Thèse de Doctorat de l'Université Lumière Lyon 2

Ecole Doctorale de Sciences Cognitives

Présentée par Julien Besle

Pour obtenir le grade de Docteur de l'Université Lyon 2

Spécialité : Sciences Cognitives - Mention : Neurosciences

Interactions audiovisuelles dans le cortex auditif chez l'homme

Approches électrophysiologique et comportementale

Soutenance publique le 22 mai 2007 devant le jury composé de :

M^{r} Pascal Barone (Examinateur)
M^{me} Nicole Bruneau (Rapporteur)
M^{r} Jean-Luc Schwartz (Rapporteur)
M^{me} Marie-Hélène Steiner-Giard (Directrice de thèse)
M^{r} Rémy Versace (Examinateur)

Table des matières

Première partie

Revue de la littérature

Chapitre 1

Premières études neurophysiologiques de la convergence audiovisuelle

Dans les études récentes sur les interactions audiovisuelles, et multisensorielles en général, il est fait mention d'un modèle "classique" de l'organisation des différents systèmes sensoriels dans lequel les informations des différentes modalités sont élaborées indépendamment avant de converger dans des aires corticales dites associatives (voir par exemple Calvert, 2001). Dans ce premier chapitre, nous passerons en revue des études qui ont cherché à définir les aires de convergence, surtout chez l'animal, sur des critères électrophysiologiques ou anatomiques. Nous verrons que lorsqu'on considère l'ensemble de ces études, ce modèle de convergence tardive ne s'impose pas de manière évidente.

1.1 Définition des aires corticales associatives en électrophysiologie

La question de la convergence des informations de plusieurs modalités sensorielles est abordée dès les premières études électrophysiologiques du cortex cérébral, principalement chez le chat, à l'aide de deux techniques électrophysiologiques. Dans la première, on recueille l'activité globale de populations de neurones à la surface du cortex de l'animal, alors que dans la seconde, on enregistre directement les potentiels d'action de cellules individuelles du cortex.

1.1.1 Études électrocorticographique (ECoG) de la convergence multisensorielle

Dans les études ECoG, les aires cérébrales de convergence sont tout d'abord définies comme les régions du cortex dans lesquelles on trouve des réponses associatives à des stimuli de plusieurs modalités. Une réponse associative se définit en général par opposition à une réponse primaire unisensorielle. Ainsi Buser et Rougeul (1956) définissent une réponse associative comme toute réponse enregistrée hors du cortex primaire, de latence plus longue et de variabilité plus grande que la réponse primaire. Si les premières réponses

associatives découvertes sont unisensorielles, on va découvrir plusieurs aires corticales répondant aussi bien à des stimulations visuelles, auditives que somesthésiques. Thompson, Johnson et Hoopes (1963) réalisent ainsi des enregistrements ECoG sur une grande partie du cortex de chats anesthésiés et définissent 4 zones polysensorielles : le gyrus suprasylvien antérieur (AMSA), le gyrus suprasylvien postérieur (PMSA), qui se trouvent tous deux entre les aires auditives et visuelles, l'aire latérale antérieure (ALA), située en arrière du cortex somesthésique primaire, et l'aire péricruciée, médiale par rapport au cortex moteur primaire. Ces aires associatives sont illustrées dans la figure 1.1.

