en moyenne après le premier onset sémantique d'une séquence, et qu'il effectuait une catégorisation visuelle rapide. L'analyse des signaux ECoG a montré que l'amplitude de l'onde P1 était reliée à la réponse du singe, du moins sur les deux premiers onsets sémantiques d'une séquence, et qu'une onde P2 apparaissait uniquement quand le singe avait catégorisé une image comme « animal » même si le stimulus était un distracteur. Toutefois cette onde P2 était plus étendue pour des réponses correctes que pour les fausses alarmes, comme si le singe *prenait conscience* de son erreur, notamment au deuxième onset sémantique de la séquence. Enfin l'analyse en item unique a mis en évidence une différence importante, notamment pour l'onde P2, entre les potentiels évoqués par de mêmes séquences reconnues et non reconnues en tant que cibles.

2) Parallèle Homme/singe

Ces résultats viennent renforcer ceux obtenus chez l'homme par(Koenig-Robert and VanRullen, 2013). Dans leur expérience, Koenig et VanRullen ont présenté des séquences SWIFT à des sujets humains.

Les séquences contenaient soit des images réelles (i.e. avec un contenu sémantique) soit des textures. Durant une première phase de 30 secondes, les sujets devaient appuyer sur un bouton s'ils reconnaissaient un objet qu'ils étaient capables de nommer ou un sur autre bouton s'ils percevaient un objet mais qu'ils n'étaient pas capables de l'identifier. Ils disposaient de toute la durée de la séquence pour donner leur réponse. L'image d'origine était ensuite présentée pendant deux secondes, puis une deuxième séquence SWIFT identique à la première mais plus courte était présentée. Les signaux EEG étaient

enregistrés pendant toute la durée de l'expérience. Ils ont alors comparé les potentiels évoqués par des images réelles reconnues, non reconnues et par les textures lors de la première période (dite naïve), puis les potentiels évoqués durant la seconde phase par les images déjà reconnues à la première période et les potentiels évoqués par les images reconnues seulement durant la seconde période (figure 1.14).

Ils ont ainsi montré une activité différentielle entre 250 et 400ms en fonction de la reconnaissance du stimulus par le sujet. Ils en concluent que c'est à cette latence qu'apparaissent les corrélats neuraux de la perception visuelle consciente chez l'homme.

Dans notre étude, l'onde P2 chez le macaque semble de la même façon caractériser la perception consciente du stimulus. Cette onde apparait également environ 250ms après l'onset sémantique, mais s'étale jusqu'à 550ms après l'onset en moyenne. Cette onde est largement distribuée dans les aires visuelles puisqu'on la retrouve pour 10 des 12 électrodes (sauf RT8 et RT 14). Contrairement à l'étude chez l'homme, pour laquelle la précision spatiale de l'electroencéphalogramme de surface ne permet pas d'attribuer cet effet à une région corticale précise, nos résultats permettent d'affirmer que cette onde est présente dans le système visuel dès les aires précoces. D'autre part, étant donné que la P1 reflète l'implication du système visuel dans le traitement feed-forward de l'entrée visuelle, le fait que dans notre étude une P2 très ample soit visible sur les électrodes peu impliquées dans ce traitement ascendant démontre une large implication de toutes les régions visuelles (non limité aux régions corticales ayant traité l'information ascendante) dans un mécanisme distribué vraisemblablement en retour (feedback).

Enfin, nos résultats ont pu montrer l'existence de l'onde P2 dans le cas des fausses alarmes, ce que ne permettait pas le protocole de Koenig et VanRullen. Cette onde P2-FA apparaît aux mêmes latences que pour les cibles réussies, mais revient plus rapidement au niveau de base suite à la deuxième présentation de l'onset sémantique. Ce résultat permet d'affiner l'interprétation de l'onde P2, dans le sens où cette onde pourrait refléter les mécanismes supportant les moments où le stimulus est perçu *en tant que* cible.

Cette expérience permet donc de proposer pour la première fois des corrélats de la conscience visuelle chez le singe du point de vue des régions impliquées et des mécanismes sous-jacents.

3) Théorie de la perception consciente : Neuronal Global Workspace

Nos résultats semblent en accord avec la théorie développée par Dehaene et al en 1998 (Dehaene et al., 1998) et reprise par DelCul et al en 2007 (Del Cul et al., 2007). Dans cette étude ils utilisent un paradigme de masquage pour étudier les corrélats de la perception visuelle consciente chez l'homme afin de valider les prédictions de leur modèle. Pour ce faire, ils présentaient un stimulus cible flashé 16ms et suivi d'un masque avec une SOA variable (Stimulus Onset Asynchrony) de 16 à 100ms. Pour des SOA très faibles, le stimulus n'est pas consciemment perçu, alors que pour des SOA plus importantes il l'est. Ainsi ils cherchent à déterminer le temps de traitement nécessaire à un stimulus pour accéder à la conscience. Ils proposent la théorie du « Neuronal Global Workspace » qui stipule que pour être perçu consciemment un stimulus doit activer des aires corticales de haut niveau telles que les aires préfrontales et réactiver en feedback un réseau global distribué parmi des aires visuelles et non visuelles (figure 1.15).

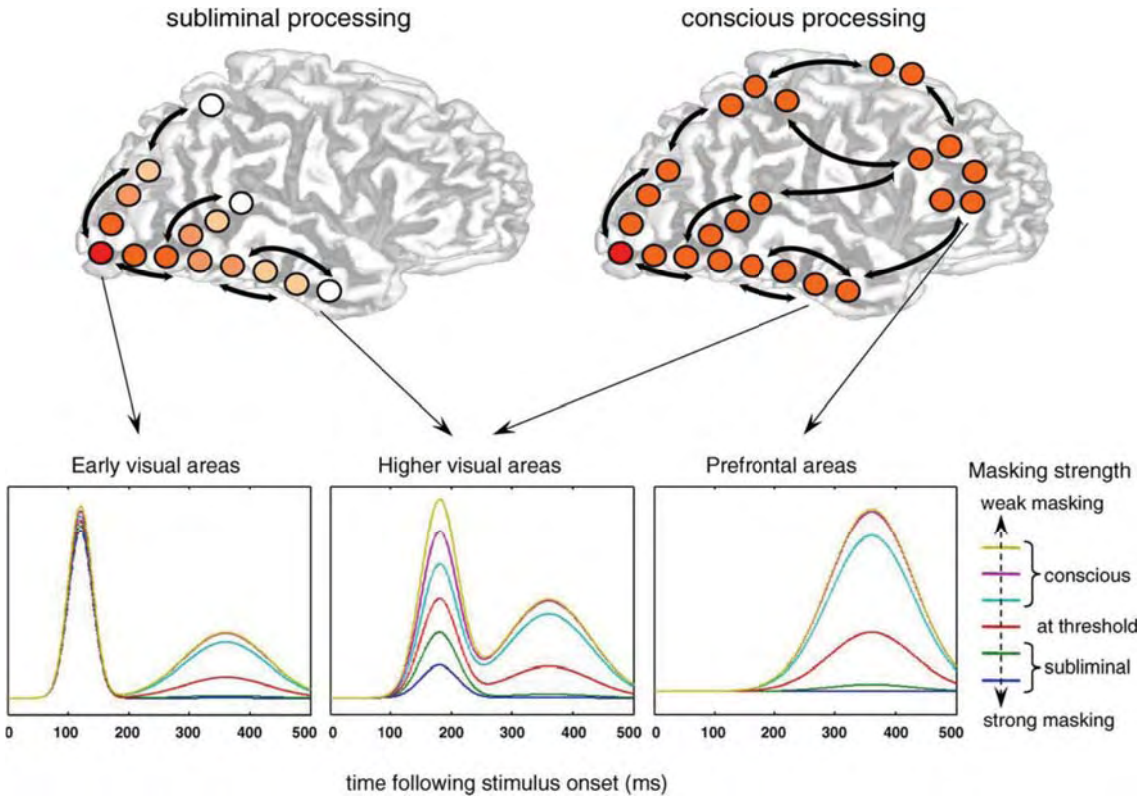


Figure 1.15 : Prédictions schématiques concernant les aires cérébrales activées par un stimulus en fonction de la puissance du masque. Plus le masque est faible (i.e. présenté avec une longue SOA), plus le stimulus peut activer des aires corticales de haut niveau (théorie du Neuronal Global Workspace). DelCul et al 2007

Les potentiels évoqués que nous avons obtenus pour les électrodes situées dans les aires visuelles (figures 1.11) paraissent tout à fait superposables avec les courbes théoriques pour les aires visuelles de bas et de haut niveau proposées par le modèle de DelCul et al.

De plus, nous avons également pu mettre en évidence par analyse en item unique une différence significative de l'activité tardive du cortex frontal (cf figure 1.16), soit environ 450 ms après l'onset sémantique. Cette activité peut être comparée à celle prédite par le modèle de DelCul et al dans les aires préfrontales, mais elle semble plus tardive (environ 400ms).

Le caractère distribué des activités corticales liées à la perception consciente observé dans notre étude semble donc supporter la théorie du « Global Neuronal Network » : afin d'être perçu consciemment, le traitement d'un stimulus ne doit pas se limiter aux seules aires visuelles, mais doit se propager dans des aires corticales moins spécialisées. Notre interprétation est donc que l'onde P2 représente l'activité de *global ignition*, « mise à

feu » de l'ensemble important de régions corticales supportant la conscience visuelle d'après la théorie de Dehaene et al (1998) et supporte donc la conscience visuelle.

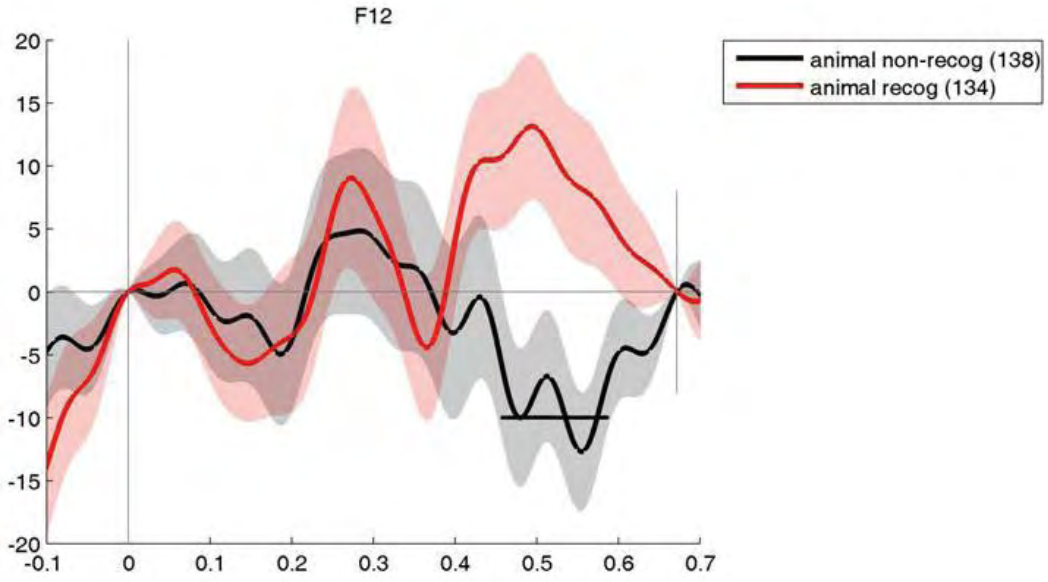


Figure 1.16 : Potentiels évoqués par des cibles nouvelles ou récentes dans deux conditions : reconnues ou non reconnues. Electrode placée dans le cortex frontal. La ligne noir indique une différence significative entre les 2 courbes (Test de Student, $p<0.05$)

Cette étude n'a pour le moment pu être menée que sur un seul singe, afin de confirmer nos résultats il faudrait pouvoir répliquer l'expérience avec un ou deux singes supplémentaires. De plus, bien que le singe ait pu apprendre à catégoriser correctement certaines images sur lesquelles il s'était trompé en premier lieu, il a montré un faible taux de correction. L'animal utilisé est un singe âgé, ce qui pourrait expliquer un manque de plasticité au niveau comportemental.

En conclusion, nos résultats montrent qu'à la suite de la première activité feedforward associée au traitement visuel du stimulus (Cauchoix et al., 2015), automatique et préconsciente, se produit ensuite une activité plus tardive largement distribuée dans les aires visuelles, associée à la perception consciente de son statut de cible. Cette activité est similaire à celle trouvée chez l'homme (Koenig-Robert and VanRullen, 2013) et coïncide avec les prédictions du Neuronal Global Workspace.

Chapitre 2: L'effet de congruence contextuelle sur la catégorisation d'objets chez l'Homme et le singe Macaque

Dans le précédent chapitre de cette thèse, nous avons mis en évidence un parallèle fort entre l'homme et le singe concernant les corrélats neuraux associés à la reconnaissance d'un objet. Nous avions alors égalisé au maximum les caractéristiques de bas niveau des stimuli pour nous concentrer sur le traitement de leurs aspects sémantiques. Dans ce second chapitre, notre démarche est toute autre. Si nous avons montré que le contenu sémantique des images conditionnait la réponse neurale chez le singe, nous ne pouvons exclure que les paramètres de bas niveau des stimuli puissent également jouer un rôle lors d'une tâche de catégorisation.

Nous sommes partis du constat que le contexte visuel pouvait influencer les performances des sujets (humains et singes) lors d'une tâche de catégorisation visuelle ultra rapide d'objet (Fize et al., 2011). Dans un tel exercice, les images sont flashées très brièvement, le système visuel doit alors extraire très rapidement les informations les plus pertinentes afin de déterminer si la cible de la tâche était ou non présente. Nous avons donc cherché à savoir quelle était la nature de ces caractéristiques pertinentes, et plus précisément si des caractéristiques de bas niveau telles que le spectre d'amplitude des images pouvait porter, au moins en partie, l'effet de congruence contextuelle observé dans les études précédentes, chez l'homme et le singe. Nous avons pour cela construit des images sans contenu sémantique mais auxquelles nous avons assigné des statistiques de scènes naturelles ou artificielles. Ces images ont ensuite été utilisées comme contextes lors d'une tâche de catégorisation visuelle.

Si dans le chapitre précédent nous montrions la similitude des signaux neuraux associés à la reconnaissance des objets chez l'homme et le singe, nous avons trouvé dans cette étude une divergence de stratégie entre ces deux espèces. Il semble ainsi que les singes

utilisent davantage les indices de bas niveau que les sujets humains. Nous ne pouvons que spéculer sur les raisons de cette divergence : est-elle due à une différence dans le système visuel, dans les stratégies cognitives, ou comme nous le pensons aux différentes expériences de vie des deux espèces. En effet nos singes ont toujours vécu en captivité, et ont donc une expérience limitée du monde, tandis que les sujets humains ont été confrontés à des situations bien plus diverses. De plus, les singes ont été longuement entraînés à des tâches de catégorisation visuelle, et ils ont pu avec le temps apprendre à se baser sur des indices de bas niveau (mais pas seulement !) pour répondre aux tâches qui leur étaient proposées.

Contextual congruency effect in natural scene categorization: Different strategies in humans and monkeys (*Macaca mulatta*)

Anne-Claire Collet [1,2]* , Denis Fize[1,2], Rufin VanRullen[1,2]

[1] Université de Toulouse; UPS; Centre de Recherche Cerveau et Cognition; Toulouse, France

[2] CNRS; CerCo; France

*Corresponding authors

E-mail : collet@cerco.ups-tlse.fr

Citation: Collet A-C, Fize D, VanRullen R (2015) Contextual Congruency Effect in Natural Scene Categorization: Different Strategies in Humans and Monkeys (*Macaca mulatta*). PLoS ONE 10(7): e0133721. doi:10.1371/journal.pone.0133721

Editor: Elsa Addessi, CNR, ITALY

Received: February 6, 2015; **Accepted:** July 1, 2015; **Published:** July 24, 2015

Abstract

Rapid visual categorization is a crucial ability for survival of many animal species, including monkeys and humans. In real conditions, objects (either animate or inanimate) are never isolated but embedded in a complex background made of multiple elements. It has been shown in humans and monkeys that the contextual background can either enhance or impair object categorization, depending on context/object congruency (for example, an animal in a natural vs. man-made environment). Moreover, a scene is not only a collection of objects; it also has global physical features (i.e phase and amplitude of Fourier spatial frequencies) which help define its gist. In our experiment, we aimed to explore and compare the contribution of the amplitude spectrum of scenes in the context-object congruency effect in monkeys and humans. We designed a rapid visual categorization task, Animal versus Non-Animal, using as contexts both real scenes photographs and noisy backgrounds built from the amplitude spectrum of real scenes but with randomized phase spectrum. We showed that even if the contextual congruency effect was comparable in both species when the context was a real scene, it differed when the foreground object was surrounded by a noisy background: in monkeys we found a similar congruency effect in both conditions, but in humans the congruency effect was absent (or even reversed) when the context was a noisy background.

Keywords: Object categorization, congruency effect, Fourier transform, amplitude spectrum

Introduction

Visual categorization appears to be a crucial ability for survival of many, especially diurnal, animal species, since they have to quickly adapt their behaviour to many different situations (to hunt a prey or escape from predators, for instance). However, because categorization implies abstraction, it was initially explored mainly in humans. In 1964 Herrnstein & Loveland [1] demonstrated for the first time a categorization capability in animals (pigeons). Following the pioneering work of Herrnstein & Loveland, other neuroscientists and behaviourists addressed the question of conceptualization and categorization in animals. The first studies investigated the existence of categories in animals at the subordinate or basic level. Thus, in 1984 Schrier et al [2] tested the existence of three concepts in macaques: “human”, “monkey” and” letter A”. They found clear evidence of transfer of the categorization rule to new items. In 1988, D’Amato & VanSant [3] also tested the existence of the concept “human” in monkeys; they showed that, if the picture training set is large enough, it allows generalization and thus, good transfer to new stimuli. But at that basic level of categorization, items within a category also share a lot of visual similarities, and abstraction abilities may not be necessary for correct performance. In 1988, Roberts and Mazmanian [4] explored the existence of a more abstract concept, “animal”, using different levels of categories in pigeons, non-human primates and humans. They showed that, after training, monkeys and pigeons could discriminate pictures well at all category levels (subordinate level: find one species of birds - the common kingfisher- among other birds; basic: find birds among other animals; super-ordinate: animals versus objects). However, they found weak categorical transfer to new stimuli in both species. Nonetheless, in that study the stimulus training set was limited (36 to 40 pairs of stimuli), and may not have been sufficient for optimal category learning, encouraging instead an association learning rather than a conceptualization. Then Fabre-Thorpe in 1998 [5(Fabre-Thorpe et al., 1998)] showed that non-human primates could categorize pictures according to the presence or absence of animals using a much larger set of stimuli (340), with a similar level of performance as

humans. Such evidence suggests that high level abstraction exists in primates, because different images of animals vary a lot in terms of physical features (body shape, size…).

Moreover, Sands et al in 1982 [6] showed that this ability could emerge even without specific training, using a comparison task in monkeys. The same observation was made by Astley and Wasserman (1992) in pigeons [7]: birds had to learn by heart to discriminate a set of positively reinforced pictures among others, and they tended to make more discrimination errors when these pictures were tested against others belonging to the same category.

Researchers then tried to assess the diagnostic features on which animals could rely to efficiently categorize pictures. First, animals have to infer the categorization rule by trial-and-error, and both pigeons [8,9 (Edwards and Honig, 1987, Aust and Huber, 2001)] and monkeys [3] appeared to better pick up relevant features for positively reinforced stimuli than for negative ones . On the other hand, the same teams showed that colours are not crucial features for categorization. Finally, while humans and monkeys can process stimuli in a global way [11], it was long held that pigeons only process visual information in a fragmented way, based on local features [9, 10], as their performance in categorization tasks can be quite robust against stimulus scrambling. However, this point of view was challenged by Wasserman and colleagues [12, 13]: they demonstrated that pigeons were sensitive to the spatial organization of objects. Furthermore the same team later found evidence that pigeons exhibited rotational invariance in object recognition, if trained with different viewpoints of the same stimulus [14], just as monkeys do [15]. Those results suggest that pigeons also can process visual stimuli in a global way.

Although many teams have worked on categorization, the neural mechanisms underlying this ability remain unclear. Nevertheless Kriegeskorte et al in 2008 [16] compared activation patterns in inferior temporal cortex in both humans and monkeys during a passive fixation task of isolated items. Despite the use of two different techniques (fMRI in humans and single cell recording in monkeys), they observed strong similarities

between humans and monkeys in the way items are grouped into categories in IT (items belonging to the same category elicited activation patterns close to each other).

In the real world, however, objects are rarely isolated but surrounded by a complex environment with multiple elements. In difficult situations our a priori knowledge of a context could help detect and recognize a target object. Contextual information can thus affect the efficiency of object recognition [17,18]). Behavioural studies in humans have explored the importance of context in facilitation and enhancement of visual processing and perceptual memory. They demonstrated that, by manipulating context, it is even possible to create false memories [19] or to predispose subjects to false recognition of objects [20].

In the studies cited above, visual exposure to contexts was long enough to allow memorization or precise exploration. But Joubert et al in 2007 [21] showed that human subjects are equally fast at categorizing a scene at a super-ordinate level (for instance man-made scenes versus natural scenes) as an object (e.g. animals versus objects). However, it takes longer to categorize scenes at the basic level, for example sea, mountain, indoor or urban scenes [22], suggesting a coarse to fine processing of such stimuli. The same super-ordinate level advantage was found in object categorization tasks: it is faster to recognize a bird as an animal than as a bird [23]. Moreover, Joubert et al (2008) [24] showed that the congruency effect between a scene and an object appears very early in an ultra-rapid categorization task where stimuli are flashed very briefly, which suggests a parallel processing of both context and object information. Davenport & Potter [25] also came to this conclusion, by comparing scene and object categorizations in humans, either with mixed or isolated stimuli. They found better performance when foreground object and background were congruent, whatever the task, i.e. whether subjects had to categorize only the foreground object, only the background, or both. Moreover, a congruency effect between context and object has also been observed in monkeys during a rapid visual categorization task [26]: task performance (categorization of the foreground object as an animal/non-animal, where only stimuli containing an

animal were the targets of the task) was affected both in terms of accuracy and reaction times, and in the same way in humans and monkeys. These results suggest that the brain mechanisms underlying the congruency effect may be similar in the two species.

The above-mentioned studies explored the time course of categorization and contextual influences, but did not directly probe the visual features of the scene images that could support these cognitive abilities. Indeed, a scene is not only a collection of objects; it also has a spatial arrangement, and global physical features which can contribute to defining its gist. Algorithms can efficiently categorize scenes on the basis of such global statistics, especially the shape of the spatial envelop [27]. This envelope corresponds to the outlines of elements of a scene which define its three-dimensional layout (like walls, buildings, mountain slopes, landscape relief etc) as well as their relations. In addition, after learning the power spectra of many different images, other algorithms were able to detect animals in different contextual scenes significantly above chance [28]. This leads to the question of what type of physical information is relevant for the human and primate visual systems to access context category. Joubert et al. in 2009 [29] designed a context categorization task in which they first equalized the amplitude spectra of their stimuli, and then destroyed the phase spectrum of the previously equalized pictures. They demonstrated that although amplitude equalization slightly impairs categorization performance in terms of reaction times, the most important diagnostic criterion relies on phase information. Loschky et al [30] also found that unlocalized amplitude information was insufficient to categorize scenes at the basic level in humans. On the other hand, Guyader et al [31] found that processing of scene category was influenced by amplitude spectrum information. They used a priming paradigm in a scene categorization task, and showed that chimeric primes built from an amplitude spectrum of scenes of the same category as the target image could speed up the categorization. Therefore, the relative contributions of these two global statistical dimensions (i.e. phase and amplitude) to the categorical processing of an image remain unclear.