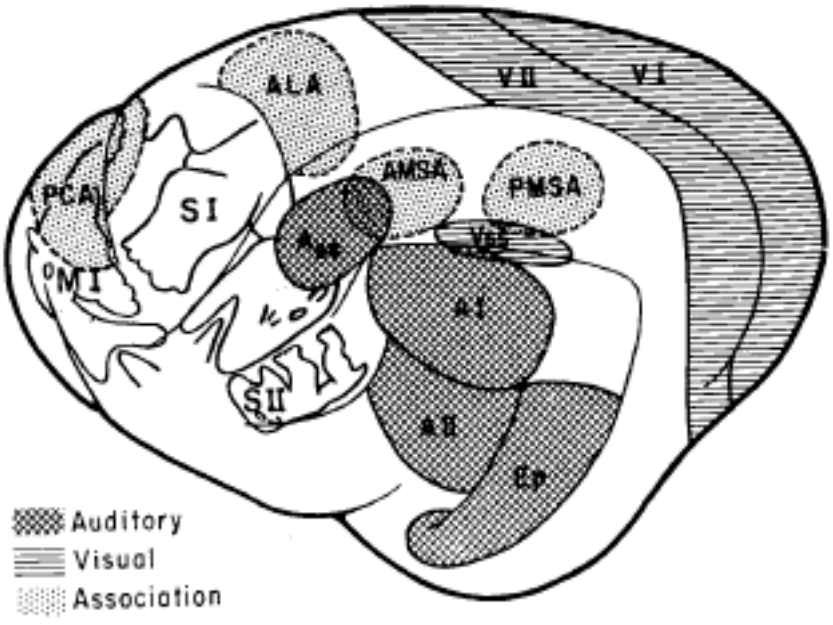

FIG. 1.1 – Localisation des aires unisensorielles et associatives chez le chat. AI : cortex auditif primaire ; AII : cortex auditif secondaire ; ALA : aire latérale antérieure ; AMSA : gyrus suprasylvien antérieur ; MI : cortex moteur primaire ; PCA : aire péricruciée ; PMSA : gyrus suprasylvien postérieur ; SI : aire somesthésique primaire ; VI : cortex visuel primaire ; VII cortex visuel secondaire ; VSS : Aire visuelle du sillon suprasylvien. D'après Thompson, Johnson et Hoopes (1963).

Les réponses dans ces aires semblent présenter les propriétés des aires associatives, dont, par exemple, une latence plus longue (35 ms après la stimulation contre 15 ms en moyenne dans le cortex visuel primaire). Dans une autre étude, Thompson, Smith et Bliss (1963) montrent, en outre, que les réponses associatives à une stimulation donnée ne sont pas corrélées aux réponses évoquées par la même stimulation dans le cortex primaire correspondant.

Afin de montrer que ces zones sont bien des zones de convergence multisensorielle, un autre critère, lié aux propriétés réfractaires des cellules nerveuses va être utilisé : l'idée est que si les informations en provenance de différentes modalités convergent vers la même population neuronale, alors la réponse à un stimulus suivant un autre stimulus devrait diminuer ou disparaitre en raison de la période réfractaire des neurones. Thompson, Smith et Bliss (1963) testent donc les réponses des aires primaires et polysensorielles à des paires

de stimulations successives de même modalité ou de modalités différentes : le résultat est que la période réfractaire des zones polysensorielles est beaucoup plus longue (il faut presque une seconde de délai pour obtenir une seconde réponse d'amplitude égale à la première) que celle des cortex sensoriels et surtout qu'elle est à peu près la même quelles que soient les modalités impliquées et que la paire soit intramodale ou intermodale. Leur conclusion est donc que les informations de différentes modalités convergent vers des cellules communes des zones polysensorielles et évoquent une réponse identique.

Notons que, dans cette étude, le délai entre les deux stimulations est choisi de façon à ce que les réponses aux deux stimuli ne se chevauchent pas (200 ms minimum pour les aires multisensorielles), si bien qu'il n'est pas question ici de stimulation réellement bimodale. De façon intéressante, l'ablation de la quasi totalité du cortex, à l'exception de ces aires associatives polysensorielles, ne supprime pas la réponse associative, ce qui suggère qu'elles reçoivent leurs entrées de zones sous-corticales.

À l'inverse, Thompson, Smith et Bliss (1963) montrent que le sillon suprasylvien (VSS dans la figure 1.1 page précédente) n'est pas une aire de convergence multisensorielle mais une aire associative spécifique au traitement visuel puisque la réponse présente une période réfractaire pour des paires de stimuli visuels, mais pas pour des paires de stimuli de deux modalités différentes (en l'occurrence audiovisuelles). Pour calculer cette période réfractaire aux délais les plus courts (5 à 40 ms), ils recourent à une analyse algébrique, dont le principe est illustré dans la figure 1.2 page suivante, qui sera reprise par beaucoup d'études multisensorielles par la suite, et qui est à la base du modèle additif utilisé dans l'analyse des interactions multisensorielles en potentiels évoqués (voir partie 7.2.1 page 107).

Utilisant cette méthode, Thompson, Smith et Bliss (1963) montrent que l'amplitude de la réponse à des paires audiovisuelles de stimuli est égale à la somme des amplitudes des réponses à des stimuli auditifs et visuels présentés séparément et concluent à une indépendance des populations neuronales générant ces réponses dans le cortex visuel du sillon supra-sylvien. Aucune tentative n'est cependant faite pour tester statistiquement la différence. Récemment, Yaka, Notkin, Yinon et Wollberg (2000) ont en effet rapporté l'existence de cellules répondant à la fois à des stimuli auditifs et visuels dans cette structure.

Les études de Thompson, Johnson et Hoopes (1963) et Thompson, Smith et Bliss (1963) suggèrent que les réponses dans les cortex associatifs polysensoriels sont totalement indifférenciées (non spécifiques), identiques d'une aire à l'autre et pourraient être dues à une convergence au niveau sous-cortical (avec l'idée que ces afférences non spécifiques court-circuiteraient les aires primaires).

Parmi les 4 aires associatives de convergence multisensorielle ainsi mises en évidence, le gyrus suprasylvien va être plus particulièrement étudié. Utilisant la même méthodologie, Rutledge (1963) trouve une asymétrie de la période réfractaire selon que la paire intermodale est auditivo-visuelle ou visuo-auditive (La période est de 150 ms dans le premier cas et de 400 ms dans le second), ce qui contraste avec l'homogénéité des périodes réfractaires rapportée par Thompson, Smith et Bliss (1963). Selon Rutledge (1963), ce résultat indiquerait une prédominance visuelle relative du gyrus suprasylvien du chat. Dans une tentative de réconcilier les deux résultats, A. S. Schneider et Davis (1974) comparent les périodes

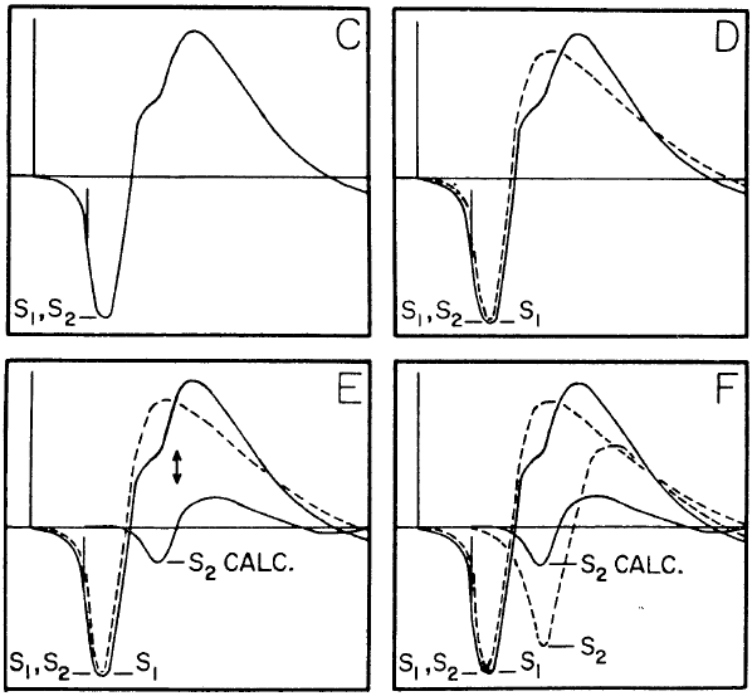

FIG. 1.2 – Illustration de la méthode algébrique utilisée par Thompson, Smith et Bliss (1963). Le but est de savoir si des réponses à deux stimuli S_1 et S_2 enregistrées sur la même électrode sont générées par deux populations neuronales distinctes : si oui, alors la somme des réponses enregistrées séparément devrait être égale à la réponse enregistrée pour la paire de stimuli S_1,S_2. À cette fin, on calcule la réponse corrigée S_2CALC qui correspond à la différence algébrique de la réponse S_1,S_2 et la réponse à S_1 seul (en tenant compte du délai le cas échéant). Si la réponse au stimulus S_2 n'est pas modifiée par la présentation du stimulus S_1, S_2CALC devrait être égal à S_2. D'après Berman (1961).

réfractaires pour des paires intramodales et intermodales de stimuli : leur conclusion est que, contrairement aux données de Thompson, Smith et Bliss (1963), l'effet réfractaire des réponses du gyrus suprasylvien pour des paires intramodales est plus fort que pour des paires intermodales, ce qui suggère une indépendance relative des entrées des différentes modalités dans cette aire de convergence.

1.1.2 Convergence audiovisuelle au niveau du neurone unitaire

Afin de préciser le caractère convergent des traitements dans le gyrus suprasylvien du chat, plusieurs équipes entreprennent d'enregistrer les réponses unitaires des neurones de cette structure à l'aide de micro-électrodes. Globalement, les cellules répondent de manière plus consistante aux stimulations visuelles (flashs) qu'aux stimulations auditives (clicks) (Bental & Bihari, 1963). Plus important, sur 109 cellules étudiées par Bental et Bihari (1963), 7 répondent significativement aux clics et aux flashs, ce qui montre clairement le caractère multisensoriel de cette structure. En général, ces cellules sont excitées (ou inhibées) de la même façon dans les deux modalités. Dans une autre étude, Dubner et Rutledge (1964) trouvent 15 % de neurones bimodaux dans le gyrus supra-sylvien du chat

non anesthésié. Lorsque les stimuli sont présentés par paires audio-somesthésiques ou visuo-somesthésiques (les résultats pour des paires audiovisuelles ne sont pas présentés), avec des délais interstimulus courts (environ 50 ms), un effet de facilitation est observé et peut se manifester de trois façons :

- diminution du seuil d'excitabilité
- diminution de la latence des réponses
- augmentation du nombre de décharges

Lorsque le délai augmente, des effets d'inhibition, rappelant ceux observés sur la surface du cortex (Rutledge, 1963), sont observés avec la même asymétrie (la période réfractaire est plus longue lorsque le premier stimulus est visuel que lorsqu'il est auditif).