In our experiment we aimed to explore the contribution of the Fourier amplitude spectrum to context-object congruency effects in humans and monkeys. To our knowledge, all previous studies comparing visual categorization in humans and monkeys concluded that processing of objects and scenes were very similar in both species.

We designed a protocol where images of isolated objects or animals were pasted either on real scene photographs (man-made or natural) or on noisy backgrounds built from Fourier transforms of 100 real scenes averaged in amplitude (and phase randomized) with a varying proportion of naturalness. The expected congruency effect was at the super-ordinate category level (i.e. natural scenes are congruent with animals and man-made scenes with objects). If at least part of this effect relies on power spectrum information, then we should observe better performances for congruent stimuli, i.e. animals pasted on a noisy background with a high degree of naturalness or objects pasted on a background with a low degree of naturalness, than for incongruent ones.

Material and Methods

Subjects:

Non-human primates:

Two male rhesus monkeys (Rx and Dy, both 20 years old) performed a rapid visual go/no-go categorization task: Animal versus Non-Animal. Both monkeys were previously trained for similar tasks. They both took part in the experiments of Fize et al, 2011 [26] and in the study of Fabre-Thorpe, 1998 [5], where monkey Dy performed a food/non-food categorization task (target category: food) and monkey Rx an animal/non-animal task (target category: animal). They were born in captivity, raised with congeners but no other species of animals or non-human primates. However, even though they had never been exposed to the animal species from the task in real life, they had previously watched wildlife documentaries during the early stage of their training for the experiment of Fabre-Thorpe, 1998 [5]. This could have helped them to form the natural category “animal”.

All procedures conformed to French and European standards concerning the use of experimental animals; all protocols used in this study (visual stimulation, training and experimental protocols, reward system, animal care) were approved by the regional ethical committee for experimentation on animals (agreement ref. MP/05/05/01/05, C2EA-14 ethical committee of Marseille). The agreement was delivered for a larger project on visual context processing in monkeys; this study was part of the project.

During the experimental period, animals are weighed every day before the session. If a loss of body mass larger than 5% is detected, dietary restriction is controlled in order to stabilize the animal's weight. If a loss of body mass of more than 10% is detected, the experiment is stopped until the animal has recovered its initial weight. A veterinarian examines each animal once a year, performing blood tests to check for SIV, hepatitis A, HSV, STLV and also performing faecal cultures. The laboratory includes an elected committee responsible for animal well-being. This committee ensures that all lab members working with animals abide by the procedures in accordance with French and European ethical standards.

Animals are housed in indoor enclosures (surface of 12m², height of 4 meters), which they share with 2 or 3 congeners of the same sex. The enclosures include perches of different heights and different adapted toys (balls, slides, swings). Each animal also has an individual cage (0.75m x 1m x 2m) that can be accessed via a tunnel from the enclosure. Monkeys are fed and watered in these cages where they spend about 6 hours every day. They are given dry macaque food (50g/kg of body weight), three fresh fruits and one vegetable. During the experimental period fruits are provided after the experimental session (regardless of performance during the session), water is also provided at the same time in an appropriate amount to complement what the animal drunk during the session. During weekends (when no experimental sessions are carried out) animals are on a normal diet including water, dry food, vegetables and fruits, provided at the same time as the other animals and in the same quantity as mentioned above. On weekends a single person is in charge of animal care, and monkeys spend only 3 hours in their cages.

Monkeys did not suffer from the experiment (since it was a psychophysical experiment, where neither surgery nor drug injection was required). Both monkeys were retired after this experiment but remained in the animal facility of the institute. One died a few months later from natural causes.

Humans:

Fifteen human subjects performed the same go/no-go visual categorisation task with the same experimental device (7 men, mean age: 27 range 22-37, 2 of them left-handed). They all had normal or corrected to normal vision. All participants gave written informed consent prior to taking part in the experiment. The study was approved by the local ethics committee ‘‘CPP Sud-Ouest et Outre-Mer I’’ under protocol number 2009-A01087-50.

Stimuli:

Vignettes

We used objects (tools, furniture, vehicles) and animals (mammals, birds, reptiles, insects, fishes) vignettes (from Hemera Photo Objects library, tiff format). Monkey Rx saw 400 vignettes, Monkey Dy 320 and humans 260. For all subjects half of the vignettes were animals, whereas half were objects. The largest dimension of each vignette was scaled to 100 pixels (6.2° of visual angle), regardless of its identity.

Backgrounds

Vignettes were randomly pasted on a background: either a photograph of a real scene (either a man-made or a natural scene, 100 exemplars of each, Real scene condition) or on a noisy background with a varying level of naturalness (from 0% to 100% by steps of 10%, Noisy background condition).

To build these noisy backgrounds we averaged Fourier amplitude spectra of 100 real photographs, randomized the corresponding phase spectrum, and then reconstructed an

image by an inverse Fourier transform. In order to create the 11 levels of naturalness, the proportion of natural and man-made scenes among the 100 photographs varied: for example, for level 20% we used 20 photographs of natural scenes and 80 of man-made scenes. Ten noisy backgrounds of each level were built from randomly chosen photographs. The photographs used to build noisy backgrounds were the same as those used in the Real scene condition.

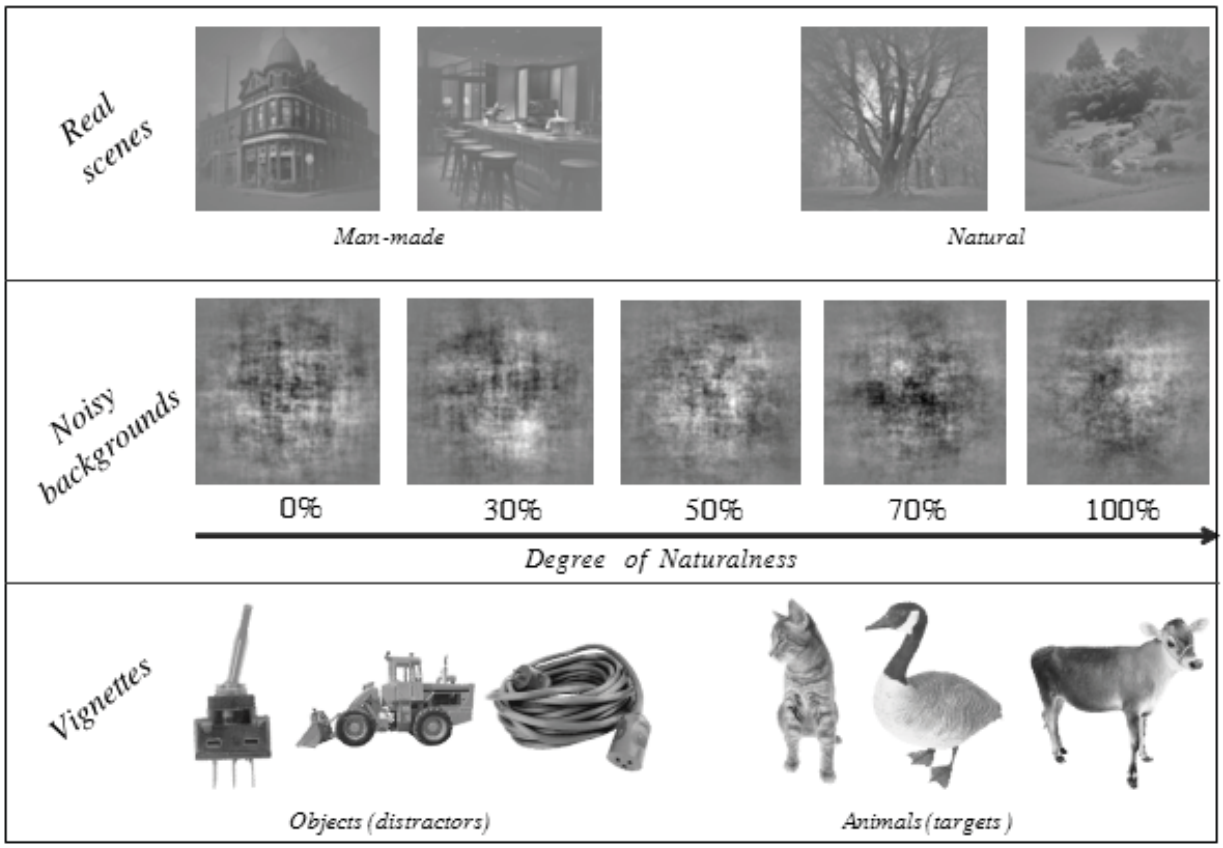


Figure 2.1: first line: examples of real scenes; second line: examples of noisy backgrounds with an increasing degree of naturalness; third line: examples of vignettes

Both backgrounds and vignettes were in greyscale (Figure 2.1), equalized in contrast and luminance. Stimuli were displayed on a grey screen (resolution 600x800 pixels). Backgrounds (size 600x600 pixels) had their borders smoothed using a Gaussian filter. Vignettes were pasted randomly in one out of nine possible positions (centre of the screen or on a circle, 3.2 degrees of eccentricity). By randomizing vignette locations and equalizing their sizes, we sacrificed the spatial and scale coherence of stimuli (e.g. a large mouse could be pasted in the middle of the sky, or a small elephant on a table top).

To maximise the influence of contextual effects, the transparency of vignettes was adjusted online based on subject's performance using a stair-case function applied to the alpha layer. We fixed the threshold at 70% of accuracy on targets, regardless of trial condition (noisy or real backgrounds). Transparency values were confined between 0 (fully visible) and 0.6.

In the "Real scene" condition, we considered targets as congruent with a natural scene and distractors with a man-made scene. The two other associations were labelled as incongruent (Figure 2).

In the "Noisy background" condition, the 11 levels of naturalness allowed us to build 11 levels of congruency: for example, a noisy background of 20% naturalness corresponds to a 20% level of congruence for targets and an 80% level for distractors (Figure 2.2).

Subsequently, data were analysed according to the congruency level.

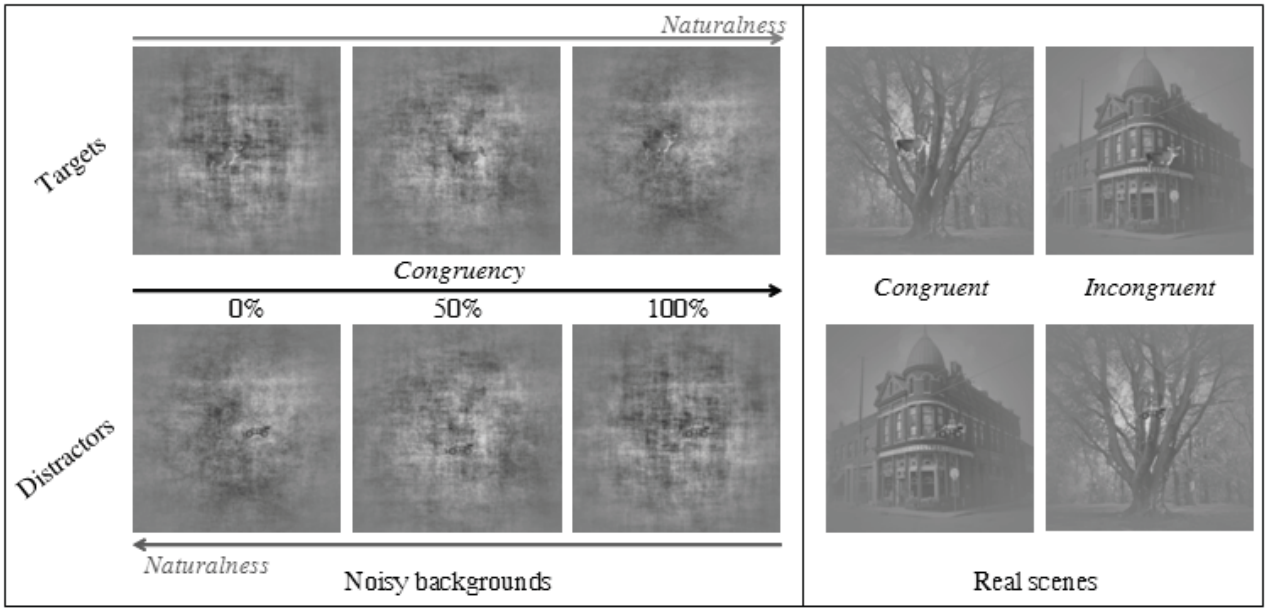


Figure 2.2: examples of final stimuli used in the experiment (a cow as target, a tractor as distractor). The congruence level increases with the degree of naturalness for targets, but decreases for distractors.

This figure also illustrates the size ratio between vignette and background, which makes the task difficult.

Procedure

Subjects (both monkeys and humans) performed a rapid visual go/no-go categorization task. They sat in a dimly lit room 50cm away from a computer screen (resolution 600x800 pixels, vertical refresh of 60 Hz). Stimulus display and behavioural response measurement were carried out using the Psychtoolbox of Matlab 7.6.0 (R2008a).

Each trial started with a fixation point (4x4 pixels) displayed at the centre of a grey screen for 800ms, immediately followed by the stimulus for 3 frames (50ms). To start the experiment, subjects placed their fingers (preferred hand) on a response pad equipped with infrared electrodes that provided timing information with millisecond precision. For each stimulus containing a target (i.e. an animal), subjects had to release the pad within 800ms; longer reaction times were considered as no-go responses. If the stimulus did not contain an animal but an object, subjects had to keep their fingers on the pad, and so inhibit any motor response.

When subjects made a mistake, the corresponding stimulus reappeared for 1500ms (negative feedback). There was no positive feedback for humans, but monkeys were rewarded with a drop of juice following each correct response (either go or no-go).

Subjects could suspend the experiment whenever they wanted, by releasing the pad. All conditions were randomly mixed.

Monkeys worked every day ad libitum during 5 weeks (Monkey Rx performed 16800 trials, and monkey Dy 18750 trials); reward was gradually increased during a session to maintain the motivation level. In order to avoid the possibility of learning a specific set of stimuli, 10 new vignettes were added every day, while the set of backgrounds remained constant. The association vignette/background was randomized.

For humans, a session was composed of 1300 trials. All trial types were equiprobable (i.e. 100 trials per congruency level in each condition). In total, the group of human subjects performed approximately the same number of trials as each monkey.

Evaluation of performance

Performance was recorded in terms of accuracy and response speed. We also computed minimal RT: we first calculated cumulative hit rate and false alarm rate across time for each subject using bins of 10ms; the minimal RT corresponds to the center of the first bin where the difference between false alarms and hits rates is significant (χ^2 test, df=1, α=0.05).We performed statistical analyses using Matlab 7.6.0 (R2008a). The total number of trials across all human subjects was comparable with the number of trials from each monkey (15 human subjects, 1300 trials each, 19500 trials in total). We thus analysed data of each monkey separately, but pooled data of all human subjects. Before pooling, we applied a vincentisation procedure (see Results for more details) to normalise reaction times across subjects. Without this normalization, the fastest human subjects would have contributed disproportionately to the early part of the cumulated d' calculation and to the latency computation.

We compared conditions within each monkey and within the pool of human subjects. To assess differences between conditions (congruent vs incongruent, and "noisy background" vs "real scene") we calculated cumulated d-prime across time in each condition and performed permutation tests to determine whether the observed difference was significant. Statistical and mathematical calculations are detailed in the results.

Results

Global performance

Subjects, both monkeys and humans, performed the categorization task significantly above chance (Monkey Dy: 75.9% correct, Monkey Rx: 87.9% and humans: 78.4%, see Table 2.1 for more details). Moreover, as previously shown in other studies (Fize et al, 2011 [26]) both monkeys and humans were able to recognize animals regardless of the

nature of the background (real scene, either natural or man-made, or noisy background). They made their decision based on the vignette identity and ignored the background.

Table 2.1 : Global performance of subjects in the two conditions

	Accuracy				Reaction times					
	Noisy Backgrounds		Real Scenes		Noisy Backgrounds			Real Scenes		
	Target	Distractors	Targets	Distractors	Minimum RT	Median RT	Mean RT	Minimum RT	Median RT	Mean RT
Monkey Dy	72%	71%	70%	88%	260ms	373ms	393ms	280ms	403ms	418ms
Monkey Rx	85%	93%	70%	93%	270ms	392ms	410ms	290ms	427ms	440ms
Humans	75%	83%	68%	87%	350ms	528ms	535ms	370ms	544ms	548ms

"Real scene" Condition

In this condition, a vignette appeared on a real scene background. In congruent trials either an animal was pasted on a natural scene or an object on a man-made scene. In incongruent trials an animal was associated with a man-made scene and an object with a natural scene. Behavioral performance is summarized in Table 2.2.

Table 2.2: Go-response rate (hits and false alarms) as well as sensitivity (d') in the "real scene" condition

	Congruent			Incongruent			Congruency effect	
	Hit rate	False alarm rate	d'	Hit rate	False alarm rate	d'	Difference	p-value
Monkey Dy	77.0%	23.8%	1.45	66.2%	34.4%	0.82	0.63	<0.001
Monkey Rx	72.2%	4.4%	2.29	68.6%	8.5%	1.86	0.43	<0.001
Humans	70.6%	11.3%	1.75	66.4%	14.3%	1.49	0.26	<0.01

Both monkeys and humans had a higher hit rate (animals correctly categorized) and a lower false alarm rate (objects incorrectly categorized) in congruent trials than in incongruent trials.

We then calculated d' (a measure of the signal detection theory) to quantify the congruency effect between the vignette and the background across response time (absolute time for both monkeys and relative time scale for humans, obtained after vincentisation).

Vincentisation

We used a vincentisation procedure to normalize reaction time distributions across subjects before pooling their data because individual variations were not negligible (median reaction times going from 478ms for the fastest subject to 584ms for the slowest, and standard deviation going from 65ms to 97ms). This method has the advantage of keeping the overall shape of the reaction time distribution while avoiding an overpowering influence of any individual participant.

We divided each subject's global reaction time distribution into 20 classes (or quantiles) of equal duration, regardless of condition, trial or response type (i.e. respectively: "Real scene" or "Noisy background" condition, congruency level of trial, hit or false alarm). We then assigned a quantile number to each trial of each subject. Finally we pooled data of all subjects within the 20 quantiles.

Cumulated d'

The d' is a sensitivity index which quantifies the ability to discern a meaningful stimulus (signal) from others (noise). It is calculated using the following formula: $d'=z(\text{hit rate})-z(\text{false alarm rate})$ where z is the inverse of the cumulative normal distribution.

The temporal evolution of d' after stimulus onset can inform us about the time course of congruency effects. To calculate cumulated d' as a function of response time (that is, the d' based on all responses given before a particular time), we first calculated cumulated hit rates and false alarm rates as a function of response time, using constant bins of actual reaction times in monkeys and quantiles of reaction time in humans. For instance, cumulated hit rate for the q^{th} bin (or quantile) in a given condition "x" was obtained by:

$$HitRate_x(q) = \frac{\sum_{bin=1}^{q} nbHits(RT \in bin)_x}{nbT\,\text{arg}\,ets_x}$$

We finally applied the d' formula to each bin or quantile and plotted the cumulated d' curves as shown in figures 3 and 4.

Permutation test

Permutation tests are a non-parametric approach to establishing the null distribution (expected distribution under the null hypothesis) of an experimental observation [32]. In our study, this observation is the value of the d-prime difference between the two conditions, congruent vs. incongruent. To establish the null distribution, we performed 1000 random permutations of the assignment of congruence labels (congruent/incongruent) to each trial. For each of the 1000 randomized data sets, we computed cumulated d' for congruent and incongruent conditions and calculated the difference between the two conditions, exactly as done with the original experimental dataset. We thus obtained a distribution of theoretical differences under the null hypothesis at each reaction time bin (for monkeys) or quantile (for humans), and compared our observed difference values to this distribution. The ranking of the observed value among the theoretical distribution provides the p-value. If our observed difference value was higher than the last percentile of the null distribution, we considered it as statistically significant.

Both humans and monkeys discriminated animals from objects better on congruent backgrounds than on incongruent ones (Figure 3). This d' difference reached statistical significance quite early (350ms after image onset for monkey Dy, 470ms for monkey Rx, and from the fourth quantile of time, i.e. 445ms SEM +/- 14ms in humans; permutation test, $p<0.01$), and remained significant until the end of the response window (limited to 800ms).

In short, humans and monkeys showed a similar contextual congruency influence in the categorization task with "Real scene" backgrounds.

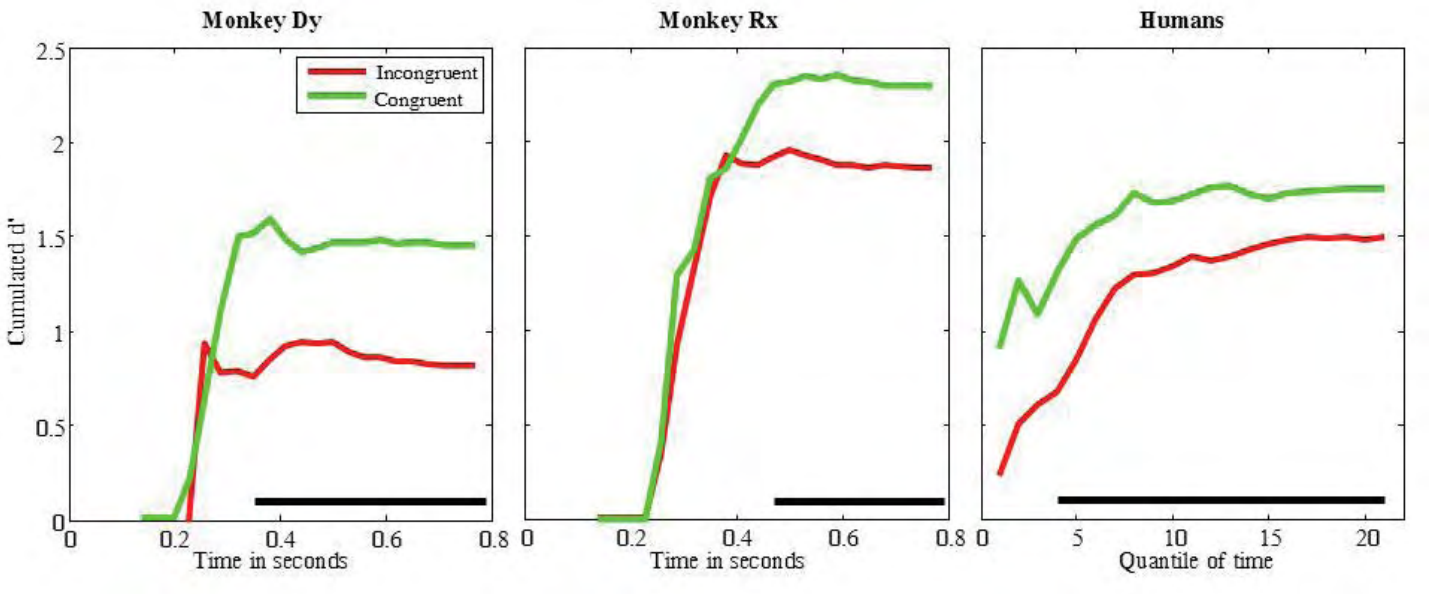


Figure 2.3: Cumulated D' as a function of time in the "Real scene" condition (bins of 30ms in monkeys, quantiles of relative time in humans: mean quantile duration =20ms SEM +/0.6ms, first quantile starts at 370ms SEM +/- 11ms). The horizontal black line indicates the significance of the difference between congruent and incongruent trials (permutation test, 1000 permutations, $\alpha<0.01$)

"Noisy Background" condition

In this condition, each background was constructed from a mixture of Fourier amplitude spectra from 100 images with a variable proportion of natural scenes and man-made scenes, and a randomized phase spectrum. Thus, the backgrounds looked like meaningless noise, but retained certain global statistical features of natural or man-made scenes (see Figure 2). Background/vignette congruency was gradually manipulated from 0% (incongruent) to 100% (congruent) by steps of 10%. In Table 3, results for the extreme conditions (0% and 100% i.e. incongruent and congruent) are summarized.