1.1.3 Aires de convergence dans le cortex frontal

Outre les 4 structures de convergence définies chez le chat, des exemples de convergence multisensorielle sont également rapportés dans le cortex frontal du singe écureuil anesthésié, par Bignall et Imbert (1969). Dans cette étude qui combine ECoG et EEG intracortical, plusieurs zones de convergence sont identifiées dans le cortex frontal : le cortex frontal post-arqué (d'après les auteurs, analogue de l'aire péricruciée chez le chat, voir la figure 1.1 page 6), le cortex orbito-frontal, l'opercule frontal et le cortex pré-arqué, ainsi que dans l'insula. Dans ces structures, contrairement aux résultats chez le chat, les latences des réponses associatives audiovisuelles sont du même ordre de grandeur que les latences observées dans les aires primaires. L'ablation des aires unisensorielles primaires ou la stimulation électrique des aires unisensorielles primaires suggèrent que le cortex frontal reçoit à la fois des entrées corticales et sous-corticales.

Un résultat analogue de convergence polysensorielle est trouvé dans une étude en sEEG (voir partie 6.4 page 92) chez l'homme (Walter, 1964) : des réponses auditives, somesthésiques et visuelles sont enregistrées dans le cortex préfrontral de patients épileptiques à des latences très précoces (environ 30 ms ; à titre de comparaison, chez l'homme, les premières réponses sensorielles corticales sont enregistrées vers 15 ms dans le cortex auditif primaire et vers 35 ms dans le cortex visuel). Comme dans les études chez l'animal, des paires de stimuli auditifs et visuels sont présentées avec un délai variant de 70 à 270 ms : aucun effet sur la période réfractaire n'est constaté et les réponses sont totalement additives. L'auteur conclut que les réponses sont dues à des projections totalement indépendantes des différents modalités sensorielles vers le cortex préfrontal.

1.1.4 Effet de l'anesthésie sur les interactions multisensorielles

Une partie de ces résultats a été obtenue chez l'animal anesthésié, or il était connu déjà à l'époque que l'anesthésie altère les réponses neuronales. Cependant, Thompson, Johnson et Hoopes (1963) trouvent des résultats identiques en diminuant la dose d'anesthésiant (chloralose) des chats. Et Thompson et Shaw (1965) confirme l'activation focale du gyrus suprasylvien chez le chat alerte par différentes modalités, bien que la réponse soit plus diffuse et moins ample que sous chloralose.

En revanche, Dubner et Rutledge (1964) montrent que les effets d'interaction pour des paires de stimuli intermodales sont plus importants à mesure que la dose de chloralose est

augmentée. Plus tard, Toldi, Fehér et Gerő (1980) compareront des réponses ECoG évoquées par des stimulations auditives, somesthésiques et visuelles dans des zones communes chez des chats sous nembutal et chloralose : alors que sous chloralose, les mêmes zones du gyrus suprasylvien que Thompson, Smith et Bliss (1963) sont activées par les trois stimuli, une toute autre configuration émerge sous nembutal.

Bien qu'elles ne remettent pas réellement en cause l'existence de ces zones polysensorielles, ces données invitent à la prudence quant aux résultats d'études chez l'animal anesthésié, certains effets, notamment d'activation par plusieurs modalités sensorielles, pouvant être exagérés sous l'effet de l'anesthésie.

1.2 Convergence audiovisuelle dans le cortex visuel

Alors que les études revues dans la partie précédente ont montré l'existence de zones de convergence multisensorielle corticale hors des cortex sensori-spécifiques, un nombre non négligeable d'études ont cherché à montrer des effets identiques dans le cortex visuel, en utilisant les mêmes méthodes.

Murata, Cramer et Rita (1965) explorent le cortex visuel primaire (cortex strié) du chat alerte avec des stimulations visuelles (lumière diffuse), auditives (claquement de main derrière l'animal) et somesthésiques (pincements/*prickles*[1]) et trouvent que 38 % des cellules répondent à des claquements avec une latence moyenne de 60 ms alors que 70% répondent à une lumière diffuse, à une latence moyenne de 35 ms. Les cellules bimodales ou trimodales (répondant aux trois modalités) montrent une certaine organisation puisqu'une cellule répondant à une stimulation auditive a plus de probabilité de répondre à une stimulation somesthésique. Étant donné la latence relativement plus longue des réponses intermodales, les auteurs concluent qu'elles sont de type associatif, en référence aux réponses enregistrées dans les cortex associatifs non spécifiques. Bental, Dafny et Feldman (1968) trouvent, chez le chat éveillé, 61% de cellules du cortex visuel primaire altèrant leur taux de décharges à la fois pour des stimuli auditifs et visuels, alors que 67 % seulement répondent à la stimulation visuelle. Ces cellules semblent montrer une tendance à altérer leur distribution de décharges dans le même sens (excitation ou inhibition) pour les stimuli des deux modalités, mais cette assertion n'est pas testée statistiquement. Ce résultat conduit les auteurs à conclure que « la théorie de spécificité des modalités ne peut être maintenue » (*"theory about modality specificity cannot be upheld"*).

Ces deux premières études ont exploré le cortex visuel en utilisant un seul type de stimulation de chaque modalité, ce qui pourrait expliquer pourquoi elles ne trouvent qu'environ 70% de cellules répondant aux stimulations visuelles dans le cortex visuel. Par ailleurs, elles ne permettent pas de conclure quant à la spécificité des réponses auditives dans le cortex visuel et restent compatibles avec l'idée que les entrées auditives dans le cortex visuel ne portent pas d'autre information que la présence d'un stimulus. Avec l'évolution des connaissances sur la spécificité et le champ récepteur (CR) des neurones visuels, d'autres équipes vont, en utilisant des stimuli plus variés et en caractérisant le CR de ces cellules,

[1]Les termes en italiques sont les termes anglais utilisés par les auteurs

non seulement trouver que la totalité des cellules du cortex visuel répondent à au moins un type de stimulation visuelle, mais également mettre en évidence une certaine correspondance entre la spécificité des cellules pour les stimulations visuelles et auditives. Ainsi, les cellules du cortex visuel primaire peuvent montrer une spécificité pour la fréquence des sons purs chez le chat anesthésié (Spinelli, Starr & Barrett, 1968). Ces cellules représenteraient 28 % des cellules visuelles et se distinguent des cellules purement visuelles par un CR plus ample.

Dans le cortex visuel extra-strié (hors cortex primaire) du chat paralysé mais non anesthésié, F. Morrell (1972) ne trouve en revanche aucune spécificité pour la fréquence mais une bonne correspondance des CR des neurones pour les stimuli auditifs et visuels, dont une majorité répondent à des stimuli en mouvement : pour 41 % des cellules, le taux de décharges est maximal lorsque le stimulus auditif se trouve dans la même position le long de l'axe horizontal que le stimulus visuel provoquant la réponse maximale. De plus, la sélectivité pour la direction du mouvement correspond dans les deux modalités. Enfin Fishman et Michael (1973) dénombrent, dans les cortex visuels strié et extra-strié, 32 % de neurones visuels sélectifs pour une fréquence auditive et 7% de neurones visuels répondant sélectivement à des chuintements plutôt qu'à des sons purs. Une correspondance des CR auditifs et visuels est trouvée le long de l'axe horizontal, mais pas vertical, ce qui confirme partiellement les résultats de F. Morrell (1972). En outre, les populations de cellules bimodales et de cellules uniquement visuelles sont organisées en colonnes corticales (Fishman & Michael, 1973).

En ECoG, Bonaventure et Karli (1968) ont enregistré une réponse auditive corticale au niveau du cortex visuel de la souris, dont la latence est plus précoce que la réponse auditiela plus précoce enregistrée à la surface du cortex auditif.

Notons qu'aucune de ces études sur le cortex visuel n'a utilisé de paires de stimuli audiovisuels, si bien qu'il n'y a, à ma connaissance, aucune donnée sur le traitement éventuel d'un évènement audiovisuel dans le cortex visuel chez l'animal.

Si les preuves d'une sensibilité du cortex visuel à des stimulations auditives ne manquent pas, on ne trouve pas de résultats analogues dans le cortex auditif : selon Stewart et Starr (1970), on ne trouve pas de cellules répondant à des stimulations visuelles dans le cortex auditif primaire de chats anesthésiés. Sur 68 cellules testées, aucune ne répond à des flashs, des points ou des barres se déplaçant dans tout le champ visuel. Toutefois, des résultats opposés ont récemment été rapportés chez le macaque alerte et actif (Brosch, Selezneva & Scheich, 2005).