Table 2.3: Go-response rate (hits and false alarms) as well as sensitivity (d') in the "Noisy background" condition

	100% Congruent			0% Congruent			Congruency effect	
	Hit rate	False alarm rate	d'	Hit rate	False alarm rate	d'	Difference	p-value
Monkey Dy	73.8%	10.5%	1.89	64.3%	14.7%	1.41	0.48	<0.001
Monkey Rx	87.2%	4.3%	2.85	84.5%	10.1%	2.29	0.56	<0.001

Humans	73.7%	16.0%	1.58	77.2%	17.1%	1.74	-0.15	NS

While the results of both monkeys follow the same trend as in the real scene condition (higher hit rate and lower false alarm rate, higher d' in congruent trials than in incongruent trials), it is not the case for human subjects, since their hit rate is higher and their d' lower in incongruent trials.

When directly comparing d' across extreme congruency levels, we found a significant difference ($p<0.001$) in both monkeys, in the same direction as in the real scene condition, i.e. d' was significantly higher in congruent trials than in incongruent trials. However, in humans, there was no significant difference between d' obtained for the two extreme congruency levels.

Finally, we computed the cumulated d' with the same method as the one used in the "real scene" condition. This was done not only for the two extreme congruency levels (0% and 100%), but also for the 9 other intermediate levels. Results are shown in figure 2.4.

Figure 2.4 shows the establishment of the congruency effect across time. In monkeys, the difference between extreme conditions (illustrated by horizontal black lines in Figure 2.4) appeared about 60ms earlier than in the real scene condition (from 290ms after stimulus onset for monkey Dy, and 410ms for monkey Rx) and remained stable across time. Moreover, the d' values of the 9 intermediate congruency levels mostly fell between the 2 extreme levels. We computed linear regressions of total d-prime as a function of congruency level. For both monkeys, there was a significant positive correlation between congruency level and d'.

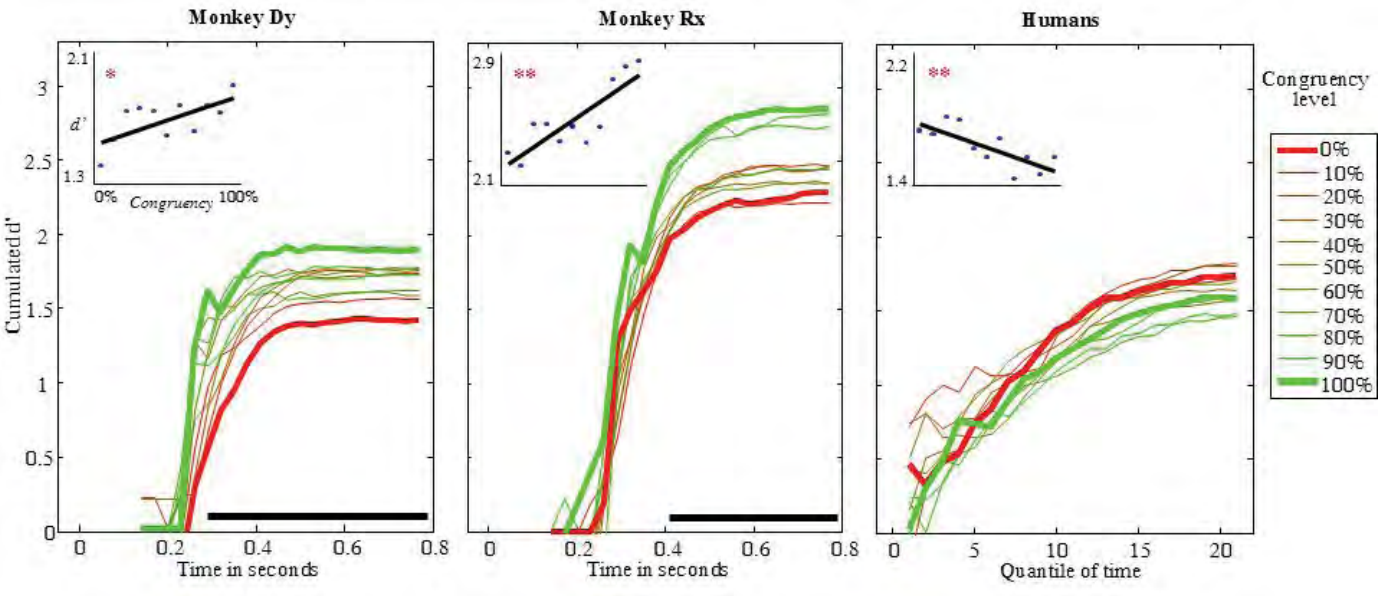


Figure 2.4: Cumulated d' as a function of time in the "Noisy Background" condition (bins of 30ms in monkeys, quantiles of relative time in humans: mean quantile duration =20ms SEM +/0.6ms, first quantile starts at 370ms SEM +/- 11ms). The thick red and green lines represent 0% and 100% congruency levels (respectively), while intermediate levels are depicted with thin lines. The horizontal black line indicates the significance of the difference between extreme congruency level trials (permutation test, 1000 permutations, $\alpha<0.01$). Inserts on top left represent linear regressions of total cumulated d' as a function of congruency level. Blue dots are observed values and black lines are the linear regression. Slopes, R^2 and p-values are respectively $s_{Dy}=0.026$, $R^2_{Dy}=0.45$ $p_{Dy}=0.024$, $s_{Rx}=0.054$, $R^2_{Rx}=0.72$, $p_{Rx}=0.001$, $s_H=-0.027$, $R^2_H=0.58$, $p_H=0.006$.

In humans, no significant difference between extreme congruency conditions was visible at any time point. From the tenth quantile onwards, d' in the incongruent condition was actually slightly higher than in the congruent condition. Computing linear regression, we observed a significant negative influence of the congruency level on subjects' performance (d').

Difference between conditions

To statistically evaluate the difference between congruency effects observed in the two conditions (real vs. noisy backgrounds), we performed random permutations of the assignment of trial condition. For each randomized data set we computed global d', we calculated d' difference (congruent vs. incongruent) within trial conditions, and finally we obtained a value of difference between conditions. We obtained a distribution of

theoretical differences according to the null hypothesis (i.e. the two trial conditions do not lead to different congruency effects) and compared our observed value to this distribution. The null hypothesis could not be rejected for either monkey (Monkey Dy: observed difference=0.15, p-value=0.23, Monkey Rx observed difference=-0.13, p-value=0.28, 1000 permutations); in other words, the congruency effect was not different in the 2 trial conditions. In humans, on the contrary, the congruency effect was significantly different in the two conditions (observed difference=0.42 p-value=0.013).

Altogether, these results imply that humans exhibited a contextual congruency effect only in the condition where the background was a real scene photograph; they did not process noisy backgrounds as real contexts. Monkeys, on the other hand, showed a similar contextual congruency effect than humans for real backgrounds, but this effect remained unchanged with noisy backgrounds.

Discussion

The main goal of this study was to determine whether or not the amplitude spectrum of a scene could support, at least in part, the contextual congruency effect found in several categorization studies, in humans and monkeys. To answer this question we designed an ultra-rapid categorization task using two distinct trial types: either the context was a real scene photograph or a noisy background built from the average of 100 scenes' amplitude spectra and random phase.

In the "Real Scene" condition, both humans and monkeys exhibited a similar congruency effect: they performed significantly better in categorizing vignettes when they were pasted on a congruent context than when they were pasted on an incongruent one. This result confirms those previously obtained by Fize et al, 2011 [26] with a similar task, and our observed congruency effect also has a similar magnitude. But, interestingly in our experiment we left aside the coherence of the object location within the scene (for instance a cow could appear in the sky of a landscape) and the object scale coherence (a bee could be as big as a rock) because vignette localization and context/vignette associations were randomized, and vignette sizes were equalized. In previous studies these two parameters were controlled. The collage of the vignette on the background was made in advance by the experimenter in order to create coherent stimuli in terms of

scale and location, whereas in our experiment, it was performed automatically and randomly by the stimulation program. Our experiment thus indicates that the congruency effect exists at a superordinate category level, and is not or sparsely affected by spatial and scale coherence.

However, because in our task we did not counterbalance the target category, we cannot generalize this congruency effect obtained with animals as targets and man-made objects as distractors to the opposite categorization task, i.e. objects as targets and animals as distractors. This caution in generalizing our conclusions is all the more warranted since some studies indicate that living things are sometimes treated as a special visual category [33].

In the "Noisy Background" condition the backgrounds had amplitude spectra comparable to those of natural or man-made scenes (or various levels in-between). Because these amplitude spectra were obtained by averaging over 100 scenes from the same superordinate category (natural vs. man-made) but from various basic-level categories (e.g. street and indoor, sea and mountain, etc.), we can assume that the amplitude component was properly isolated from other potential physical features, including spatial layout which could be typical of scenes at the basic level [27].

Surprisingly we obtained different results in humans and monkeys: monkeys exhibited a similar congruency effect in the Noisy background condition as in the Real Scene condition while humans seemed to process noisy backgrounds in a different way than real scenes.

Monkeys better discriminated an animal from an object when the noisy background was built from amplitude spectra average of real scenes from the congruent category (man-made scenes for objects and natural scenes for animals) than on backgrounds built from incongruent category pictures. Interestingly, d' congruency effects in the "Noisy Background" condition were not significantly different from those obtained in the "Real Scene" condition. So it seems that the congruency effect observed in monkeys relies in large part on image amplitude spectrum at the superordinate level. Moreover, regressing monkeys' performance on the proportion of naturalness in these noisy backgrounds

suggested that this effect can be progressively modulated, i.e. the congruency effect was proportional to the ratio of congruent over incongruent amplitude spectrum information.

On the contrary, humans' performance was affected in a different way in this second condition. First we found no significant d' difference between extreme congruency levels, i.e. an absence of congruency effect, although the same observers had displayed significantly positive congruency effects for real scene backgrounds. Further a linear regression taking into account the intermediate congruency levels showed that noisy backgrounds could actually give rise to a negative congruency effect: subjects were better able to discriminate animals from objects when the noisy backgrounds contained more incongruent physical features. The congruency effect obtained in the "Real Scene" condition thus cannot be explained by the amplitude spectrum of the scenes. Since there was no semantic information in the background (i.e. nothing to recognize or categorize), we might suggest that subjects processed it as a texture. Consequently, the negative influence of congruency on performance could be due to a type of camouflage effect: it might be easier to distinguish a curvilinear shape (i.e. an animal) on a background containing a lot of straight lines (i.e. features of a man-made scene) and conversely. This phenomenon has indeed been examined in visual search studies [34,35]; they concluded that it takes longer to find a target on a complex background whose features are similar to those of the target. In our experiment, presentation time was constant and very brief (50ms) and response time was limited (800ms), so this predicted decreased performance for background-congruent trials might be revealed as a lower d' rather than an increased RT.

A congruency effect, positive or negative, implies three different mechanisms: object processing, scene processing and interactions between scene and object. Object processing has been extensively explored in both humans and monkeys [16, 36]; since objects were not manipulated in our experimental design, our results do not challenge those previously obtained for object categorization. To our knowledge, scene processing has been mainly studied in humans, in terms of category relevant features [27-30]. Moreover, background-object interactions and congruency effects have rarely been investigated in terms of physical features of the scene stimuli. Our study brought to light

that there might be a difference between humans and monkeys in background processing, object-background interaction or both.

We can hypothesize that the observed difference between humans and monkeys may not be due to a difference in visual processing in itself, but rather to different visual experiences of the world. Congruency effects in humans would rely on associations between the to-be-categorized item and contextual elements which usually co-occur in the real world [17, 20, 37]**;** but because our monkeys were born and have always lived in captivity, they have very little experience of such real world co-occurrences; their visual system may thus need to analyze photographs of scenes as a collection of physical features. This notion raises the question of what representation monkeys may have of the scenes in our experiment. Many studies aimed to explore understanding of pictures by non-human primates. There are three possible ways of reading a picture, as defined in Fagot et al, 2000 [38]: the first one is the *independence* mode: the monkey does not link a picture to the real object; the second one is the *confusion* mode: the monkey confuses the picture with the real object; and the third is the *equivalence* mode: the monkey recognizes a picture as a representation of the real object. To determine which mode is used by monkeys, Truppa et al in 2009 [39] designed a Matching to Sample protocol with capuchin monkeys, and observed that these New World monkeys were able to match a real object with its photograph using the equivalence mode. In 2008, Parron et al. [40] tested this ability in pictorially naïve baboons, gorillas and chimpanzees. They first trained the animals to grasp a piece of banana presented against a pebble, and then tested them using stimuli pairs (a real object and a photograph or two photographs). Animals never mistook a real piece of banana for its photograph, but when a photograph of banana was presented against a pebble, both baboons and gorillas grasped it and tried to eat it, which suggests that they used the *confusion* mode. Lastly Pokorny and de Waal proved in 2009 [41] that capuchin monkeys were able to recognize photographs of group mate faces in an *equivalence* mode. Fagot et al. 2010 [42] suggest that recognition of a photograph by a monkey is a dynamic learning process. A monkey frequently exposed to photographs of well-known objects can learn to distinguish them. But on the other hand, it is possible that repeated exposure to pictures leads monkeys to adopt the independence mode. This could be the case in our study because 1) our monkeys were exposed to photographs for

years, and 2) they never faced in real life most of the animals or objects we showed. In that case, monkeys would just read pictures as a collection of features, and learn to categorize items based on certain features without associating any meaning to the picture itself.

Acknowledgements

We would like to thank Camille Lejards, Emilie Rapha and Grégory Marsal for animal care and help in experiment organization.

Bibliography

1- Herrnstein RJ, Loveland DH (1964) Complex visual concept in the pigeon. Science 146(3643): 549-551.

2- Schrier AM, Angarella R, Povar ML (1984) Studies of concept formation by stumptailed monkeys: Concepts humans, monkeys, and letter A. J Exp Psychol Anim Behav Process 10(4): 564-584.

3- D'Amato MR, Van Sant P (1988) The person concept in monkeys (Cebus apella). J Exp Psychol Anim Behav Process 14(1): 43-55.

4- Roberts WA, Mazmanian DS (1988) Concept learning at different levels of abstraction by pigeons, monkeys, and people. J Exp Psychol Anim Behav Process 14(3): 247-260.

5- Fabre-Thorpe M, Richard G, Thorpe SJ (1998) Rapid categorization of natural images by rhesus monkeys. Neuroreport 9(2): 303-308.

6- Sands SF, Lincoln CE, Wright AA (1982) Pictorial similarity judgments and the organization of visual memory in the rhesus monkey. J Exp Psychol Gen 111(4): 369-389.

7- Astley SL, Wasserman EA (1992) Categorical discrimination and generalization in pigeons: All negative stimuli are not created equal. J Exp Psychol 18(2): 193-207

8- Edwards CA, Honig WK (1987) Memorization and "feature selection" in the acquisition of natural concepts in pigeons. Learning and motivation(18): 235-260.

9- Aust U, Huber L (2001) The role of item- and category-specific information in the discrimination of people- vs. nonpeople images by pigeons. Anim Learn Behav 29(2): 107-119.10- Cerella J(1980) The pigeon's analysis of pictures. Pattern Recog 12:1-6

11- Schrier AM, Brady PM (1987) Categorization of natural stimuli by monkeys (Macaca mulatta): effects of stimulus set size and modification of exemplars. J Exp Psychol Anim Behav Process 13(2): 136-143.

12- Wasserman EA, Kirkpatrick-Steger K, Van Hamme LJ, Biederman I (1993) Pigeons are sensitive tot he spatial organisation of complex visual stimuli. Psychological Science 4(5):336-341

13- Kirkpatrick-Steger K, Wasserman EA, Biederman I (1996) Effects of spatial rearrangement of object components on picture recognition in pigeons. J Exp Ana of Behav 65: 465-475

14- Wasserman EA, Gagliardi JL, Cook BR, Kirkpatrick-Steger K, Astley SL, Biederman I (1996) The pigeon's recognition of drawings of depth-rotated stimuli. J Exp Psych 22(2):205-221

15- Logothetis NK, Pauls J, Bülthoff HH, Poggio T (1994) View-dependent object recognition by monkeys. Curr Biol 4:401-414

16- Kriegeskorte N, Mur M, Ruff DA, Kiani R, Bodurka J et al. (2008) Matching categorical object representations in inferior temporal cortex of man and monkey. Neuron 60(6): 1126-1141.

17- Bar M (2004) Visual objects in context. Nat Rev Neurosci 5(8): 617-629.

18- Gronau N, Neta M, Bar M (2008) Integrated Contextual Representation for Objects' Identities and Their Locations. J Cogn Neurosci 20(3): 371-388.

19- Miller MB, Gazzaniga MS (1998) Creating false memories for visual scenes. Neuropsychologia 36(6): 513-520.

20- Palmer SE (1975) The effects of contextual scenes on the identification of objects. Mem Cognit 3(5): 519-526.

21- Joubert OR, Rousselet GA, Fize D, Fabre-Thorpe M (2007) Processing scene context: Fast categorization and object interference. Vision Res 47(26): 3286-3297.

22- Rousselet GA, Joubert OR, Fabre-Thorpe M (2005) How long to get to the “gist” of real-world natural scenes? Visual Cogn 12(6): 852-877.

23- Mace MJ, Joubert OR, Nespoulous JL, Fabre-Thorpe M (2009) The time-course of visual categorizations: you spot the animal faster than the bird. PLoS ONE 4(6): e5927.

24- Joubert OR, Fize D, Rousselet GA, Fabre-Thorpe M (2008) Early interference of context congruence on object processing in rapid visual categorization of natural scenes. J Vis 8(13): 1-18.

25- Davenport JL, Potter MC (2004) Scene consistency in object and background perception. Psychol Sci 15(8): 559-564.

26- Fize D, Cauchoix M, Fabre-Thorpe M (2011) Humans and monkeys share visual representations. Proc Natl Acad Sci U S A 108(18): 7635-7640.

27- Oliva A, Torralba A (2001) Modeling the shape of the scene: A holistic representation of the spatial envelope. Int J Comput Vision 42(3): 145-175.

28- Torralba A, Oliva A (2003) Statistics of natural image categories. Network 14(3): 391-412.

29- Joubert OR, Rousselet GA, Fabre-Thorpe M, Fize D (2009) Rapid visual categorization of natural scene contexts with equalized amplitude spectrum and increasing phase noise. J Vis 9 (1): 1-16.

30- Loschky LC, Sethi A, Simons DJ, Pydimarri TN, Ochs D et al. (2007) The importance of information localization in scene gist recognition. J Exp Psychol Hum Percept Perform 33(6): 1431-1450.

31- Guyader N, Chauvin A, Peyrin C, Herault J, Marendaz C (2004) Image phase or amplitude? Rapid scene categorization is an amplitude-based process. C R Biol 327(4): 313-318.

32- Siegel S (1956) Nonparametric statistics for the behavioral sciences: McGraw Hill. Chapter 5: The randomization for matched pairs.

33- Gaffan D, Heywood CA (1993) A spurious category-specific visual agnosia for living things in normal human and non-human primates. J Cogn Neurosci 5(1): 118-128

34- Neider MB, Zelinsky GJ (2006) Searching for camouflaged targets: Effects of target-background similarity on visual search. Vision Res 46(14): 2217-2235.

35- Wolfe JM, Oliva A, Horowitz TS, Butcher SJ, Bompas A (2002) Segmentation of objects from backgrounds in visual search tasks. Vision Res 42(28): 2985-3004.

36- Kiani R, Esteky H, Mirpour K, Tanaka K (2007) Object category structure in response patterns of neuronal population in monkey inferior temporal cortex. J Neurophysiol 97(6): 4296-4309.

37- Oliva A, Torralba A (2007) The role of context in object recognition. Trends Cogn Sci 11(12): 520-527.

38- Fagot J, Martin-Malivel J, Dépy D (2000) What is the evidence for an equivalence between objects and pictures in birds and nonhuman primates? . In: Fagot J, editor. Picture perception in animals: Psychology press.

39- Truppa V, Spinozzi G, Stegagno T, Fagot J (2009) Picture processing in tufted capuchin monkeys (Cebus apella). Behav Processes 82(2): 140-152.

40- Parron C, Call J, Fagot J (2008) Behavioural responses to photographs by pictorially naive baboons (Papio anubis), gorillas (Gorilla gorilla) and chimpanzees (Pan troglodytes). Behav Processes 78(3): 351-357.

41- Pokorny JJ, de Waal FB (2009) Monkeys recognize the faces of group mates in photographs. Proc Natl Acad Sci U S A 106(51): 21539-21543.

42- Fagot J, Thompson RK, Parron C (2010) How to read a picture: Lessons from nonhuman primates. Proc Natl Acad Sci U S A 107(2): 519-520.

Chapitre 3 : L'importance de la congruence sémantique dans la catégorisation multimodale chez l'homme

Après avoir exploré dans le précédent chapitre les caractéristiques des stimuli participant à l'effet de congruence contextuelle dans le domaine visuel chez l'homme et le singe, nous nous proposons d'étudier dans ce dernier chapitre la congruence sémantique en multisensoriel, chez l'homme uniquement. Comme dans les deux précédentes études, nous avons utilisé une tâche de catégorisation (Animal versus Instrument de musique) et nous avons attaché une grande importance au contrôle des différents paramètres physiques des stimuli, dans les domaines auditif et visuel. Nous avons pour cela choisi de réutiliser la technique SWIFT dans le domaine visuel, et d'utiliser une technique de randomisation de *snippets* dans le domaine auditif. Ce faisant, nous avons égalisé un maximum de paramètres de bas niveau afin de déterminer l'importance du gain multisensoriel dû aux caractéristiques sémantiques des stimuli.

Nos stimuli étant très bruités, les performances des sujets étaient assez basses en condition unisensorielle, mais augmentaient significativement en condition multisensorielle congruente, sans jamais toutefois atteindre un optimum. Nous avons donc mis en évidence un gain multisensoriel important dû aux aspects sémantiques, bien que nous ayons constaté que les deux catégories choisies n'étaient pas équivalentes. Cette étude bénéficierait d'une réplication en EEG afin de déterminer à quelle latence et dans quelles zones cérébrales se fait l'intégration des différentes composantes sensorielles d'un même objet.

I) Introduction

En conditions écologiques, la majorité des évènements perceptuels que nous rencontrons nous fournissent simultanément des informations via différentes modalités sensorielles. Par exemple, lorsque nous évoluons dans une ville, nous voyons les voitures, nous entendons le bruit de leurs moteurs et sentons l'odeur des pots d'échappement, nous percevons également les piétons, leurs conversations, les effluves des cuisines nous parviennent lorsque nous passons devant un restaurant… L'environnement est riche, mais nous combinons automatiquement les différents aspects sensoriels provenant d'une même source et effectuons une ségrégation des évènements provenant de sources différentes, afin d'avoir une représentation complète et cohérente du monde qui nous entoure. Dans de telles situations, les aspects temporels, spatiaux et sémantiques sont autant d'indices traités par les différents systèmes sensoriels afin d'identifier l'objet.