1.3 Convergence corticale chez l'homme : premières études en potentiels évoqués (PE)

Mises à part de rares données en EEG intracérébrale (Walter, 1964), les données neurophysiologiques anciennes sur la convergence audiovisuelle chez l'homme proviennent essentiellement de l'EEG de scalp. Le but des études d'EEG ayant utilisé des stimuli bimodaux n'était pas tant de définir les structures de convergence multisensorielle que d'étudier la

spécificité des différentes ondes des PE par rapport aux différentes modalités sensorielles. La localisation des structures cérébrales à l'origine des potentiels enregistrés sur le scalp est en effet difficile en raison de la diffusion des potentiels électriques dans les tissus cérébraux et extra-cérébraux. Par contre, ces études ont fourni des informations précieuses sur la latence de la convergence des informations auditives et visuelles chez l'homme.

Dès les années 60, Ciganek (1966) étudie la réponse à un flash précédé d'un clic à un délai variant de 40 à 250 ms. L'analyse est analogue celle utilisée chez le chat en ECoG (voir figure 1.2 page 8) : la réponse corrigée au flash suivant un clic est comparée à la réponse au flash présenté seul. L'amplitude des 6 premières ondes (jusqu'à une latence d'environ 170 ms) ne varie pas, donc ces 6 ondes sont censées être spécifiques à la modalité visuelle. Néanmoins, l'onde VII (vers 180 ms) est significativement diminuée lorsque le délai est de 250 ms, ce qui indique qu'elle n'est pas spécifique d'une modalité et que les entrées auditives et visuelles convergent à ce stade (le montage bipolaire entre Oz et Pz utilisé dans cette étude rend difficile la comparaison de ces ondes avec ce qu'on connait aujourd'hui des potentiels évoqués visuels).

C'est la spécificité sensorielle de la réponse positive au vertex vers 200 ms, évoquée à la fois par un stimulus auditif et un stimulus visuel, qui a sans doute été la plus débattue, sans doute parce qu'elle apparait à une latence charnière entre les ondes plus précoces considérées comme spécifiques et les réponses suivantes, considérées comme non spécifiques, telle la P300. Bien qu'il ait été montré que la réponse auditive au vertex vers 200 ms possède des générateurs dans le cortex auditif (Vaughan & Ritter, 1970), au moins deux études ont cherché à étudier les interactions entre les réponses au vertex évoquées par plusieurs modalités : en testant toutes les paires de stimuli intra et intermodales auditives, visuelles et somesthésiques, H. Davis, Osterhammel, Wier et Gjerdingen (1972) montrent que l'inhibition de la réponse à la deuxième stimulation est moindre pour les paires intermodales que pour les paires intramodales (le délai entre les composantes auditive et visuelle étant de 500 ms). Cependant, la réponse au second stimulus de la paire n'était pas corrigée par la méthode algébrique, ce qui limite l'interprétation. Dans une étude avec des paires visuo-auditives et auditivo-visuelles, Peronnet et Gerin (1972) montrent, en utilisant la correction algébrique, que l'inhibition due à la période réfractaire est moindre en intermodal qu'en intramodal, pour un délai de 250 ms. Ces deux études vont donc dans le sens d'une spécificité relative des réponses auditives et visuelles, sans que néanmoins soit exclue l'existence d'une composante non spécifique à cette latence.

Ces études en EEG de scalp suggèrent donc que la convergence des informations auditives et visuelles n'a pas lieu avant environ 200 ms dans les aires corticales. D'autres études plus récentes, utilisant d'autres types de protocoles ainsi que des analyses plus sensibles, ont mis en défaut cette idée. Elles seront passées en revue dans le chapitre 4

1.4 Convergence sous-corticale

Alors que la notion de réponse associative non spécifique (commune à plusieurs modalités) s'est plutôt développée avec les études sur le cortex cérébral, celle d'interaction audiovisuelle lors du traitement d'une stimulus multisensoriel proprement dit va émerger

des études de la convergence dans des structures sous-corticales, en particulier au niveau du colliculus supérieur.