Les études portant sur la reconnaissance d'objets sont nombreuses dans le domaine visuel, modalité sensorielle dominante chez l'homme (Colavita, 1974), mais le sont nettement moins dans les domaines auditif et audiovisuel. Bien que les études sur la congruence sémantique se développent ces quinze dernières années, les travaux explorant l'effet de congruence sémantique dans la reconnaissance d'objets multi-sensoriels restent encore rares.

Les effets de la congruence sémantique ont été étudiés au sein des modalités visuelle et auditive. Ces études s'intéressaient notamment à l'effet du contexte sur la catégorisation d'un objet désigné comme cible de la tâche. Dans le domaine visuel, Davenport et Potter en 2004 (Davenport and Potter, 2004), ont par exemple montré que l'on était meilleur et plus rapide pour reconnaître un objet situé au premier plan quand l'arrière plan était sémantiquement congruent. En 2011, Fize et al (Fize et al., 2011) ont confirmé ces résultats lors d'une tâche de catégorisation rapide, animal versus non-animal à la fois chez l'homme et le singe. Ils ont donc démontré que la congruence entre le contexte et l'objet cible était un facteur crucial lors de la catégorisation ultrarapide. Dans la modalité auditive, des effets similaires ont été observés. Ballas et Mullins en 1991 (Ballas and Mullins, 1991) ont ainsi montré que l'identification d'un son ambigu pouvait être biaisée

si on insérait le son cible au sein d'une séquence de sons environnementaux sémantiquement incongruents. En 2008, Niessen et al (Niessen et al., 2008) ont également comparé les performances d'identification de sons ambigus (c'est-à-dire pouvant être attribués à au moins deux sources différentes) précédés ou non d'un son sémantiquement relié à l'une des sources possibles. Ils ont ainsi montré que les sujets étaient plus enclins à identifier le son comme provenant de la source congruente avec le son amorce.

L'effet de la congruence sémantique a aussi été démontré en multimodal, où l'information dans une modalité sensorielle peut jouer le rôle de contexte pour la détection d'une cible dans une autre modalité sensorielle (Doehrmann and Naumer, 2008). Ainsi la présentation de scènes visuelles ou de photographies d'objets facilite l'identification de sons sémantiquement reliés aux images (Özcan and Van Egmond, 2009). Cette facilitation se retrouve aussi bien en termes de performances d'identification qu'en termes de temps de réaction. En 2008, Schneider et collègues (Schneider et al., 2008) ont comparé les performances d'identification d'objets visuels ou sonores, soit isolément, soit appariés au sein d'une modalité ou encore appariés entre modalités. Comme attendu, leur première expérience a montré que les sujets étaient meilleurs dans le domaine visuel que dans le domaine auditif lors de l'identification d'un stimulus isolé. Dans une deuxième expérience les sujets devaient juger la taille d'un objet (auditif ou visuel) précédé d'un stimulus amorce (auditif ou visuel) congruent ou incongruent. Ils ont alors montré que les sujets étaient toujours plus lents quand le stimulus amorce était incongruent, mais ceci quelle que soit la modalité sensorielle. Dans ces études, la congruence sémantique aide les sujets à identifier un objet, probablement grâce à la redondance de l'information dans les deux modalités. Toutefois, du fait de la dissociation temporelle entre les stimuli auditifs et visuels, ces travaux ne peuvent mettre en lumière d'éventuels mécanismes d'intégration multisensorielle.

En 2004, Laurienti et al ont mis au point une expérience très simple de reconnaissance de la couleur (Laurienti et al., 2004). Les sujets devaient discriminer entre rouge, bleu et vert. Les stimuli visuels étaient des disques colorés, les stimuli auditifs, l'énonciation des mots « rouge », « bleu » ou « vert ». Les essais pouvaient être soit unimodaux, soit bimodaux congruents, soit bimodaux incongruents. Dans le cas des essais bimodaux, les stimuli

auditif et visuel étaient présentés simultanément. Grâce à ce paradigme expérimental, ils ont mis en évidence une facilitation dans les essais bimodaux congruents par rapport aux essais unimodaux et unimodaux incongruents. De plus ils ont montré une violation du *Race model* (Miller, 1986) dans la condition audiovisuelle congruente. Le Race model donne la probabilité optimale de la détection d'une cible audiovisuelle en fonction des probabilités de réponse dans chacune des modalités, dans le cas où il n'y aurait pas d'intégration des signaux auditifs et visuels. Lorsque le taux de réponses observé au cours du temps est supérieur au Race model, alors on considère que des mécanismes intégratifs ont été mis en jeu. Cette étude ne manipule pas de larges catégories, mais se limite aux couleurs. En 2010, Chen et Spence (Chen and Spence, 2010) ont étudié l'impact des sons sur l'identification d'images masquées. Ils ont pour cela utilisé des images et des sons d'objets complexes. Ils ont ainsi montré que la congruence sémantique d'un son influençait l'identification d'images flashées puis masquées par les participants. La facilitation de la reconnaissance était d'autant plus importante que le son était présenté simultanément avec l'image. La présentation d'un son incongruent, au contraire, altérait les performances en deçà de la condition contrôle (sans stimulus sonore) et générait davantage de fausses reconnaissances. Murray et collaborateurs ont de surcroît montré que la présentation d'un son congruent lors d'une tâche de reconnaissance d'images (i.e. l'image a-t-elle déjà été présentée ?) améliorait la mémorisation de l'image, mais l'effet était inverse lorsque le son présenté était incongruent ou sans contenu sémantique (pure tone)(Murray et al., 2004, Lehmann and Murray, 2005). Enfin Molholm et collègues ont montré que lors d'une tâche de reconnaissance d'objet en go/no-go (i.e. « pressez le bouton quand vous voyez et/ou entendez la vache »), les sujets étaient plus rapides quand le stimulus était bimodal congruent que lorsque le stimulus était bimodal incongruent ou unimodal (Molholm et al., 2004). Ils ont ainsi obtenu une violation significative du Race model dans la condition bimodale congruente.

Toutes ces équipes de recherche se sont intéressées à l'effet de la congruence sémantique des stimuli dans les différentes modalités sensorielles, toutefois elles n'ont pas exploré la possibilité de la contribution de paramètres physiques de plus bas niveau à cet effet de congruence. En revanche Werner et Noppeney en 2010, ont tenté de se pencher davantage sur la contribution éventuelle des paramètres physiques de stimuli au

gain multisensoriel chez des sujets humains(Werner and Noppeney, 2010). Leurs stimuli étaient constitués de films de manipulation d'outils ou d'instruments de musique, et des bandes son correspondantes. Mais ils ont également créé des séquences de bruit blanc, ainsi que des séquences de bruit visuel. Ainsi ils ont pu comparer la condition congruente film+son correspondant avec des conditions film+bruit blanc ou bien bruit visuel+son, les sujets étant engagés dans une tâche de catégorisation outils versus instrument de musique. Ils ont montré une augmentation des performances des sujets en condition audiovisuelle, aussi bien en termes de taux de réussite, de temps de réaction que de d', en comparaison avec les conditions unimodales ou unimodales+bruit. Toutefois ils n'ont pas exploré l'effet de l'incongruence sémantique, et les bruits visuels ou sonores étaient tous identiques et ne portaient pas d'information physique correspondant à celles des films ou bandes sons.

Dans la présente étude, nous avons donc cherché à isoler l'effet de congruence sémantique des stimuli lors d'une tâche de catégorisation en go/no-go Animal versus Instrument de musique, chez l'homme. Nous avons donc décidé de stimuler en permanence les deux modalités sensorielles, visuelle et auditive, grâce à des séquences où le contenu sémantique était intégré dans du bruit, bruit qui était construit à partir des images ou bandes sons utilisées comme cibles. Ce faisant, nous avons pu égaliser au cours du temps la luminance, le contraste et les fréquences spatiales dans le domaine visuel, ainsi que le contenu fréquentiel avec l'amplitude associée dans le domaine auditif.

II) Matériel et méthodes

A) Les sujets

Quinze sujets ont participé à l'expérience, 7 femmes, 3 gauchers, avec un âge moyen de 29ans (24 à 40 ans). Tous ont rapporté avoir une vision normale ou corrigée et une audition normale. Ils ont été soumis à une tâche de catégorisation en go/no-go, « Animal » versus « Instrument de musique » (l'une ou l'autre des catégories étant la cible de la tâche).

B) Les stimuli

1) Les catégories

Nous avons utilisé les catégories « Animal » et « Instrument de musique ». Nous avons sélectionnés 25 items sémantiques dans chaque catégorie, identiques dans les modalités visuelle et auditive. Nous avons ensuite effectué 500 tirages aléatoires de paires de stimuli pour constituer nos essais.

2) La construction des séquences d'évènements

Les stimuli étaient des séquences audio-visuelles de 10 secondes, composées majoritairement de bruit, au cours desquelles étaient présentés des évènements auditifs, visuels, audio-visuels congruents (congruence au niveau de l'item) et/ou audio-visuels incongruents (incongruence au niveau catégoriel), apparaissant dans un ordre et à des temps aléatoires. Un évènement correspondait à l'apparition du son ou de l'image d'origine dans la séquence bruitée (cf paragraphes 3&4). Deux évènements étaient séparés de 1 seconde au minimum (entre les 2 onsets), le premier évènement apparaissait au minimum 500ms après le début de la séquence et le dernier, au plus tard 800ms avant la fin de la séquence. Une séquence pouvait contenir de un à cinq évènements.

Prenons l'exemple d'une séquence dont les deux entités sémantiques étaient une vache et un piano. Le tirage aléatoire a conduit à 5 évènements dans cette séquence, ayant lieu aux temps 1, 3.7, 5, 6.3, et 8 (en secondes à partir du début de la séquence). Ces

évènements étaient respectivement : piano-visuel, piano-audio, vache-audiovisuel (audiovisuel congruent), piano-visuel/vache-audio et piano-visuel (audiovisuel incongruent).

3) Les séquences visuelles

Les séquences visuelles ont été construites en utilisant la technique SWIFT mise au point par Koenig et VanRullen (Koenig-Robert and VanRullen, 2013) via Matlab R2013a. Cette technique permet de détruire et reconstruire une image de manière cyclique tout en conservant le contraste, la luminance et les fréquences spatiales stables au cours du temps.

A partir des deux images nous avons construit deux séquences SWIFT, à des fréquences différentes. Les fréquences étaient déterminées par la durée entre deux évènements dans la même modalité et la même catégorie.

Si l'on reprend l'exemple précédent, nous avons donc construit une séquence « vache » et une séquence « piano » (cf figure 3.1), la vache devant apparaître de manière visible une seule fois, à 5s, et le piano trois fois à 1s, 6.3s et 8s.

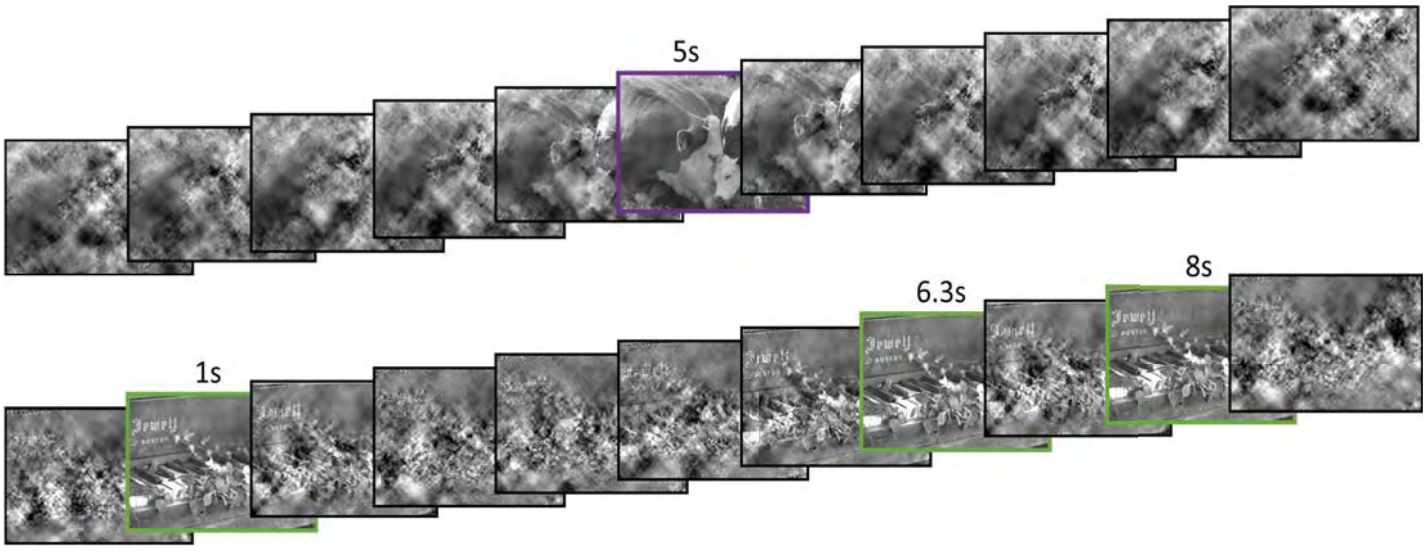


Figure 3.1 : exemple de séquences SWIFT, en haut vache, et en bas piano. Les onsets sémantiques ont été encadrés en couleur.

Ces deux séquences ont ensuite été superposées en transparence en fixant la couche alpha à 0.5. La séquence visuelle résultante contient donc quatre évènements, chacun altéré par le bruit de l'autre catégorie (cf figure 3.2).

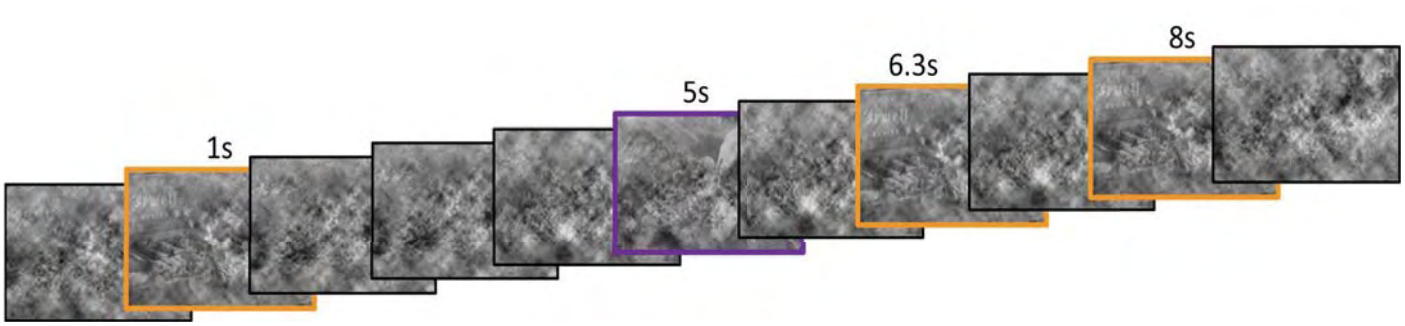


Figure 3.2 : exemple de la séquence SWIFT finale. Les onsets sémantiques ont été encadrés en couleur.

4) Les séquences auditives

Les séquences auditives ont été construites en utilisant une technique de randomisation d'extraits de sons. Nous avons tout d'abord sélectionné des séquences sonores de 300ms dont le contenu sémantique était aisément reconnaissable (cinq sujets, non inclus dans les 15 ayant passé l'expérience, ont été soumis à un test de reconnaissance de ces sons). Les sons ont été égalisés en amplitude, et les silences éventuels ont été supprimés grâce au logiciel Praat. La fréquence d'échantillonnage sonore était de 44100Hz.

Nous avons ensuite découpés ces séquences de 300ms en N extraits (snippets) de durée égale pour un même son, mais pouvant varier de 0.5ms à 6ms entre les sons. L'ordre de ces snippets a été ensuite randomisé de manière à détruire l'enveloppe du son, et nous avons créé des séquences de 10 secondes, au sein desquelles apparaissait la séquence originale aux temps déterminés par la séquence des évènements. Le choix de la durée du snippet a été effectué de manière à casser la rythmicité intrinsèque des sons, tout en maximisant la variabilité du bruit. Ainsi le cri d'un criquet par exemple, contenait une telle rythmicité intrinsèque qu'il a fallu choisir la durée minimale de 0.5ms afin de le rendre méconnaissable après randomisation.

Revenons à l'exemple précédent. Nous avons donc choisi deux extraits sonores de la vocalisation d'une vache et de quelques notes de piano. Puis nous avons découpé ces deux sons en snippets de 2 ms pour la vache et 6 ms pour le piano.

Voici les tracés des deux sons originaux, et d'extraits de même durée, randomisés :

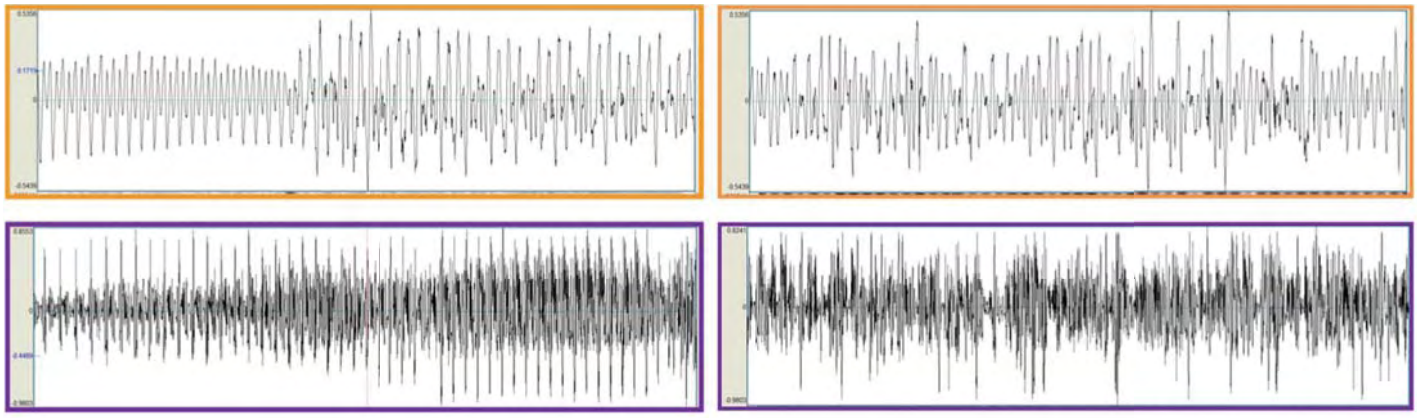


Figure 3.3 : tracés des sons originaux à gauche, et de séquences randomisées à droite. En orange, sons de piano et en violet de beuglement de vache.

Après avoir créé deux séquences de 10 secondes, contenant du bruit (snippets randomisés) avec, aux temps donnés par la séquence d'évènements, les sons originaux, nous superposons les deux. Concernant notre exemple nous obtenons alors une séquence auditive contenant 3 évènements : piano à 3.7s, vache à 5 et 6.3s.

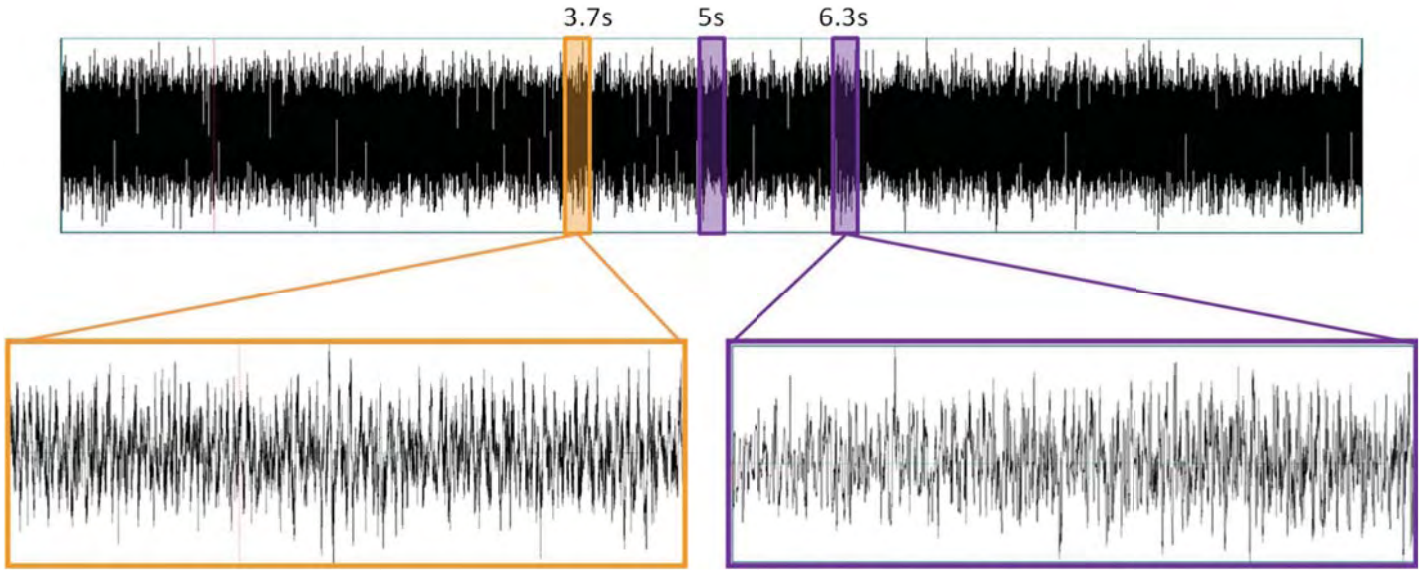


Figure 3.4 : Tracé de la séquence piano-vache, avec zoom sur les évènements auditifs, « piano » en orange, « vache » en violet

On constate que l'ajout de bruit rend le tracé du son méconnaissable, toutefois perceptuellement, le contenu sémantique du son restait identifiable, bien que moins net.

Toutes ces opérations réalisées dans les domaines auditif et visuel nous ont permis d'égaliser un maximum de paramètres de bas niveau, d'avoir un niveau de bruit constant rendant la tâche plus difficile et de stimuler en continu les modalités auditive et visuelle.

C) Le protocole

Nous avons utilisé le logiciel de psychophysique EventIDE afin de garantir une bonne synchronisation audiovisuelle (+/-2ms). Les sujets étaient placés dans un box noir, à 50cm de l'écran de stimulation, tête fixe. La fréquence de rafraichissement de l'écran était de 60Hz (fréquence à laquelle les séquences SWIFT avaient été construites). Le son était dispensé par des écouteurs intra-auriculaires Etymotic^{R}, le volume réglé de manière confortable pour chaque participant.

Les sujets ont été soumis à une tâche de catégorisation go/no-go « Animal » versus « Instrument de musique ». Ils ont chacun effectué deux sessions, contenant 10 blocs chacune, chaque bloc étant constitué de 25 essais. Au total ils ont donc réalisé 500 essais, et n'ont vu qu'une seule fois chaque séquence. Au début de chaque bloc était donnée l'instruction sur la catégorie cible.

Pour démarrer un bloc ou un essai, le sujet devait presser la touche « E » pour les droitiers et « I » pour les gauchers avec la main non dominante. Ensuite durant l'essai, ils avaient pour consigne de presser la barre espace aussi vite que possible avec leur main dominante, chaque fois que la catégorie cible apparaissait, quelle que soit la modalité (de zéro à cinq appui(s) possible(s) dans un essai, cf figure 3.5). Tous les appuis étaient enregistrés au cours des essais.

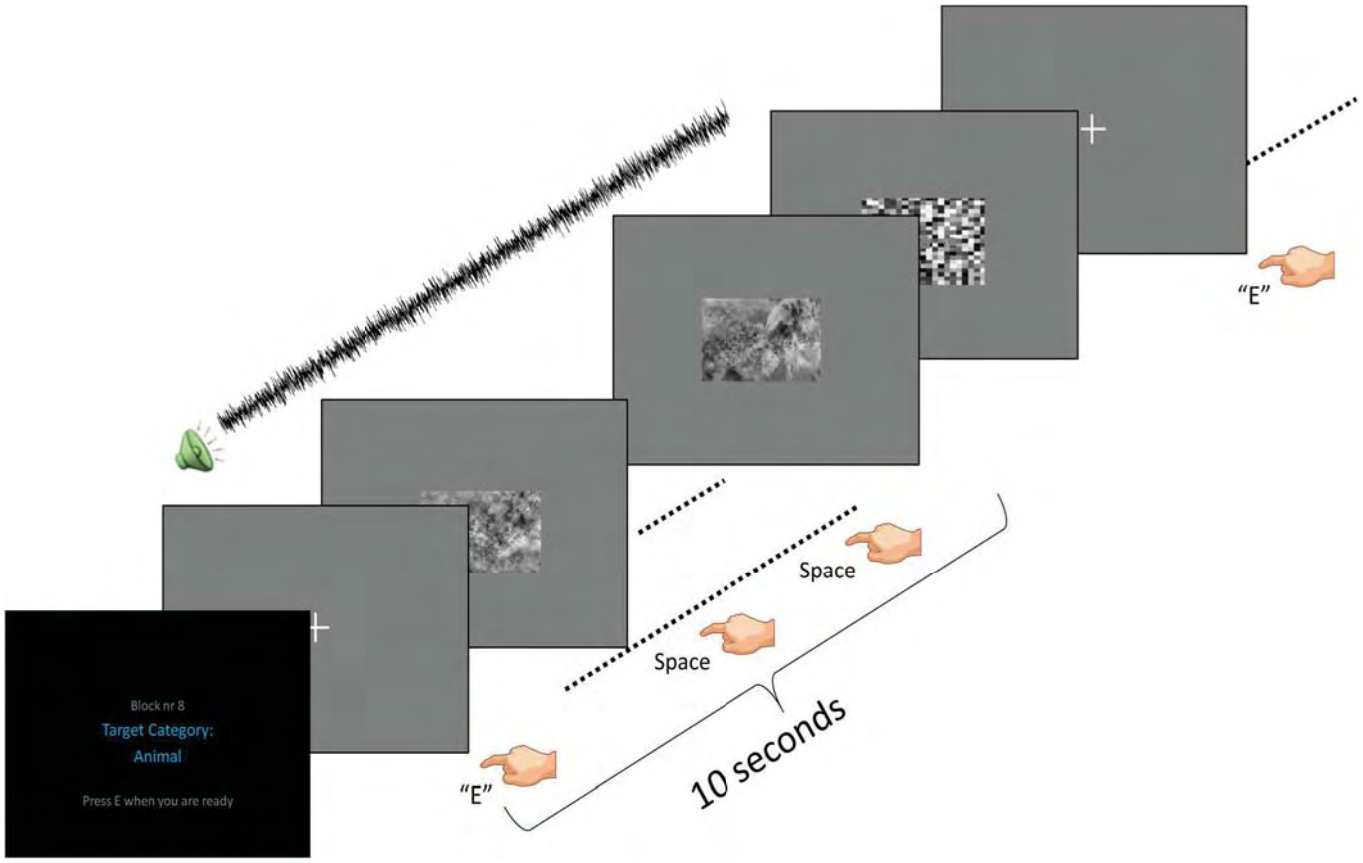


Figure 3.5 : Déroulement d'un essai (au début d'un bloc)

D) Analyses

Les résultats ont été analysés avec Matlab R2013a. Les sujets ont été groupés grâce à une technique de vincentisation des données. Les données (précision et temps de réaction) ont été analysées selon la modalité de stimulation (Audio seul : A, Visuel seul : V, Audiovisuel congruent : AV, Audiovisuel incongruent : AVi) et selon la catégorie (Animal ou Instrument de musique).

1) Vincentisation

Pour grouper nos sujets, nous avons découpé leur répartition de temps de réaction (toutes réponses confondues) en 20 bins de durée égale au sein d'un même sujet. Nous avons ensuite moyenné les sujets au sein de chaque bin. Cette méthode nous a permis de nous affranchir des différences interindividuelles en termes de rapidité. Nous avons toujours supprimé les temps de réaction inférieurs à 200ms après l'onset d'un évènement, considérant qu'il s'agissait d'anticipations.

Les données seront donc présentées par la suite en fonction des bins de temps, et non en fonction des temps absolus.

2) Probabilité cumulée de réponses

Nous avons calculé le taux réponses (réponses go correctes : *hits* et fausses alarmes *FA)* cumulé en fonction du temps, c'est-à-dire pour les hits au bin de temps T :

$$p(hits)_T = \frac{\sum_{t=1}^{T} hits}{nombre\ total\ de\ cibles}$$

3) Théorie de la détection du signal et d' cumulé

Une des mesures communément utilisées dans l'analyse de tâches de catégorisation est le d', mesure issue de la théorie de la détection du signal. Le d' permet d'évaluer à quel point le signal (ici la catégorie cible) émerge parmi le bruit (les distracteurs de la tâche), en prenant en compte à la fois le taux de hits et le taux de fausses alarmes.

d'= z(hits)-z(FA) où z représente la fonction cumulative de la fonction normale inverse.

Nous avons ainsi pu calculer le d' cumulé en fonction des bins de temps de réaction grâce aux valeurs des probabilités cumulées de hits et fausses alarmes obtenues comme précédemment décrit.

4) Race model

Nous avons utilisé le *race model* pour évaluer l'importance de l'intégration multi-sensorielle lors de notre tâche. Le race model est un modèle de sommation probabilistique qui évalue la probabilité optimale de réponse dans la condition multi-sensorielle s'il n'y a pas d'intégration multi-sensorielle. Il se calcule à partir des probabilités de réponses observées dans les modalités uni-sensorielles comme suit :

$p(AV)_{race\ model} = p(A)+p(V)-p(A)*p(V)$

Nous avons calculé ce race model au cours du temps avec les probabilités cumulées uni-sensorielles et comparé la courbe obtenue à la courbe des réponses multi-sensorielles observées. Si la courbe observée était significativement au dessus de la courbe du Race model, cela nous permettait de conclure qu'il y avait intégration multi-sensorielle, et donc une interaction des modalités auditive et visuelle.

Enfin nous avons combiné les mesures du d' et du Race model, afin de calculer un d' théorique optimal dans le cas où il n'y aurait pas d'intégration multisensorielle :

$d'_{race\ model} = z(p(\text{hits AV})_{race\ model}) - z(p(\text{FA AV})_{race\ model})$

Et nous avons comparé de la même manière le d' observé et le d' théorique afin de déterminer l'importance de l'intégration multi-sensorielle.

III) Résultats

A) Résultats généraux : effet de congruence

Tous les sujets ont correctement réalisé la tâche de catégorisation. Les cibles étaient moins bien détectées chez tous les sujets dans la modalité visuelle seule (45% de hits) contre les modalités auditive seule (65% de hits) et audiovisuelle congruente (75% de hits). Toutefois c'est dans les modalités visuelle seule et audiovisuelle congruente que les sujets étaient les plus rapides (TR moyen=450ms). La modalité audiovisuelle incongruente se situe entre les modalités uni-sensorielles, tant en termes de temps de réaction qu'en termes de taux de réussite (notons que pour ces essais, nous ne pouvions calculer de taux de fausses alarmes, puisqu'une des deux modalités contenait obligatoirement la catégorie cible). Enfin la modalité multi-sensorielle conduit à une meilleure détectabilité du signal que les modalités uni-sensorielles.

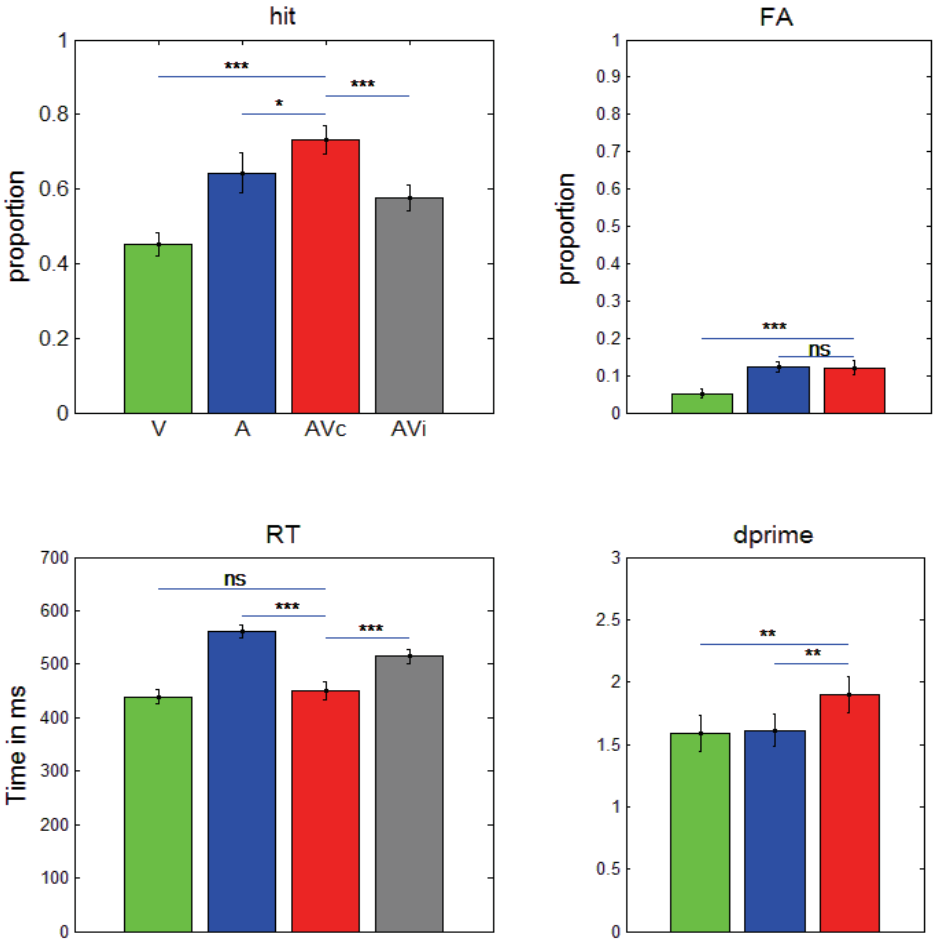


Figure 3.6 : De gauche à droite et de haut en bas, Taux moyens de hits et de fausses alarmes dans les différentes modalités sensorielles, temps de réaction moyens et d' moyens.

Significativité donnée par un test de Friedman pour mesures répétées. Valeurs corrigées pour comparaisons multiples.

Comparaisons réalisées entre la modalité audiovisuelle congruente et les autres modalités sensorielles.

La condition multi-sensorielle congruente présente donc un avantage tant en termes de taux de réussite, de temps de réaction et de d' comparée aux conditions unisensorielles et multisensorielle incongruente.

B) L'intégration multi-sensorielle : le Race model

Les analyses qui suivent se concentrent sur la modalité multi-sensorielle congruente. Nous avons donc calculé les taux de hits et fausses alarmes cumulés par bin de temps de réaction, dans les modalités uni-sensorielles et multi-sensorielle congruente afin de calculer le race model, et le d' cumulé. Les catégories « Animal » et « Instrument de musique » sont combinées dans ces analyses.

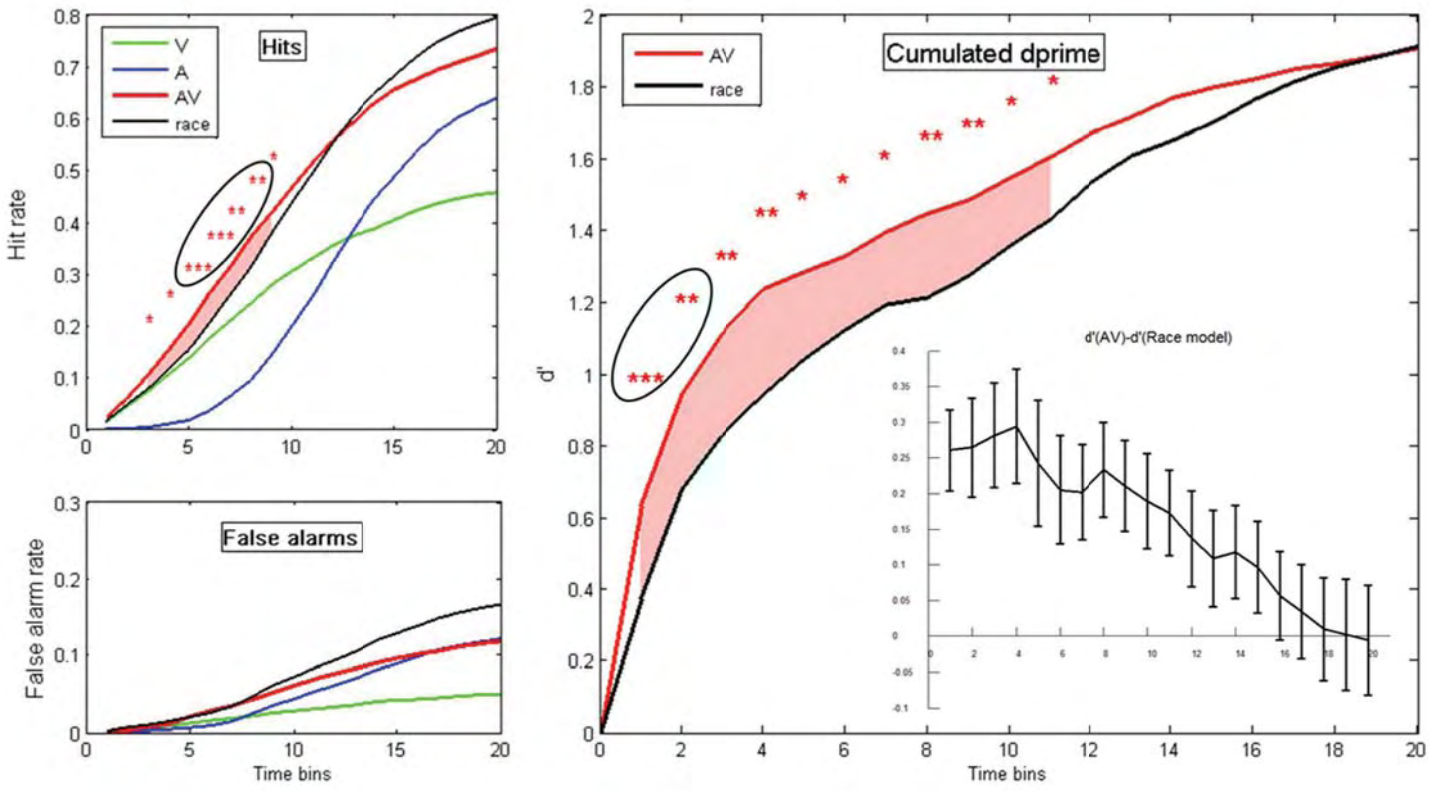


Figure 3.7 : A gauche :Taux de hits et de fausses alarmes, cumulés par bin de temps de réaction. A droite : d' cumulé par bin de temps de réaction pour la condition audio-visuelle congruente
V= visuel seul, A=Audio seul, AV= Audio-visuel congruent, Race= prédiction du race model. L'aire en rouge pale donne la zone de significativité de la différence entre la courbe AV et la courbe race. Les astérisques entourés sont les points qui résistent à la correction pour comparaisons multiples.

Nous observons donc une violation du Race model dans la condition multi-sensorielle congruente, violation robuste à une correction pour comparaison multiple dans les bins 5 à 8 concernant les hits (soit environ de 350ms à 470ms après l'onset sémantique), et dans les deux premiers bins concernant le d', soit de 200 à 260ms environ après l'onset. Toutefois il est intéressant de noter que le d' observé est toujours supérieur ou égal au d' prédit par le Race model, il est très supérieur dans les temps précoces et la différence tend à diminuer à des latences plus importantes. De plus aucune violation du Race model n'est observée concernant les fausses alarmes en condition multi-sensorielle congruente, ce qui semble indiquer que la condition multi-sensorielle ne prédispose pas simplement à répondre davantage, mais à répondre mieux, surtout pour les essais les plus rapides.

C) L'importance de la congruence sémantique

Dans les essais audiovisuels congruents, nous avons trouvé une violation significative du Race model, tant en termes de taux de hits que de d'. Mais qu'en est-il concernant les essais audiovisuels incongruents ? Nous avons donc calculé le Race model à partir des probabilités de hits dans une modalité et de fausses alarmes dans l'autre modalité :

$$p(\text{hits AVi})_{\text{race model}} = \frac{1}{2}\,[p(\text{hits V})+p(\text{FA A})-p(\text{hits V})xp(\text{FA A}) + p(\text{hits A})+p(\text{FA V})-p(\text{hits A})*p(\text{FA V})]$$

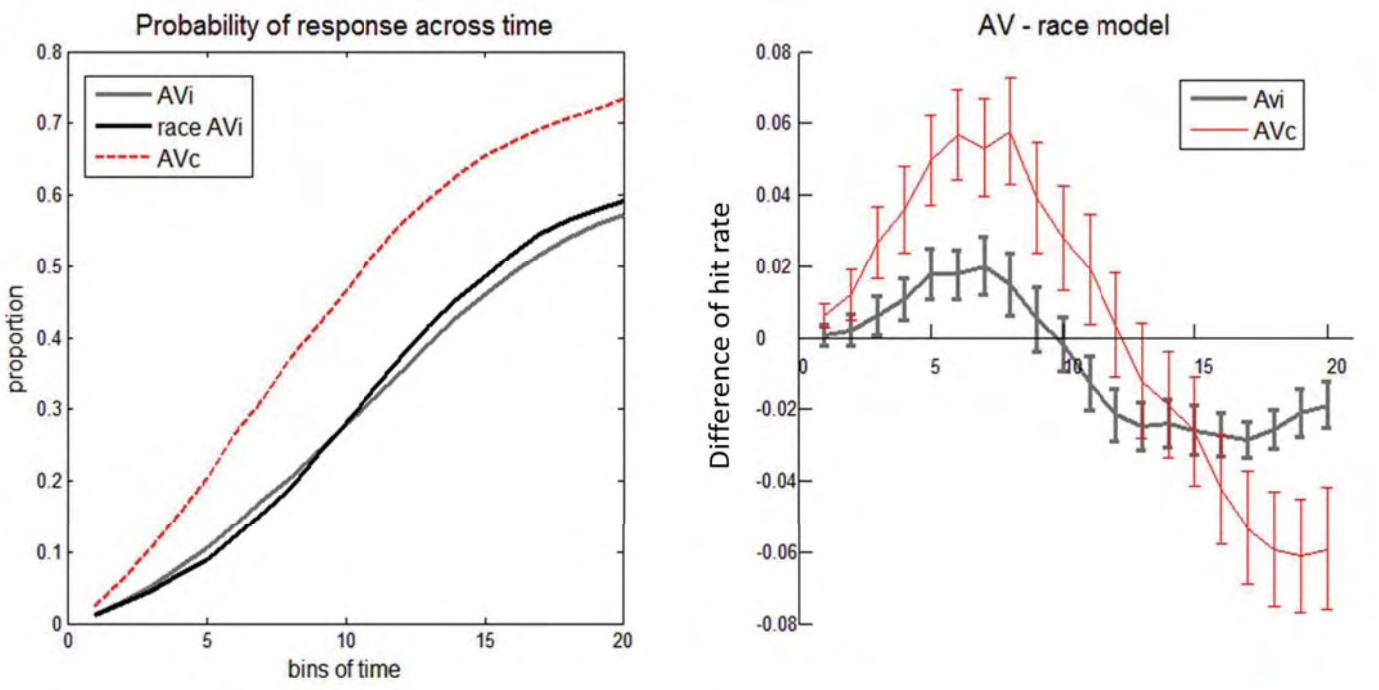


Figure 3.8 : A gauche : Taux de hits observés pour les essais incongruents et comparaison avec le race model. A droite : En rouge sont montrés les hits pour les essais audiovisuels congruents (pour référence), et la différence avec le race model pour essais congruents (différence entre les courbes AV et race de la figure 3.7-hits)

Nous ne constatons aucune violation significative du race model pour les essais incongruents. Quand les stimuli dans les modalités auditive et visuelle ne sont pas congruents, il semble donc qu'il n'y ait pas d'intégration multisensorielle, mais plutôt que les deux stimuli sont traités en parallèle. De plus, les essais incongruents conduisent à un taux de réussite bien plus faible que les essais congruents. La stimulation conjointe des deux modalités ne peut pas, à elle seule, expliquer le plus fort taux de réussite dans les essais audio-visuels congruents.

D) Les différences entre catégories

1) Taux de réussite et race model

Nous avons dans cette étude, utilisé deux catégories, que nous avons pris soin de contrebalancer au sein de chaque sujet. Toutefois nous avons voulu tester les différences éventuelles qu'il pouvait exister entre ces deux catégories.

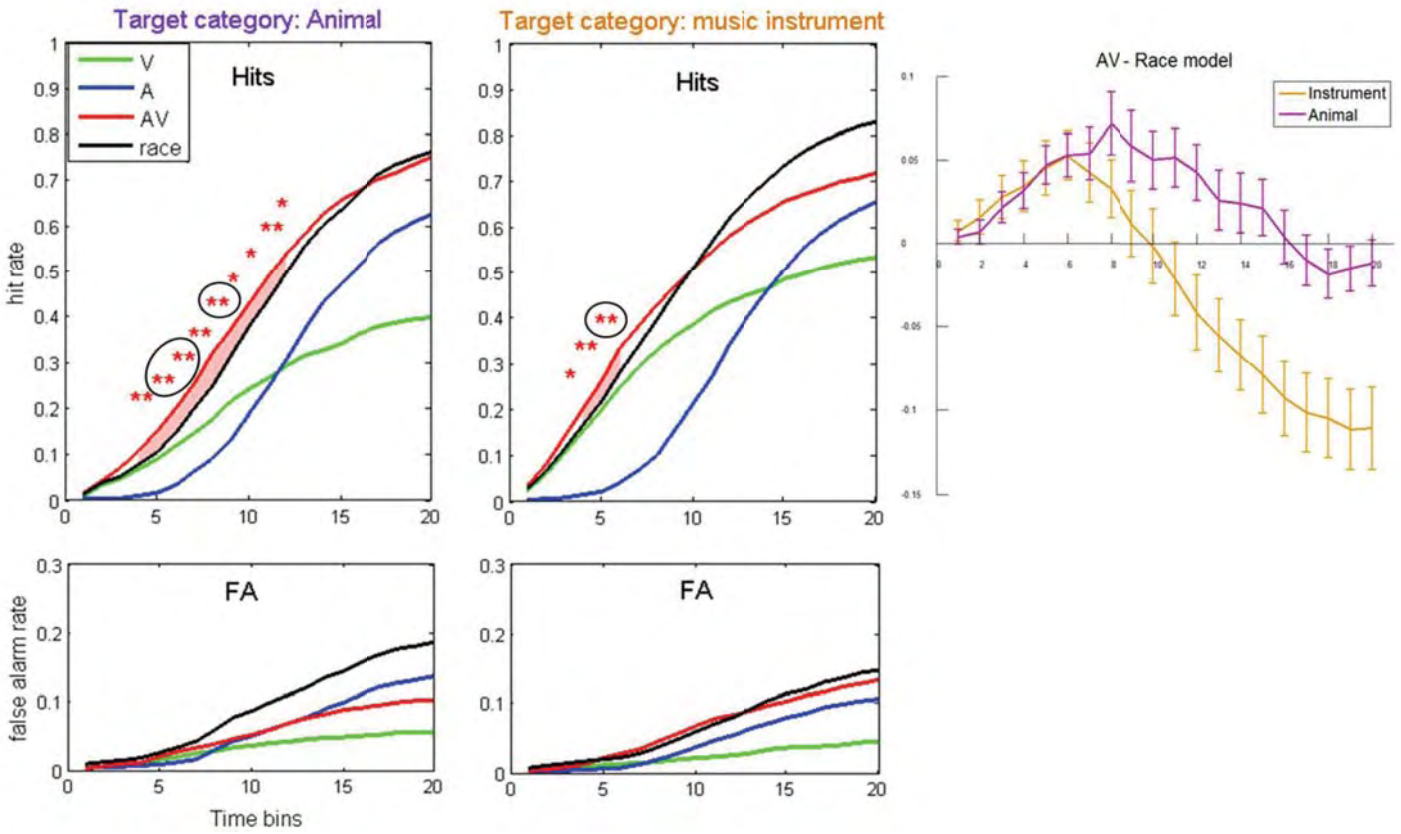


Figure 3.9 : De gauche à droite : taux de hits et de fausses alarmes dans les différentes modalités pour les deux catégories, Tracé de la différence entre les taux de hits dans la modalité audiovisuelle et du taux de hits prédit par le race model. Les astérisques donnent la significativité du test de Mann-Whitney sur échantillons appariés. Les astérisques entourés correspondent aux points dont la significativité résiste à la correction pour comparaisons multiples

La première différence que nous pouvons noter entre les catégories est le taux de réponses correctes dans la modalité visuelle seule. Les sujets ont en effet répondu significativement plus aux images d'instruments de musique qu'aux images d'animaux (test de Mann-Whitney pour échantillons appariés, $p<0.001$). Pour les modalités auditive et audiovisuelle en revanche, il n'y a pas de différence significative entre les deux catégories concernant le taux de hits. Dans le cas des fausses alarmes, les instruments de musique en ont engendré significativement plus que les vocalisations animales (test de Mann-Whitney, $p<0.01$). Il n'y a pas de différence entre les autres modalités.