1.4.1 Colliculus Supérieur / Tectum optique

Le colliculus est une structure sous-corticale qui reçoit, dans ses couches les plus profondes, des entrées de divers noyaux et relais sensoriels ascendants appartenant aussi bien aux modalités visuelle, auditive et somesthésique (Edwards, Ginsburg, Henkel & Stein, 1979). Elle a rapidement été considérée comme une structure de convergence multimodale pour plusieurs raisons :

- sa lésion provoque des déficits dans des comportements d'orientation vers des stimuli aussi bien visuels qu'auditifs ou somesthésiques (par exemple G. E. Schneider, 1969)
- on trouve dans les couches profondes du colliculus supérieur des cellules répondant non seulement à des stimuli auditifs, visuels, mais également des cellules répondant à deux voire à trois modalités (Horn & Hill, 1966), les couches superficielles étant chez la plupart des espèces dédiées uniquement à la modalité visuelle. Ce résultat a été répliqué chez toutes les espèces mammifères étudiées, mais est également valable pour sa structure analogue chez des espèces aviaires et reptiliennes, le tectum optique (poule : Cotter, 1976, chouette : Knudsen, 1982, iguane : Stein & Gaither, 1983).
- ces cellules montrent une préférence pour les stimuli complexes en mouvement, aussi bien auditifs que visuels (Gordon, 1973 ; Wickelgren, 1971)
- la stimulation électrique de certaines cellules du colliculus supérieur du chat provoque des mouvements controlatéraux des organes récepteurs tels que la tête, les yeux et les pavillons des oreilles (Harris, 1980, cité par Harris, Blakemore & Donaghy, 1980).

Tous ces résultats suggèrent qu'il s'agit d'une structure impliquée dans des comportements d'orientation vers un stimulus, qu'il soit visuel ou auditif, et que cette capacité serait un caractère ancestral commun au moins aux vertébrés terrestres. Toutefois des différences importantes dans la répartition des cellules multisensorielles ont été trouvées chez différentes espèces. La proportion de cellules multisensorielles est de 1 à 2% chez le hamster (Chalupa & Rhoades, 1977), de 8% chez le macaque (Cynader & Berman, 1972) et de 50 à 60% chez le chat (par exemple Meredith & Stein, 1986b). Elle peut même atteindre 90% des cellules chez la chouette ou le cochon d'Inde et s'étendre aux couches superficielles (Knudsen, 1982 ; King & Palmer, 1985), dans lesquelles les cellules sont spécifiques à la modalité visuelle chez les autres espèces. Ces différences importantes pourraient être liées à des différences de niche écologique : par exemple, la chouette est un prédateur nocturne dont la perception repose majoritairement sur des indices auditifs spatiaux.

Les mécanismes neuronaux qui sous-tendent cette convergence multisensorielle ont été étudiés sous deux aspects : celui de la correspondance des représentations spatiales de différentes modalités et celui de l'interaction des réponses lors d'une stimulation multisensorielle.

Les expériences concernant les caractéristiques spatiales de la réponse des cellules des couches profondes du colliculus supérieur ont en général rapporté une correspondance spatiale des CR auditifs et visuels : une cellule auditive et une cellule visuelle proches l'une

de l'autre, ou une cellule audiovisuelle, répondent de façon maximale à des stimuli auditifs et visuels provenant d'une même position de l'espace. Cette correspondance a été observée chez un grand nombre d'espèces (hamster : Chalupa & Rhoades, 1977, souris : Dräger & Hubel, 1975 ; Gordon, 1973, cochon d'Inde : King & Palmer, 1983, chouette : Knudsen, 1982, chat : Wickelgren, 1971). De plus, il a en général été montré que le colliculus supérieur est organisé de façon spatiotopique, les cellules proches ayant des champs récepteurs auditifs et/ou visuels proches.

Cette relation entre représentations auditive et visuelle de l'espace dans le colliculus supérieur peut cependant être plus complexe chez certaines espèces : les études citées plus haut ont en effet étudié les champs récepteurs visuels alors que l'animal garde les yeux dans une position de repos, c'est-à-dire le regard orienté dans l'axe de la tête. Il n'est donc pas possible de dire si cette correspondance est conservée si les yeux changent d'orientation dans l'orbite. Harris et coll. (1980) montrent que, chez le chat, les champs récepteurs des cellules du colliculus sont invariantes dans le référentiel rétinien en ce qui concerne la modalité visuelle, et dans le référentiel de la tête en ce qui concerne la modalité auditive. Donc si l'animal oriente son regard sur le côté, la correspondance des champs récepteurs n'est pas maintenue. Mais ces auteurs montrent également que l'orientation de la tête suit naturellement de près l'orientation des yeux chez le chat, ce qui a pour effet de maintenir la correspondance des représentations spatiales.

À l'inverse, les primates sont capables d'orienter leur regard pendant un long moment sans bouger la tête. Jay et Sparks (1984) montrent que selon l'orientation du regard, le champ récepteur auditif des cellules du colliculus supérieur varie dans le référentiel de la tête afin de compenser l'orientation du regard. En moyenne cependant, cette variation est inférieure à l'angle des yeux dans les orbites, ce qui indique que plusieurs systèmes de coordonnées co-existent dans le colliculus supérieur du macaque (Jay & Sparks, 1987).