Si l'on compare désormais les courbes de performance en audiovisuel et le race model, nous constatons des différences remarquables. Dans le cas où la catégorie cible était l'animal, les performances observées sont significativement supérieures au race model sur 8 bins de temps, dont trois résistent à la correction pour comparaisons multiples, et

les performances sont supérieures ou égales au race model sur l'ensemble des pas de temps pris en compte. Concernant la catégorie Instrument de musique en revanche, nous observons une violation du race model seulement pour trois bins de temps dans les réponses précoces, dont un point résistant à la correction pour comparaisons multiples, et une différence très négative pour les réponses tardives (les performances observées dans les 3 derniers bins de temps de réactions sont significativement inférieures à celles prédites par le Race model après correction pour comparaisons multiples).

Enfin sur le graphique représentant la différence entre les performances observées dans la modalité audiovisuelle et le race model, il est surprenant de constater que les courbes des deux catégories sont parfaitement superposées dans les cinq premiers bins de temps de réaction.

2) d' observé et d' prédit par le race model

Après avoir examiné les performances des sujets pour les deux catégories, il est nécessaire de calculer le d' dans la condition audiovisuelle et le d' prédit par le Race model. Nous pouvons à nouveau constater (figure 3.10) que les deux catégories ne sont pas équivalentes.

En ce qui concerne la catégorie Animal, le d' observé est significativement supérieur au d' prédit par le race model sur l'ensemble des temps de réaction analysés, même après correction pour comparaisons multiples (test de Mann-Whitney pour échantillons appariés).

Dans le cas de la catégorie Instrument de musique, les résultats sont plus contrastés. Nous observons une violation du Race model durant les 5 premiers bins de temps de réaction, mais seuls deux point survivent à la correction pour comparaisons multiples. Dans les quatre derniers bins de temps de réaction, le d' prédit par le race model est significativement supérieur au d' observé, toutefois cette significativité ne résiste pas à la correction pour comparaisons multiples.

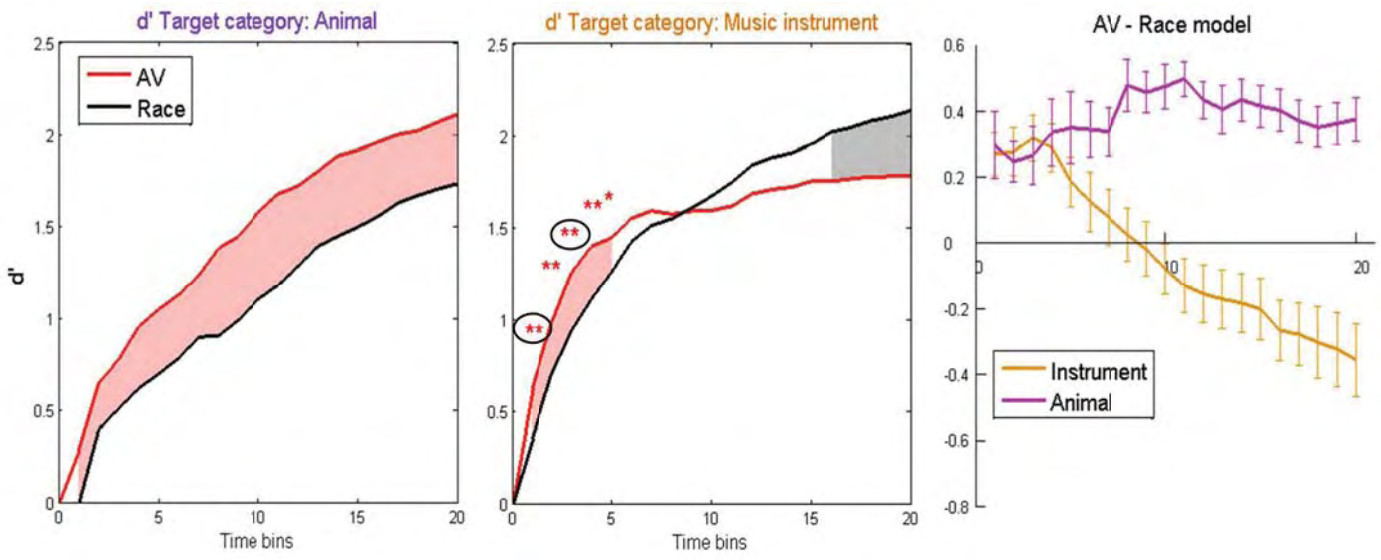


Figure 3.10 : De gauche à droite d' cumulé par pas de temps de réaction, courbe réelle et courbe prédite par le race model pour la catégorie Animal, pour la catégorie Instrument de musique, et représentation de la différence entre ces courbes. Pour le panel « Animal », les astérisques de significativité n'ont pas été représentés car présentes sur l'ensemble de la courbe, pour le panel « Instrument » les astérisques représentent la significativité avec un test de Mann-Whitney pour échantillons appariés, les astérisques entourés sont les points résistants à la correction pour comparaisons multiples.

Enfin si l'on compare les différences entre le race model et les valeurs observées pour les deux catégories, on constate que dans le cas de la catégorie cible « Animal », la différence de d' observé et d' prédit est positive et stable (entre 0.3 et 0.5), alors que pour la catégorie « Instrument » cette différence d'abord positive, décroît après le 4ème bin de temps de réaction et devient négative au dixième bin.

Le gain multisensoriel est donc équivalent pour les deux catégories dans des temps précoces, mais se différencie rapidement à des latences plus importantes.

IV) Discussion

Dans cette étude nous avons mis en évidence l'avantage de la condition multimodale congruente comparée aux conditions unimodales et bimodale incongruente grâce à un protocole de catégorisation en go/no-go. De plus nous avons tenté d'égaliser un maximum de paramètres de bas niveau dans les stimuli afin d'isoler l'effet de la congruence sémantique. La facilitation en condition bimodale congruente se retrouve à tous les niveaux de performances comportementales : les taux de réussite, les temps de réaction et le d'. Nous avons de plus obtenu des violations significatives du Race model, tant en termes de hits qu'en terme de d'.

En 2012 Otto et Mamassian ont remis en question la théorie du Race model (Otto and Mamassian, 2012). Ils ont ainsi élaboré le « *noisy race model* ». Ce modèle se base sur des modèles de prise de décision perceptuelle. La prise de décision implique l'accumulation d'indices sensoriels dans le temps, et ce processus est partiellement corrompu par du bruit. Lorsque deux modalités sensorielles sont stimulées conjointement, le bruit lié aux processus neuraux augmente. En augmentant le bruit, la variance de la distribution des temps de réaction pour la prise de décision augmente également. Ce qui implique une accélération des temps de réaction courts et un ralentissement des temps de réaction les plus lents. Selon ce modèle, donc, une violation du Race model ne signifie pas nécessairement que des processus intégratifs de l'information aient été mis en jeu. Les déviations par rapport au Race model s'expliqueraient alors par une augmentation de la variance due au bruit (cf figure 3.11).

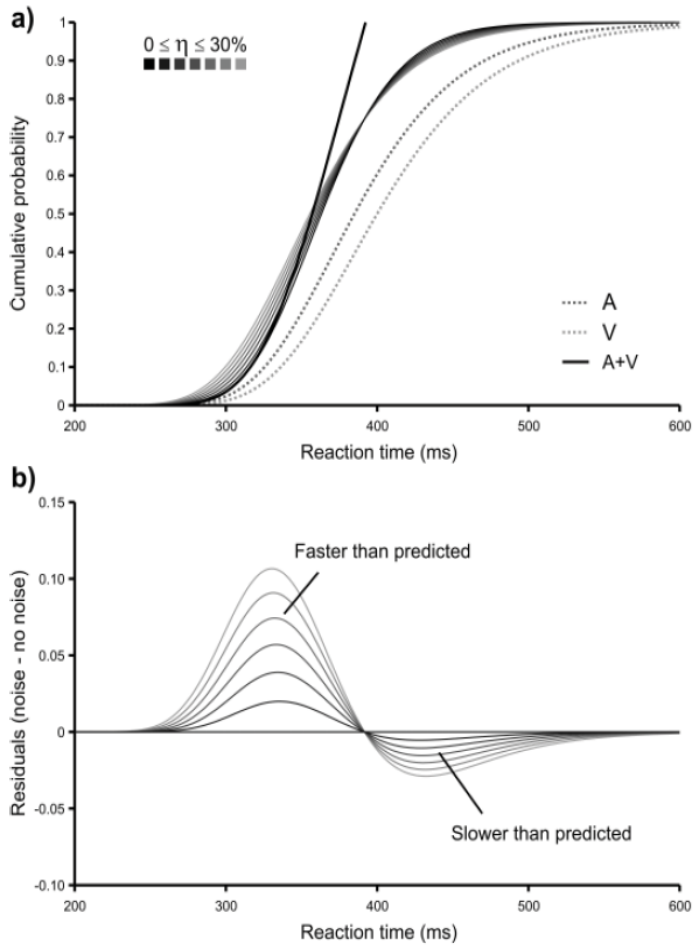


Figure 3.11 : *Figure tirée de Otto et Mamassian, manuscrit non publié.(Otto and Mamassian)*

a) Prédictions du Race model avec ajout de différentes valeurs de bruit η
b) Déviations entre le Race model non bruité et le Race model bruité

Toutefois ce modèle a été construit pour des tâches de détection où les performances des sujets atteignent 100% dans toutes les modalités. De fait, la facilitation audiovisuelle n'est visible qu'en termes de temps de réaction, mais pas de taux de réussite. Notre expérience impliquait une tâche avec une charge cognitive plus élevée puisqu'il s'agissait de catégoriser des stimuli. Les sujets ont correctement réalisé cette tâche, et bien qu'ayant des faibles taux de réussite dans la modalité visuelle. Ils n'ont donc pas seulement détecté les évènements mais ont correctement catégorisé les stimuli. De plus, dans notre étude, la facilitation audiovisuelle s'observe aussi bien en termes de temps de réaction qu'en taux de réussite. Enfin, nous avons choisi de calculer le d' cumulé, qui donne une mesure fiable de la détectabilité d'un signal dans du bruit, dans notre étude le bruit serait les distracteurs. Sur la figure suivante, le signal A serait nos distracteurs et le signal B, les cibles. On applique un bruit gaussien autour de ces deux signaux. Le d' mesure la distance entre les sommets de chacune des gaussiennes, divisé par l'écart-type du bruit.

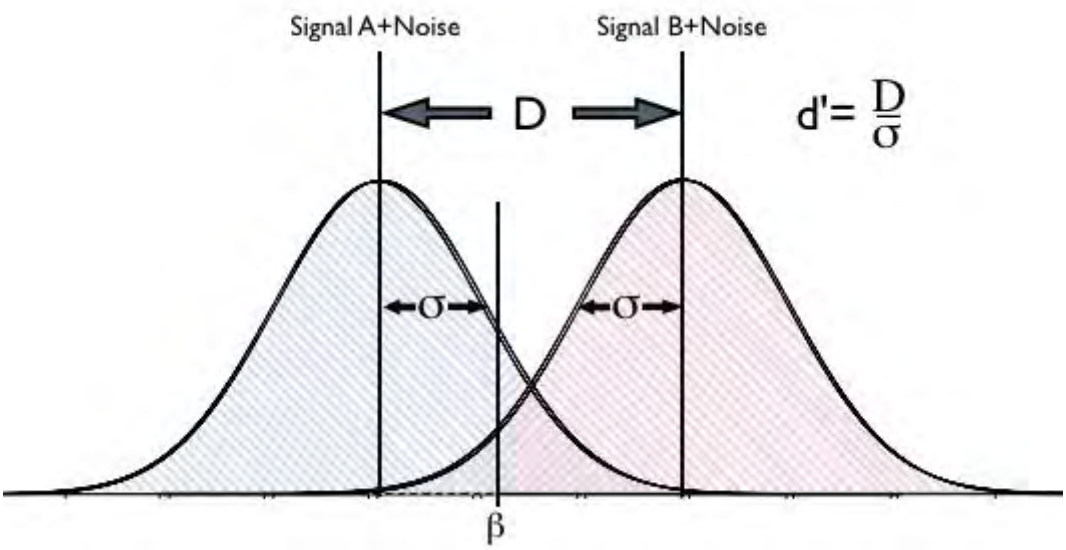


Figure 3.12: Modèle du d'

Le modèle d'Otto et Mamassian pourrait expliquer les données que nous obtenons en ce qui concerne les hits, notamment pour la catégorie « Instrument de musique », et dans une moindre mesure, puisque les déviations du Race model ne sont pas significatives, pour les essais incongruents. Toutefois leur modèle ne peut expliquer les résultats obtenus pour la catégorie « Animal », puisque nos données empiriques sont toujours supérieures ou égales aux réponses prédites. De plus, si le bruit augmente concernant les hits dans la condition audiovisuelle, il doit augmenter également pour les fausses alarmes. Donc le calcul du d' nous permet de prendre en compte le bruit aussi bien pour les données empiriques que pour les données prédites. Et quelle que soit la catégorie, nous observons une violation du Race model, a minima pour les temps de réaction les plus précoces.

Nous ne pouvons donc éliminer avec certitude l'hypothèse du bruit du modèle d'Otto et Mamassian concernant le taux de hits obtenu dans notre expérience pour la catégorie Instruments de musique, mais cette hypothèse ne peut expliquer intégralement les résultats obtenus pour l'autre catégorie et pour le d'.

La différence de facilitation multi-sensorielle entre les catégories pourrait s'expliquer par une asymétrie lors de la création des stimuli. Nous avons constaté que les sujets avaient des performances moindres en modalité visuelle pour la catégorie Animal. Or nous avons volontairement choisi des images non canoniques des items afin de maximiser le taux

d'erreurs. Toutefois, les images prises isolément, et même intégrées au sein de séquences SWIFT (avant superposition) sont aisément identifiables dans les deux catégories. Il est donc difficile de prévoir a priori les performances des sujets. Nous pourrions répliquer cette expérience en choisissant des images jugées plus difficiles pour la catégorie Instrument de musique. Nous pourrions ainsi vérifier si la facilitation multi-sensorielle est d'autant plus importante que les stimuli dans chacune des modalités sont malaisément identifiables. Cette hypothèse semble écologiquement valable, et en accord avec le principe *d'Inverse Effectiveness* (Holmes, 2009) : l'apport d'informations dans différentes est d'autant plus bénéfique que les informations prises isolément sont peu aisément identifiables.

Enfin, il serait intéressant de répliquer cette expérience avec de nouveaux sujets en enregistrant conjointement l'électroencéphalogramme afin de mesurer, via la technique des potentiels évoqués, la facilitation multi-sensorielle due à la congruence sémantique.

Discussion générale

I) Résumé des résultats

Dans cette thèse nous avons donc exploré différents aspects de la catégorisation visuelle et audio-visuelle en nous attachant à distinguer les aspects de haut niveau et de bas niveau des stimuli. Nous avons tout d'abord tenté d'isoler les corrélats neuraux de catégorisation d'images en tant que cibles de la tâche chez le singe. Afin de nous focaliser sur l'activité liée à la reconnaissance du contenu sémantique des stimuli, nous avons égalisé plusieurs paramètres de bas niveau des stimuli (tels que la luminance globale, les contrastes globaux et locaux et les fréquences spatiales) grâce à la technique SWIFT. Nous avons ainsi montré une activité électrophysiologique tardive et maintenue dans l'ensemble des aires visuelles enregistrées, directement corrélée à la réponse comportementale de l'animal. Nous avons également démontré que cette activité ne dépendait pas des caractéristiques physiques des stimuli et ne pouvait pas s'expliquer non plus par la réponse motrice du singe.

Dans une seconde partie nous avons voulu déterminer chez l'homme et chez le singe, si des caractéristiques de bas niveau du contexte visuel pouvaient interférer avec cette capacité de catégorisation. Pour cette étude nous avons donc créé des arrière-plans bruités, sans contenu sémantique, avec différents niveaux de naturalité. Les caractéristiques de bas niveau manipulées étaient cette fois le spectre d'amplitude de la transformée de Fourier des images d'origine. Nous avons comparé les performances de catégorisation de vignettes collées soit sur un arrière-plan bruité soit sur une photographie de scène réelle. Ce faisant nous avons pu évaluer la part de l'effet de congruence contextuelle portée par le spectre d'amplitude des scènes. Nos résultats semblent indiquer que l'homme et le singe adoptent des stratégies différentes : les sujets humains ne sont pas sensibles à la congruence si le contexte n'a pas de contenu sémantique, tandis que les singes montrent un effet similaire de congruence quel que soit la nature du contexte (i.e. réel ou bruité). Toutefois nous avons voulu nuancer ces résultats par l'expérience perceptuelle de nos deux espèces : si les sujets humains ont été confrontés dans la réalité à de nombreuses et très diverses situations, les singes quant à

eux, ont toujours vécu en captivité, dans un environnement sensoriel assez pauvre. Il est possible que la stratégie différente adoptée par les singes résulte davantage d'une exposition répétée à des images sur écran, et que des singes qui auraient eu une vie sauvage adoptent une stratégie plus proche de l'homme.

Enfin dans notre dernière étude, chez l'homme, nous avons également exploré les effets de congruence, mais cette fois entre deux modalités sensorielles en tentant d'isoler les composantes sémantiques des stimuli. Dans le domaine visuel nous avons réutilisé la technique SWIFT, et dans le domaine auditif nous avons randomisé des fragments de sons pour créer un bruit au sein duquel nous avons inséré les sons originaux. Nous avons pu ainsi mettre en évidence une facilitation audiovisuelle pour les essais congruents, plus ample que l'effet prédit par le Race model. L'effet de facilitation était d'autant plus important que les stimuli étaient difficilement perçus en condition unimodale. Bien que le Race model soit aujourd'hui remis en question par certains auteurs (Otto and Mamassian, 2012), les violations que nous obtenons ne semblent pas pouvoir s'expliquer par la théorie que ces auteurs ont développée.

II) La contribution des statistiques de haut et bas niveaux chez l'homme et le singe

Les caractéristiques de haut niveau (i.e. sémantiques) et de bas niveau des stimuli ont été explorées séparément tout au long de cette thèse. Dans des images ou des sons réels, les deux sont nécessairement imbriqués. Et hélas, nous ne pouvons donc pas les isoler de manière symétrique ! En effet grâce à des techniques de *scrambling* nous pouvons détruire l'information sémantique des stimuli pour ne garder que certaines statistiques physiques des images (ce que nous avons fait dans le chapitre 2), mais la seule alternative dont nous disposons pour isoler le contenu sémantique des stimuli est d'égaliser autant que possible les statistiques des images (comme nous avons tenté de le faire dans les chapitres 1 et 3).

A) Le « haut niveau » peut-il tout expliquer chez l'homme ?

Est-il raisonnable d'imaginer que le cerveau humain ne traite que les composants des stimuli portant une information sémantique et ignore les caractéristiques de bas niveau ? Cela semble exagéré. Notre hypothèse dans le second chapitre était qu'en modulant le spectre d'amplitude d'images chimériques on pourrait moduler la performance de catégorisation de nos sujets. On supposait alors que certaines caractéristiques de bas niveau visuel pouvaient interférer dans une tâche dont la charge cognitive est relativement élevée. Nous avions formulé cette hypothèse à partir de travaux antérieurement réalisés dans le laboratoire (Joubert et al., 2009) ou par d'autres équipes de recherche (Guyader et al., 2004). L'étude de Joubert et collègues suggérait que le système visuel utilisait le spectre d'amplitude d'une image pour accélérer la catégorisation, tandis que l'étude de Guyader et collègues montrait que lors d'une tâche de catégorisation de scène, l'effet de *priming* ne s'observait que lorsque le *prime* et la scène à catégoriser partageaient le même spectre d'amplitude. De plus, des études computationnelles avaient désigné le spectre d'amplitude comme la statistique la plus représentative de la catégorie de la scène (Torralba and Oliva, 2003).

Figure D.1 : Intensité moyenne et signature spectrale d'un set de scènes selon les objets qu'elles contiennent (tirée de Torralba et Oliva, 2003)

Cette étude avait ainsi montré que le spectre d'amplitude d'une scène était un bon prédicteur de la présence ou non d'un item, et que des algorithmes de catégorisation utilisant cette statistique avaient d'excellentes performances.

Nos résultats, bien que n'infirmant pas ces travaux, montrent que le système visuel humain ne fonctionne pas comme un algorithme de catégorisation.

Il est donc possible que les statistiques de bas niveau, seules, ne fournissent pas d'information suffisante à notre cerveau pour que nous puissions les analyser.

Dans la troisième étude de cette thèse nous avons montré que le gain multisensoriel s'expliquait (au moins en partie) par les informations sémantiques des stimuli. Dans notre expérience, toutefois, nous n'avions aucun essai purement unisensoriel puisque les deux modalités étaient toujours stimulées. Cependant grâce aux résultats obtenus par Werner et Noppeney (Werner and Noppeney, 2010), nous pouvons raisonnablement supposer que l'ajout de bruit, auditif ou visuel, ne fait pas varier significativement les performances. En effet dans leur étude, ils comparaient les performances de catégorisation « outils » versus « instrument de musique » de sujets humains dans deux conditions « unimodales ». Dans le premier cas, seule une modalité sensorielle était stimulée, dans le second, ils ajoutaient du bruit dans la seconde modalité (A versus An dans le domaine auditif, V versus Vn dans le domaine visuel sur la figure D.2 ci-dessous).

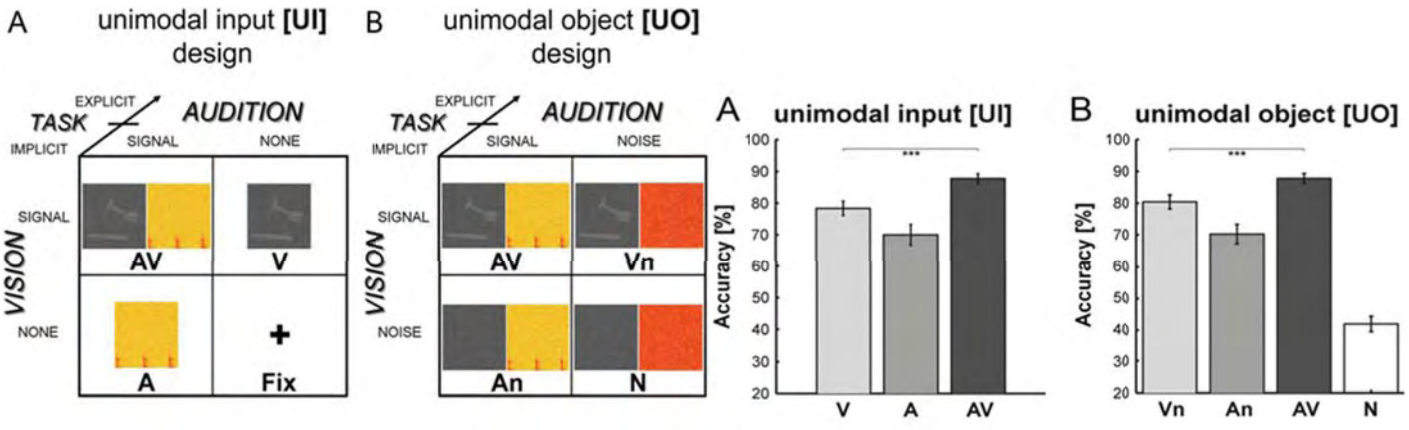
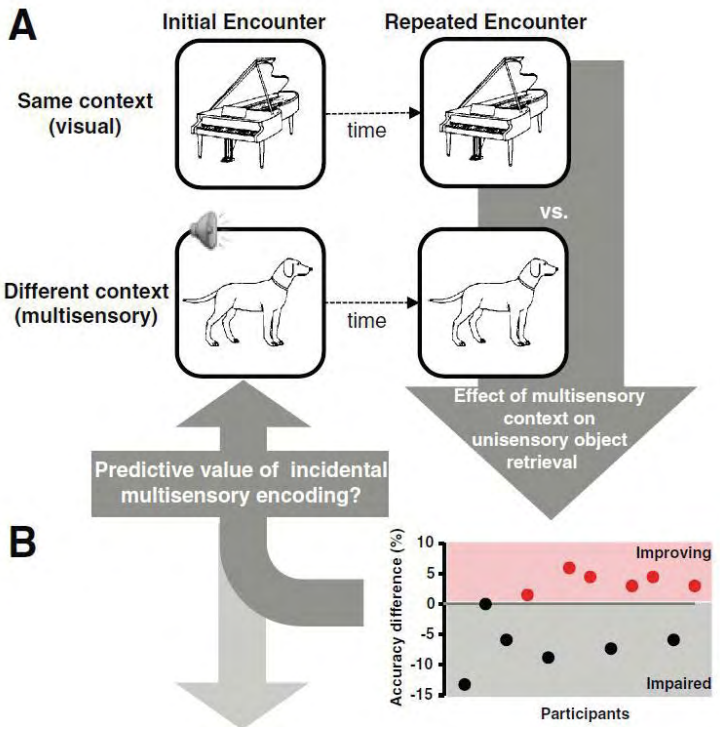


Figure D.2: Stimuli utilisés et résultats obtenus par Werner et Noppeney, 2010

L'ajout de bruit ne modifie pas les performances des sujets à la tâche de catégorisation (une augmentation de 2% des performances en condition Vn par rapport à V peut être vue, mais elle est négligeable). Nous pouvons donc supposer qu'il en irait de même avec

notre protocole, et que les sujets auraient des taux de réussite similaires dans une tâche où seule l'une des deux modalités serait stimulée en comparaison avec nos essais unimodaux où seule une modalité dispense l'information sémantique (bien que les deux modalités soient stimulées).

Dans une série d'études Murray et collaborateurs ont exploré l'effet de la stimulation d'une deuxième modalité sensorielle dans une tâche d'apprentissage/recollection d'objets (Murray et al., 2004, Lehmann and Murray, 2005, Thelen and Murray, 2013, Thelen et al., 2015a, Thelen et al., 2015b). La tâche consistait pour les sujets à dire si un stimulus (une image ou un son) était nouveau ou avait déjà été présenté. La première présentation d'un stimulus pouvait être soit unisensorielle, soit multisensorielle congruente (image de chien+son d'aboiement), multisensorielle incongruente (image de chien+son de cloche) ou multisensorielle neutre (image de chien +*pure tone* ou carrés en niveaux de gris + aboiement). La seconde présentation du stimulus était toujours unisensorielle. La deuxième modalité n'était donc jamais pertinente pour la tâche. Ils ont ainsi montré que seule la présentation d'un stimulus multisensoriel congruent présentait systématiquement un avantage pour la phase de rappel du stimulus unisensoriel, et cela uniquement en termes de taux de réussite, mais pas de temps de réaction. Dans une brève communication en 2015, Thelen et collègues ont rapporté des résultats pour la condition image+*pure tone.*



Ils montrent ainsi que la moitié de leurs sujets présentent un effet de facilitation tandis que les autres voient leurs performances dégradées lorsque la première présentation du stimulus était accompagnée d'un *pure tone*.

Figure D.3 : Protocole et effet de facilitation ou dégradation multisensorielle chez les participants, extrait de Thelen et al, 2015

Mais nous pouvons supposer que cet effet de gain ou de dégradation multisensoriel(le) est dans ce cas davantage dû à un biais attentionnel, le bip ayant attiré l'attention ou distrait le participant par effet de surprise. Nous pouvons également émettre l'hypothèse que ces effets, puisque non cohérents entre participants, ne sont dus qu'à du bruit.

Dans notre expérience, les sujets devaient être attentifs aux deux modalités, et catégoriser l'objet sémantique, et puisque les deux modalités étaient toujours stimulées, il ne pouvait y avoir d'effet de « surprise » dû à l'apparition d'un stimulus dans l'une ou l'autre des modalités. De plus, ayant pris soin d'égaliser un maximum de paramètres physiques dans les stimuli, nous pouvons croire que l'effet que nous obtenons est bel et bien purement dû à la congruence sémantique, et donc aux caractéristiques de haut niveau de nos stimuli.

Si, dans l'état actuel des connaissances en perception visuelle et auditive, nous ne pouvons exclure totalement une contribution des paramètres de bas niveau des stimuli lors de leur catégorisation par l'être humain, elle semble très faible.

B) Les singes sont-ils davantage sensibles au « bas niveau » ?

Comme nous venons de le voir, l'Homme semble être très peu influencé par les caractéristiques de bas niveau des stimuli et se focalise sur leur signification. Et les primates non humains ? Ils sont, à n'en pas douter, capable de reconnaître également le contenu sémantique des stimuli auxquels ils sont confrontés ! Pour preuve, quelles que soient les conditions dans lesquelles ils ont dû catégoriser les images, ils ont toujours réussi la tâche. Un autre argument pour penser qu'ils associaient les images avec l'objet qu'elles représentaient vient des observations que l'on a pu effectuer pendant les sessions d'expérimentation : tous nos singes montraient des signes de peur ou de dégoût lorsque leur était présentées des images d'araignée et/ou de serpent. Une peur instinctive, et que l'on pourrait qualifier d'archaïque, puisqu'ils n'avaient jamais été confrontés à ces animaux dans la réalité. Quoi qu'il en soit, l'image pour eux avait du sens, et dans ce cas représentait une menace.

Dans notre première étude, en outre, le singe répondait correctement à la tâche. Ses réponses étaient totalement corrélées à l'apparition de l'*onset* sémantique dans les séquences, bien que bon nombre de paramètres de bas niveau aient été égalisés. Le singe basait donc bien sa réponse sur le contenu sémantique du stimulus et non sur ses paramètres de bas niveau. Les primates non humains ne catégorisent pas comme des algorithmes, et utilisent les caractéristiques de haut niveau des stimuli.

Toutefois notre seconde étude suggère qu'ils seraient plus sensibles aux paramètres physiques des stimuli que les hommes. Une étude de Einhäuser et collègues en 2006 (Einhäuser et al., 2006) avait mis en évidence un phénomène similaire. Ils ont comparé les mouvements des yeux de deux macaques rhésus et de sept sujets humains lors d'une tâche d'exploration libre de photographies de scènes naturelles. Ces photographies pouvaient être soit intactes, soit modifiées en contraste (de -60% à +100% par pas de 20%) en six points choisis aléatoirement. Ils ont ainsi montré que pour les images intactes, l'homme et le singe étaient attirés de manière similaire par les points les plus saillants de l'image. Mais lorsqu'on augmente le contraste en certains points, les singes vont davantage les fixer que les hommes. Les sujets humains continueraient ainsi à focaliser leur attention sur les zones donnant du sens à l'image, tandis que les singes sont

attirés par les zones les plus contrastées. Kayser et collègues ont confirmé ce résultat en introduisant des patchs de *pink noise* dans des scènes naturelles lors d'une tâche d'exploration d'image chez le macaque uniquement(Kayser et al., 2006). Si les patchs avaient le même contraste que le reste de l'image, les animaux avaient tendance à les éviter et à se concentrer sur les endroits les plus saillants de la scène. Si par contre leur contraste était augmenté par rapport au reste de l'image, les animaux les exploraient plus longtemps que prédit par le hasard. Dans cette étude, il est à noter que les patchs de bruit étaient bien plus nombreux que dans celle d'Einhäuser, altérant fortement l'image. De plus elle ne fournit aucun élément de comparaison avec l'homme. Comme nous, Einhäuser et collègues se sont interrogés sur la raison de cette différence de stratégie entre l'homme et le singe, évoquant la possibilité de l'exposition répétée à des images pour les animaux. Toutefois ils spécifient dans leur étude que les animaux n'avaient jamais été impliqués dans une expérience utilisant des photographies de scènes naturelles, contrairement à nos singes. Ils mentionnent également la pauvreté de l'environnement visuel des animaux de laboratoire comme explication plausible.

Dans l'état des connaissances actuelles tant sur la perception visuelle que sur les capacités cognitives des primates non humains, nous ne pouvons que faire des suppositions sur la cause de cette divergence. Réside-t-elle dans une différence intrinsèque des processus perceptuels ? Ou bien, et c'est l'hypothèse qui semble la plus probable, elle serait due à une différence d'expérience. On pourrait imaginer que le système visuel, par défaut, traiterait les statistiques de bas niveau des images indépendamment de leur localisation et de l'information sémantique qu'elles supportent. Mais avec la diversité des expériences de vie, nous nous concentrons sur les éléments non pas les plus saillants en termes de contraste et luminance, mais les plus informatifs sémantiquement. Les animaux de laboratoire n'ont fait face qu'à un environnement visuel très appauvri : couleurs uniformes et invariantes (les murs des laboratoires ne jaunissent pas en automne), très peu de formes différentes (les barreaux des cages, les formes simplistes des jouets pour animaux, les écrans), et en général ils n'ont côtoyé que des congénères et des humains. Leur expérience du monde est limitée, et celle de la nature réduite à néant. Quel sens peuvent-ils alors associer aux images que nous leur présentons ? En aucun cas nous ne voulons ici remettre en question la capacité de

catégorisation et d'abstraction des primates non humains. Mais les animaux nés et ayant toujours vécu dans l'environnement aseptisé et monotone des laboratoires ont pu développer des stratégies basées sur le bas niveau. Il faudrait pouvoir soumettre des animaux ayant une plus grande expérience visuelle à cette tâche afin d'avoir des éléments de comparaison plus fiables. Nous pourrions par exemple utiliser les singes de la station de primatologie de Rousset. Ils vivent en extérieur, bien que captifs, et nous pourrions adapter le dispositif mis en place avec des babouins (*Papio papio*) par Joël Fagot aux macaques. Les singes choisissent librement d'aller travailler, les contacts avec l'expérimentateur (et donc le stress) sont réduits au minimum. Si ces animaux développent la même stratégie que nos macaques, nous pourrions en tirer des conclusions sur leur fonctionnement cognitif. Dans le cas contraire, il ne pourrait s'agir que d'un biais lié aux conditions de vie des primates en laboratoire.

III) Ouvertures

A) Apport des neurosciences cognitives comparées et la question de la généralisation

Dans ce travail nous avons étudié les facultés de catégorisation des singes. L'intérêt de travailler sur l'animal, au delà de l'évaluation des capacités cognitives des animaux per se, est de pouvoir généraliser les résultats trouvés au fonctionnement du cerveau humain. Notamment le travail avec l'animal permet des explorations plus invasives du système nerveux. Il est alors tentant, d'autant plus quand on s'intéresse aux fonctions perceptives, de généraliser et de supposer que le cerveau humain fonctionne comme celui d'autres primates. Toutefois, notre deuxième étude a montré qu'il fallait être prudent : si dans la globalité l'effet de congruence contextuelle est le même chez l'homme et chez le singe, il semble que les mécanismes le sous-tendant soient différents. D'autres auteurs avant nous avaient également pointé des divergences entre l'homme et le singe en termes de processus cognitifs, dans la majorité des cas, en faveur de l'homme. En 1997 Fagot et Deruelle ont montré par exemple que la préséance du global sur le local démontrée chez l'homme (Navon, 1977) ne se retrouvait pas chez le babouin (Fagot and Deruelle, 1997). En effet si les hommes sont meilleurs pour reconnaître une grande lettre formée de

petites lettres, et s'ils ont davantage de difficulté à identifier les petites lettres qui forment la grande, l'effet est complètement inverse chez le babouin ! Les systèmes visuels du singe et de l'homme sont pourtant très proches. Et l'hypothèse de Navon sur cet effet de préséance se plaçait au niveau des mécanismes perceptuels : l'aspect global d'un objet serait plus saillant que ses attributs locaux pour le système visuel. Or Fagot et Deruelle montrent l'inverse chez le babouin, deux hypothèses alors se dessinent : soit les mécanismes de la perception sont différents chez les deux espèces, soit il s'agit d'un processus cognitif plus que perceptif. Encore une fois, cette seconde hypothèse semble plus plausible tant les systèmes sont proches aux niveaux anatomique et cellulaire.

Une autre étude comparant les capacités de mémoire de travail de l'homme et du chimpanzé a montré la supériorité indubitable de nos plus proches cousins dans ce domaine (Inoue and Matsuzawa, 2007) ! Cette expérience, très médiatisée, consistait en l'apparition très brève des chiffres de 1 à 9 à des positions aléatoire sur un écran, remplacés par des carrés blancs. Le sujet devait alors toucher les carrés dans l'ordre croissant de la suite.

Figure D.5 : un chimpanzé réalisant la tâche (figure tirée de Inoue et Matzusawa, 2007)

Six chimpanzés et 9 sujets humains ont été entraînés pour cette tâche. Les 3 chimpanzés les plus jeunes étaient plus rapides et plus précis que les hommes. Fait plus surprenant encore, les performances de ces jeunes singes ne décroissaient pas avec le raccourcissement du temps de présentation (de 650ms à 210ms) et restaient stables autour de 80% de réussite. Les sujets humains voyaient par contre leurs performances fortement affectées par le raccourcissement du temps de présentation (chutant à 35% en moyenne pour une présentation de 210ms). Il a été proposé que ces performances hors du commun fussent dues à l'entraînement intensif auquel avaient été soumis les jeunes

chimpanzés. Deux étudiants ont alors été soumis à un entrainement similaire (Cook and Wilson, 2010). Ils ont atteint des performances similaires aux jeunes chimpanzés, rassurant notre ego humain. Sauf que pendant ce temps un des jeunes chimpanzés, ayant continué à s'exercer lui aussi, était capable de réaliser la tâche après un flash de seulement 60ms (REF) et ça aucun humain n'en semble capable aujourd'hui (Humphrey, 2012). Si nos capacités cognitives dépassent en général celles de nos cousins primates, il est des cas où ce sont eux qui ont des capacités supérieures aux nôtres.

L'utilisation des primates non humains dans la recherche en neurosciences a permis de grandes avancées dans notre compréhension du fonctionnement du cerveau, particulièrement dans le domaine de la vision, il faut rester précautionneux quand il s'agit de généraliser à l'homme les résultats trouvés chez le singe.

B) La catégorisation : un prétexte pour étudier la reconnaissance des objets

Dans les trois études de cette thèse, nous avons développé des protocoles expérimentaux basés sur la catégorisation visuelle ou auditive. Catégoriser un objet, qu'il soit auditif ou visuel, revient à le reconnaître à un niveau global. Nous savons désormais que chez l'homme, il est plus rapide de catégoriser au niveau super-ordonné (i.e. animal versus non-animal par exemple) qu'au niveau de base (oiseau versus autres animaux/objets)(Mace et al., 2009). La catégorisation « forcée », c'est-à-dire lorsque l'expérimentateur fixe les catégories à reconnaître au préalable, en opposition à la catégorisation libre où le participant forme lui-même ses catégories, offre de nombreux avantages. Le premier est que l'on peut transposer les protocoles directement aux animaux, nul besoin de maîtriser le langage pour réaliser la tâche. Le second, c'est que l'on a accès à des mécanismes rapides, puisque justement il n'y a pas besoin de verbaliser la réponse à la tâche, elle peut se faire en pressant un bouton. Les réponses ne sont pas biaisées par l'expérience personnelle du sujet. Ceci est très important dans le domaine auditif où la valence émotionnelle des sons est forte (Bergman et al., 2009).

Pourtant, force est de constater que les études de catégorisation auditive « forcée » sont rares ! Contrairement au domaine visuel, où les études en catégorisation rapide sont nombreuses, nous nous sommes heurtés, lors de la mise en place de notre troisième expérience, à une certaine pauvreté bibliographique concernant le temps de présentation minimum pour qu'un son soit identifiable ou catégorisable (Suied et al., 2014), la latence nécessaire pour catégoriser un son, les mécanismes sous-tendant la catégorisation auditive et le codage des catégories auditives (Staeren et al., 2009). Ces rares études nous ont confortés dans l'idée que les mécanismes de la catégorisation auditive et leur rapidité étaient semblables à ce qui se passe dans le domaine visuel. Toutefois il serait profitable d'explorer de manière plus contrôlée la catégorisation des sons, en psychophysique et en électroencéphalographie. De telles expériences seraient rapides et peu coûteuses à mettre en place, et permettraient de fournir un bon étalon pour des études ultérieures.

Mais étudier la catégorisation, c'est en réalité se poser la question de la reconnaissance presque instantanée des objets qui nous entourent, malgré leur infinie variabilité. Regrouper ces objets en catégories est un processus paradoxal : d'un côté on simplifie le problème en réduisant la variabilité, en s'abstrayant de l'unicité de chaque objet, de l'autre, cela suppose une capacité d'abstraction puisque justement la variabilité physique au sein des catégories les plus larges est immense ! Différentes théories ont émergé depuis les années 80 pour tenter d'expliquer comment notre cerveau arrivait à reconnaitre un nombre infini d'objets en très peu de temps. On peut citer les théories sur la reconnaissance d'objets : la théorie de la reconnaissance par composantes (« geons ») qui suppose qu'un objet est la somme de plusieurs parties, de forme simple, et que la combinaison spatiale de ces formes permet d'extraire l'identité de l'objet (Biederman, 1987), le modèle *image-based* qui suppose la mise en mémoire d'un grand nombre d'images, et la comparaison à ces exemplaires lors de la vision d'un nouvel objet. D'autres modèles s'intéressent davantage à la catégorisation : modèle par prototype (Posner and Keele, 1968), les objets sont catégorisés en fonction de leur similarité à un prototype stocké en mémoire, le modèle par *decision-boundary* (Ashby, 2000) où l'espace des représentations est subdivisé en régions correspondant à la réponse du sujet. Mais tous ces modèles sont principalement computationnels et ne rendent pas compte de ce qui se passe dans la réalité. Nous n'en sommes qu'aux balbutiements des découvertes sur la

reconnaissance et la catégorisation des objets. Si dans un premier temps les études ont séparé les deux champs d'investigation (i.e. la reconnaissance et la catégorisation), les théories tendent maintenant à converger. De nombreuses questions subsistent sur les mécanismes impliqués dans la catégorisation. Nous savons que les représentations des objets sont encodées de manière distribuée dans le cortex inférotemporal, et que les patterns d'activation sont similaires au sein d'une catégories (Kriegeskorte et al., 2008a). Toutefois les réponses catégorielles se retrouvent davantage au niveau du cortex préfrontal (Freedman et al., 2002), mais ces réponses pourraient être liées à la décision du sujet plus qu'à ses représentations catégorielles.

La transduction d'un signal visuel ou auditif en une représentation abstraite reste donc mystérieuse. Dans cette thèse, nous avons manipulé différents paramètres des stimuli afin de tenter de déterminer si des caractéristiques physiques de bas niveau pouvaient interférer avec un processus cognitif de haut niveau. Si nous avons découvert que le spectre d'amplitude des stimuli visuels pouvait moduler le processus de catégorisation chez le singe de laboratoire, nous n'avons pas pu mettre en évidence un tel phénomène chez l'homme. Dans notre espèce, il semblerait que les processus de traitement visuel se focalisent sur les informations sémantiquement pertinentes, et ignorent le bas niveau des stimuli, comme une adaptation à un environnement complexe.

Références bibliographiques

(2001) Neuroscience. Sunderland: Sinauer Associates.
Alain C, Arnott SR, Hevenor S, Graham S, Grady CL (2001) "What" and "where" in the human auditory system. PNAS 98:12301-12306.
Arguin M, Bub D, Dudek G (1996) Shape Integration for Visual Object Recognition and Its Implication in Category-Specific Visual Agnosia. Visual Cogn 3:221-275.
Ashby G (2000) A stochastic version of general recognition. J of Mathematicl Psychology 44:310-329.
Astley SL, Wasserman EA (1992) Categorical discrimination and generalization in pigeons: all negative stimuli are not created equal. J Exp Psychol Anim Behav Process 18:193-207.
Aust U, Huber L (2001) The role of item- and category-specific information in the discrimination of people- vs. nonpeople images by pigeons. Anim Learn Behav 29:107-119.
Ballas JA, Mullins T (1991) Effects of context on the identification of everyday sounds. Human performance 4:199-219.
Bar-Yosef O, Nelken I (2007) The effects of background noise on the neural responses to natural sounds in cat primary auditory cortex. frontiers in computational neuroscience 1.
BAR M (2004) Visual objects in context. Nature Neuroscience 5:617-629.
Bergman P, Sköld A, Västfjäll D, Fransson N (2009) Perceptual and emotional categorization of sounds. J Accoust Soc Am 126:3156-3167.
Biederman I (1987) Recognition-by-components: a theory of human image understanding. Psychol Rev 94:115-147.
Biederman I, Mezzanotte RJ, Rabinowitz JC (1982) Scene perception: detecting and judging objects undergoing relational violations. Cogn Psychol 14:143-177.
Bizley JK, Cohen YE (2013) The what, where and how of auditory-object perception. Nat Rev Neurosci 14:693-707.
Boucart M, Despretz P, Hladiuk K, Desmettre T (2008) Does context or color improve object recognition in patients with low vision? Vis Neurosci 25:685-691.
Brainard DH (1997) The Psychophysics Toolbox. Spat Vis 10:433-436.
Cauchoix M, Crouzet S, Fize D, Serre T (2015) Fast ventral stream neural activity enables rapid visual categorization. bioRxiv 017897.
Chen Y-C, Spence C (2010) When hearing the bark helps to identify the dog: Semantically-congruent sounds modulate the identification of masked pictures. Cognition 114:389-404.
Colavita FB (1974) Human sensory dominance. Perception & psychophysics 16:409-412.
Cook P, Wilson M (2010) Do young chimpanzees have extraordinary working memory? Psychonomic Bulletin & Review 17:599-600.
Crick F, Koch C (1990) Towards a neurobiological theory of consciousness. Semin Neurosci 2:263-275.
Crick F, Koch C (1998) Consciousness and neuroscience. Cereb Cortex 8:97-107.

D'Amato MR, Van Sant P (1988) The person concept in monkeys (Cebus apella). J Exp Psychol Anim Behav Process 14:43-55.

Davenport JL, Potter MC (2004) Scene consistency in object and background perception. Psychol Sci 15:559-564.

Deese J (1959) On the prediction of occurrence of particular verbal intrusions in immediate recall. Journal of Experimental Psychology 58:17-22.

Dehaene S, Kerszberg M, Changeux JP (1998) A neuronal model of a global workspace in effortful cognitive tasks. Proc Natl Acad Sci U S A 95:14529-14534.

Del Cul A, Baillet S, Dehaene S (2007) Brain Dynamics Underlying the Nonlinear Threshold for Access to Consciousness. PLoS Biol 5:e260.

Doehrmann O, Naumer MJ (2008) Semantics and the multisensory brain: how meaning modulates processes of audio-visual integration. Brain Res 1242:136-150.

Doehrmann O, Naumer MJ, Volz S, Kaiser J, Altmann CF (2008) Probing category selectivity for environmental sounds in the human auditory brain. Neuropsychologia 46:2776-2786.

Edwards CA, Honig WK (1987) Memorization and "feature selection" in the acquisition of natural concepts in pigeons. learning and motivation 18:235-260.

Einhäuser W, Kruse W, Hoffmann K-P, König P (2006) Differences of monkey and human overt attention under natural conditions. Vision Res 46:1194-1209.

FABRE-THORPE M (2003a) Visual categorization: accessing abstraction in non-human primates. The royal society 358:1215-1223.

Fabre-Thorpe M (2003b) Visual categorization: accessing abstraction in non-human primates. Philos Trans R Soc Lond B Biol Sci 358:1215-1223.

Fabre-Thorpe M, Richard G, Thorpe SJ (1998) Rapid categorization of natural images by rhesus monkeys. Neuroreport 9:303-308.

Fagot J, Deruelle C (1997) Processing of global and local visual information and hemispheric specialization in humans (Homo sapiens) and baboons (Papio papio). J Exp Psychol Hum Percept Perform 23:429-442.

FAGOT J, THOMPSON RKR, PARRON C (2010) How to read a picture: Lessons from nonhuman primates. PNAS107:519-520.

Fize D, Cauchoix M, Fabre-Thorpe M (2011) Humans and monkeys share visual representations. Proc Natl Acad Sci U S A 108:7635-7640.

Freedman DJ, Riesenhuber M, Poggio T, Miller EK (2002) Visual categorization and the primate prefrontal cortex: neurophysiology and behavior. J Neurophysiol 88:929-941.

Gaffan D, Heywood CA (1993) A spurious category-specific visual agnosia for living things in normal human and nonhuman primates. J Cogn Neurosci 5:118-128.

Gaver WW (1993) What in the World Do We Hear?: An Ecological Approach to Auditory Event Perception Ecol Psychol 5:1-29.

Giordano BL, McDonnell J, McAdams S (2010) Hearing living symbols and nonliving icons: category specificities in the cognitive processing of environmental sounds. Brain Cogn 73:7-19.

Girard P, Koenig-Robert R (2011) Ultra-Rapid Categorization of Fourier-Spectrum Equalized Natural Images: Macaques and Humans Perform Similarly. PLoS ONE 6:e16453.

Gould J, Chalupa L, Lindsley D (1974) Modifications of pulvinar and geniculo-cortical evoked potentials during visual discrimination learning in monkeys Electroencephalography and Clinical Neurophysiology 36:639-649.
Gronau N, Neta M, Bar M (2008) Integrated Contextual Representation for Objects' Identities and Their Locations. J Cogn Neurosci 20:371-388.
Gross CG, Bender DB, Rocha-Miranda CE (1969) Visual receptive fields of neurons in inferotemporal cortex of the monkey. Science 166:1303-1306.
Gross CG, Rocha-Miranda CE, Bender DB (1972) Visual properties of neurons in inferotemporal cortex of the Macaque. J Neurophysiol 35:96-111.
Guyader N, Chauvin A, Peyrin C, Herault J, Marendaz C (2004) Image phase or amplitude? Rapid scene categorization is an amplitude-based process. C R Biol 327:313-318.
Gygi B, Kidd GR, Watson CS (2007) Similarity and categorization of environmental sounds. Percept Psychophys 69:839-855.
Gygi B, Shafiro V (2011) The incongruency advantage for environ,ental sounds presented in natural auditory scenes. J Exp Psychol Hum Percept Perform 37:551-565.
Hayes KJ, Hayes C (1953) Picture perception in a home-raised Chimpanzee. Journal of comparative and physiological psychology 46:470-474.
Herrnstein RJ, Loveland DH (1964) Complex visual concept in the pigeon. Science 146:549-551.
Hollingworth A, Henderson JM (1998) Does consistent scene context facilitate object perception? J Exp Psychol Gen 127:398-415.
Holmes NP (2009) The principle of inverse effectiveness in multisensory integration: some statistical considerations. Brain Topogr 21:168-176.
Hubel DH, Wiesel TN (1968) Receptive fields and functional architecture of monkey striate cortex. J Physiol 195:215-243.
Hubel DH, Wiesel TN (1970) Stereoscopic vision in macaque monkey. Cells sensitive to binocular depth in area 18 of the macaque monkey cortex. Nature 225:41-42.
Hubel DH, Wiesel TN (1974) Uniformity of monkey striate cortex: a parallel relationship between field size, scatter, and magnification factor. J Comp Neurol 158:295-305.
Hubel DH, Wiesel TN (1977) Ferrier lecture. Functional architecture of macaque monkey visual cortex. Proc R Soc Lond B Biol Sci 198:1-59.
Hubel DH, Wiesel TN (1979) Brain mechanisms of vision. In: The mind's eye: Readings from Scientific American, vol. 241 (Wolfe, J. M., ed), pp 40-52 New York: Freeman.
Humphrey N (2012) ‘This chimp will kick your ass at memory games – but how the hell does he do it?’. Trends Cogn Sci 16:353-355.
Imbert M (1999) Etude du cortex cérébral des primates: comparaison des aires visuelles chez le macaque et chez l'homme. Primatologie 2:1-28.
Inoue S, Matsuzawa T (2007) Working memory of numerals in chimpanzees. Curr Biol 17:R1004-1005.
JOUBERT OR, FIZE D, ROUSSELET GA, FABRE-THORPE M (2008) Early interference of context congruence on object processing in rapid visual categorization of natural scenes. Journal of Vision 8:1-18.
JOUBERT OR, ROSSELET GA, FIZE D, FABRE-THORPE M (2007) Processing scene context: Fast categorization and object interference. Vision research 47:3286-3297.
Joubert OR, Rousselet GA, Fabre-Thorpe M, Fize D (2009) Rapid visual categorization of natural scene contexts with equalized amplitude spectrum and increasing phase noise. J Vis 9 1-16.

Kaspar K, Hassler U, Martens U, Trujillo-Barreto N, Gruber T (2010) Steady-state visually evoked potential correlates of object recognition. Brain Res 1343:112-121.
Kayser C, Nielsen KJ, Logothetis NK (2006) Fixations in natural scenes: Interactions of image structure and image content. Vision Res 46:2535-2545.
KIANI R, ESTEKY H, MIRPOUR K, TANAKA K (2007a) Object category structure in response patterns of neural population in monkey inferior temporal cortex. Journal of neurophysiology 97:4296-4309.
Kiani R, Esteky H, Mirpour K, Tanaka K (2007b) Object category structure in response patterns of neuronal population in monkey inferior temporal cortex. J Neurophysiol 97:4296-4309.
Kirkpatrick-Steger K, Wasserman EA (1996) The what and the where of the pigeon's processing of complex visual stimuli. J Exp Psychol Anim Behav Process 22:60-67.
Koenig-Robert R, VanRullen R (2012) Isolating and tracking the neural correlates of object recognition. soumis.
Koenig-Robert R, VanRullen R (2013) SWIFT: A novel method to track the neural correlates of recognition. Neuroimage 81:273-282.
Kriegeskorte N, MUR M, Ruff DA, Kiani R, Bodurka J, Esteky H, Tanaka K, Bandettini PA (2008a) Matching categorical object reprensentations in inferior temporal cortex of man and monkey. Neuron 60:1126-1141.
Kriegeskorte N, Mur M, Ruff DA, Kiani R, Bodurka J, Esteky H, Tanaka K, Bandettini PA (2008b) Matching categorical object representations in inferior temporal cortex of man and monkey. Neuron 60:1126-1141.
Kuffler SW (1953) Discharge patterns and functional organization of mammalian retina. J Neurophysiol 16:37-68.
Lamy D, Salti M, Bar-Haim Y (2009) Neural Correlates of Subjective Awareness and

Unconscious Processing: An ERP Study. Journal of Cognitive Neuroscience 21:1435–1446.
Las L, Stern EA, Nelken I (2005) Representation of tone in fluctuating maskers in the ascending auditory system. J Neurosci 25:1503-1513.
Laurienti PJ, Kraft RA, Maldjian JA, Burdette JH, Wallace MT (2004) Semantic congruence is a critical factor in multisensory behavioral performance. Exp Brain Res 158:405-414.
Lebrun N, Clochon P, Etevenon P, Lambert J, Baron JC, Eustache F (2001) An ERD mapping study of the neurocognitive processes involved in the perceptual and semantic analysis of environmental sounds and words. Cogn Brain Res 11:235-248.
Lehmann S, Murray MM (2005) The role of multisensory memories in unisensory object discrimination. Cogn Brain Res 24:326-334.
Logothetis N (1998) Object vision and visual awareness [In Process Citation]. Curr Opin Neurobiol 8:536-544.
Logothetis NK, Sheinberg DL (1996) Visual Object Recognition. Annu Rev Neurosci 19:577-621.
Logothetis NK, Vetter T, Hurlbert A, Poggio T (1994) View-based models of 3D object recognition and class-specific invariances. pp 1-11: MIT - Artificial Intelligence Laboratory.
Loschky LC, Sethi A, Simons DJ, Pydimarri TN, Ochs D, Corbeille JL (2007) The importance of information localization in scene gist recognition. J Exp Psychol Hum Percept Perform 33:1431-1450.

Mace MJ, Joubert OR, Nespoulous JL, Fabre-Thorpe M (2009) The time-course of visual categorizations: you spot the animal faster than the bird. PLoS ONE 4:e5927.
McGurk H, MacDonald J (1976) Hearing lips and seeing voices. Nature 264:746-748.
Miller J (1986) Timecourse of coactivation in bimodal divided attention. Perception & psychophysics 40:331-343.
Miller MB, Gazzaniga MS (1998) Creating false memories for visual scenes. Neuropsychologia 36:513-520.
Molholm S, Ritter W, Javitt DC, Foxe JJ (2004) Multisensory Visual–Auditory Object Recognition in Humans: a High-density Electrical Mapping Study. Cereb Cortex 14:452-465.
Murray MM, Michel CM, Peralta RGd, Ortigue S, Brunet D, Andino SG, Schniderb A (2004) Rapid discrimination of visual and multisensory memories revealed by electrical neuroimaging. NeuroImage 21:125-135.
Navon D (1977) Forest before trees: the precedence of global features in visual perception. Cogn Psychol 9:353-383.
Neider MB, Zelinsky GJ (2006) Searching for camouflaged targets: Effects of target-background similarity on visual search. Vision Res 46:2217-2235.
Nelken I (2008) Processing of complex sounds in the auditory system. Current Opinion in Neurobiology 18:413-417.
Niessen ME, VanMaanen L, Andringa TC (2008) Disambiguating through context. International journal of semantic computing 2.
OLIVA A, TORRALBA A (2001) Modeling the shape of the scene: A holistic representation of the spatial envelope. International journal of computer vision 42:145-175.
OLIVA A, TORRALBA A (2007) The role of context in object recognition. Trends un cognitive sciences 11:520-527.
Otto TU, Mamassian P Noise and (multi-) sensory integration: When Drawbacks mimic benefits. Unpublished manuscript.
Otto TU, Mamassian P (2012) Noise and Correlations in Parallel Perceptual Decision Making. Curr Biol 22:1391-1396.
Özcan E, Egmond RV (2009) The effect of visual context on the indentification of ambiguous environmental sounds. Acta Psychologica 131:110-119.
Özcan E, Van Egmond R (2009) The effect of visual context on the identification of ambiguous environmental sounds. Acta Psychologica 131:110-119.
Palmer SE (1975) The effects of contextual scenes on the identification of objects. Mem Cognit 3:519-526.
Parron C, Call J, Fagot J (2008) Behavioural responses to photographs by pictorially naive baboons (Papio anubis), gorillas (Gorilla gorilla) and chimpanzees (Pan troglodytes). Behav Processes 78:351-357.
Pokorny JJ, de Waal FB (2009) Monkeys recognize the faces of group mates in photographs. Proc Natl Acad Sci U S A 106:21539-21543.
Portilla J, Simoncelli E (2000a) A parametric texture model based on joint statistics of complex wavelet coefficients. International Journal of Computer Vision 40:49-71.
Portilla J, Simoncelli EP (2000b) A parametric texture model based on joint statistics of complex wavelet coefficients. International journal of computer vision 40:49-71.
Posner MI, Keele SW (1968) On the genesis of abstract ideas. Journal of Experimental Psychology 77:353-363.

Remy F, Vayssière N, Pins D, Boucart M, Fabre-Thorpe M (2014) Incongruent object/context relationships in visual scenes: Where are they processed in the brain? Brain Cogn 84:34-43.

Roberts WA, Mazmanian DS (1988) Concept learning at different levels of abstraction by pigeons, monkeys, and people. J Exp Psychol Anim Behav Process 14:247-260.

Rosch E, Mervis CB, Gray WD, Johnson DM, Boyes-Braem P (1976) Basic objects in natural categories. Cogn Psychol 8:382-439.

Rousselet GA, Joubert OR, Fabre-Thorpe M (2005) How long to get to the "gist" of real-world natural scenes? Visual Cogn 12:852-877.

Rousselet GA, Mace MJ, Fabre-Thorpe M (2003) Is it an animal? Is it a human face? Fast processing in upright and inverted natural scenes. J Vis 3:440-455.

Sands SF, Lincoln CE, Wright AA (1982) Pictorial similarity judgments and the organization of visual memory in the rhesus monkey. J Exp Psychol Gen 111:369-389.

Schneider TR, Engel AK, Debener S (2008) Multisensory identification of natural objects in a two-way crossmodal priming paradigm. Exp Psychol 55:121-132.

Schrier AM, Angarella R, Povar ML (1984) Studies of concept formation by stumptailed monkeys: Concepts humans, monkeys, and letter A. J Exp Psychol Anim Behav Process 10:564-584.

Schrier AM, Brady PM (1987) Categorization of natural stimuli by monkeys (Macaca mulatta): effects of stimulus set size and modification of exemplars. J Exp Psychol Anim Behav Process 13:136-143.

Seyfarth RM, Cheney DL, Marler P (1980) Vervet monkey alarm calls: semantic communication in a free-ranging primate. Anim Behav 28:1070-1094.

Staeren N, Renvall H, De Martino F, Goebel R, Formisano E (2009) Sound Categories Are Represented as Distributed Patterns in the Human Auditory Cortex. Curr Biol 19:498–502.

Suied C, Agus TR, Thorpe SJ, Mesgarani N, Pressnitzer D (2014) Auditory gist: Recognition of very short sounds from timbre cues. J Accoust Soc Am 135:1380-1391.

Tanaka K (1996) Inferotemporal cortex and object vision. Annu Rev Neurosci 19:109-139.

Thelen A, Matusz PJ, Murray MM (2015a) Multisensory context portends object memory. Curr Biol 24.

Thelen A, Murray MM (2013) The Efficacy of Single-Trial Multisensory Memories. Multisensory Res 26:483-502.

Thelen A, Talsma D, Murray MM (2015b) Single-trial multisensory memories affect later auditory and visual object discrimination. Cognition 138:148-160.

Thompson RK (1995) Natural and Relational Concepts in Animals. In: Comparative Approaches to Cognitive Science (Roitblat, H. L. and Meyer, J. A., eds), pp 175-224 Cambridge, MA: MIT Press.

Thorpe SJ, Fabre-Thorpe M (2001) Neuroscience. Seeking categories in the brain. Science 291:260-263.

Tian B, Reser D, Durham A, Kustov A, Rauschecker JP (2001) Functional specialization in rhesus monkey auditory cortex. Science 292:290-293.

Torralba A, Oliva A (2003) Statistics of natural image categories. Computation in neural systems 14:391-412.

Truppa V, Spinozzi G, Stegagno T, Fagot J (2009) Picture processing in tufted capuchin monkeys (Cebus apella). Behav Processes 82:140-152.

Tsunada J, Cohen YE (2014) Neural mechanisms of auditory categorization: from across brain areas to within local microcircuits. frontiers in Neuroscience 8.

Tsunada J, Lee JH, Cohen YE (2011) Representation of speech categories in the primate auditory cortex. J Neurophysiol 105 2634-2646.

Vogels R (1999a) Categorization of complex visual images by rhesus monkeys. Part 1: behavioural study. Eur J Neurosci 11:1223-1238.

Vogels R (1999b) Effect of image scrambling on inferior temporal cortical responses. Neuroreport 10:1811-1816.

Warren RM (1970) Perceptual restoration of missing speech sounds. Science 167:392-393.

Wasserman EA, Cook BR, Kirkpatrick-Steger K, Astley SL, Biederman I (1996) The Pigeon's Recognition of Drawings of Depth-Rotated Stimuli. J Exp Psychol Anim Behav Process 22:205-221.

Wasserman EA, Kirkpatrick-Steger K, Van Hamme LJ, Biederman I (1993) Pigeons Are Sensitive to the Spatial Organization of Complex Visual Stimuli. Psychological Science 4:336-341.

Werner S, Noppeney U (2010) Distinct Functional Contributions of Primary Sensory and Association Areas to Audiovisual Integration in Object Categorization. J Neurosc 30:2662–2675.

Whittingstall K, Logothetis NK (2009) Frequency-Band Coupling in Surface EEG Reflects Spiking Activity in Monkey Visual Cortex. Neuron 64:281-289.

Wolfe JM, Oliva A, Horowitz TS, Butcher SJ, Bompas A (2002) Segmentation of objects from backgrounds in visual search tasks. Vision Res 42:2985-3004.

Woodman GF, Kang MS, Rossi AF, Schall JD (2007) Nonhuman primate event-related potentials indexing covert shifts of attention. Proc Natl Acad Sci U S A 104:15111-15116.

Yoshikubo S (1985) Species discrimination and concept formation by rhesus Monkeys. Primates 26:285-299.

Zayan R, Vauclair J (1998) Categories as paradigms for comparative cognition. Behavioural Processes 42:87-99.

Dans cette thèse, nous nous sommes proposé d'explorer les contributions relatives des caractéristiques de haut et de bas niveau des stimuli dans la catégorisation d'objet. Ce travail comporte trois études, chez l'homme et le singe. L'originalité de cette thèse réside donc dans la construction des stimuli. Notre première étude a visé à caractériser les corrélats neuraux de la reconnaissance d'images chez le singe en ECoG. Pour cela nous avons développé un protocole de catégorisation où les stimuli étaient des séquences visuelles dans lesquelles les contours des objets (information sémantique, caractéristique de haut niveau) étaient modulés cycliquement grâce à la technique SWIFT (créée par Roger Koenig et Rufin VanRullen) alors que la luminance, les contrastes et les fréquences spatiales (caractéristiques de bas niveau) étaient conservées. Grâce à une analyse en potentiels évoqués, nous avons pu mettre en évidence une activité électrophysiologique tardive en « tout ou rien » spécifique de la reconnaissance de la cible de la tâche par le singe. Mais parce que les objets sont rarement isolés en conditions réelles, nous nous sommes penchés dans une deuxième étude sur l'effet de congruence contextuelle lors de la catégorisation d'objets chez l'homme et le singe. Nous avons comparé la contribution du spectre d'amplitude d'une transformée de Fourier à cet effet de congruence chez ces deux espèces. Nous avons révélé une divergence de stratégie, le singe semblant davantage sensible à ces caractéristiques de bas niveau que l'homme. Enfin dans une dernière étude nous avons tenté de quantifier l'effet de congruence sémantique multisensorielle dans une tâche de catégorisation audiovisuelle chez l'homme. Dans cette étude nous avons égalisé un maximum de paramètres de bas niveau dans les deux modalités sensorielles, que nous avons toujours stimulées conjointement. Dans le domaine visuel, nous avons réutilisé la technique SWIFT, et dans le domaine auditif nous avons utilisé une technique de randomisation de *snippets*. Nous avons pu alors constater un gain multisensoriel important pour les essais congruents (l'image et le son désignant le même objet), s'expliquant spécifiquement par le contenu sémantique des stimuli. Cette thèse ouvre donc de nouvelles perspectives, tant sur la cognition comparée entre homme et primate non humain que sur la nécessité de contrôler les caractéristiques physiques de stimuli utilisés dans les tâches de reconnaissance d'objets.

Mots clés : Catégorisation, caractéristiques de bas niveau, contenu sémantique, effet de congruence

In this thesis, we explored the relative contributions of high level and low level features of stimuli used in object categorization tasks. This work consists of three studies in human and monkey. The originality of this thesis lies in stimuli construction. Our first study aimed to characterize neural correlates of image recognition in monkey, using ECoG recordings. For that purpose we developped a categorization task using SWIFT technique (technique created by Roger Koenig and Rufin VanRullen). Stimuli were visual sequences in which object contours (semantic content, high level feature) were cyclically modulated while luminance, contrasts and spatial frequencies (low level features) remained stable. By analyzing evoked potentials, we brought to light a late electrophysiological activity, in an « all or none » fashion, specifically related to the target recognition in monkey. But because in real condition objects are never isolated, we explored in a second study contextual congruency effect in visual categorization task in humans and monkeys. We compared the contribution of Fourier transform amplitude spectrum to this congruency effect in the both species. We found a strategy divergence showing that monkeys were more sensitive to the low level features of stimuli than humans. Finally, in the last study, we tried to quantify multisensory semantic congruency effect, during a audiovisual categorization task in humans. In that experiment, we equalized a maximum of low level features, in both sensory modalities which were always jointly stimulated. In the visual domain, we used again the SWIFT technique, whereas in auditory domain we used a snippets randomization technique. We highlighted a large multisensory gain in congruent trials (i.e. image and sound related to the same object), specifically linked to the semantic content of stimuli. This thesis offers new perspectives both for comparative cognition between human and non human primates and for the importance of controlling the physical features of stimuli used in object recognition tasks.

Key words : Categorization, low level features, semantic content, congruency effect