Que les champs récepteur auditifs et visuels soient alignés ou qu'il existe des mécanismes neuronaux ou comportementaux de compensation des différents systèmes de coordonnées n'indique toutefois pas comment vont interagir les réponses à des stimuli auditifs et visuels lorsqu'ils sont présentés ensemble. Cette question a été étudiée principalement chez le chat (par exemple Meredith & Stein, 1983) et le cochon d'Inde (par exemple King & Palmer, 1985) anesthésiés. Chez ces deux espèces, plusieurs types d'interaction sont rencontrés :

- une cellule bimodale (c'est-à-dire répondant aux deux stimuli présentés séparément), peut voir son taux de décharge ou la durée de sa réponse augmenter au-delà de la réponse unimodale la plus forte, et même au-delà de la somme des réponses aux stimuli présentés séparément, lorsque les deux stimuli (par exemple auditifs et visuels) sont présentés simultanément au même endroit.
- une cellule bimodale peut voir sa réponse diminuer en-deçà de sa réponse unimodale maximale dans les mêmes conditions. Cette forme d'interaction est plus rarement observée, en tous cas chez le chat anesthésié.
- une cellule unimodale peut voir sa réponse augmenter ou diminuer si on ajoute un stimulus de l'autre modalité, dans les mêmes conditions que précédemment.

Ces interactions multisensorielles ont parfois été appelées multiplicatives en raison du fait qu'elles sont souvent supérieures à la somme des réponses aux stimuli unimodaux.

Ces différents types d'interaction peuvent être rencontrés dans la même cellule, selon les caractéristiques des stimuli. Différentes règles d'intégration, proposées notamment par Stein et Meredith (1993), expliquent ces différents types d'interaction.

- selon la "règle d'efficacité inverse", moins les stimuli auditifs et visuels sont efficaces présentés isolément, plus l'augmentation relative de leur taux de décharges sera grande s'ils sont combinés (Meredith & Stein, 1983, 1986b), à tel point que deux stimuli, apparemment inefficaces présentés séparément, peuvent évoquer une réponse s'il sont présentés simultanément. Cette règle s'expliquerait, selon ces auteurs, par le fait que la contribution de plusieurs modalités est d'autant plus nécessaire à la détection d'un stimulus que les stimuli unimodaux sont difficiles à détecter séparément. Notons qu'elle pourrait aussi s'expliquer par le caractère non linéaire de la réponse neuronale en fonction de l'intensité des stimuli.
- selon la "règle de coïncidence spatiale", les interactions varient en fonction de la correspondance spatiale des sources des stimuli (King & Palmer, 1985 ; Meredith & Stein, 1986a). Ainsi l'augmentation de la réponse est moindre si les stimuli auditifs et visuels proviennent de sources différentes mais restent dans leurs CR respectifs. En revanche, l'augmentation se transforme en diminution si l'un des stimuli sort de son CR. Cette règle de congruence spatiale est censée garantir l'unicité spatiale des stimuli lorsqu'ils sont perçus simultanément par différentes modalités.
- selon la "règle de coïncidence temporelle", les interactions varient en fonction de la correspondance temporelle des stimuli. De manière générale, plus les stimuli sont séparés dans le temps, moins l'interaction est importante, qu'il s'agisse d'une augmentation ou d'une diminution (Meredith, Nemitz & Stein, 1987). Cependant, l'interaction optimale ne correspond pas forcément à la coïncidence temporelle des stimuli : selon King et Palmer (1985), elle correspondrait à la différence de latence d'arrivée des informations auditives et visuelles au colliculus supérieur. En revanche, selon Meredith et coll. (1987), le délai optimal correspondrait plutôt à la différence de latence des périodes de décharge maximale, qui varient d'un neurone à l'autre et peuvent être différentes selon les modalités. Quoiqu'il en soit, il existe une certaine tolérance à la disparité temporelle puisque des interactions importantes ont lieu lorsque le délai dépasse de plus de 200 ms le délai optimal. Cette tolérance permettrait à l'organisme de réagir à un stimulus audiovisuel quelle que soit sa distance par rapport au stimulus, malgré la différence de vitesse de conduction du son et de la lumière dans l'air.

Bien que l'existence de telles interactions aient été établies principalement chez le chat et le cochon d'inde anesthésiés, et que d'importantes différences interspécifiques existent dans la structure multisensorielle du colliculus supérieur, Cynader et Berman (1972) mentionnent des augmentations de la réponse à des stimulations visuelles par la présentation concomitante d'une stimulation auditive dans le colliculus supérieur du macaque. En outre, des résultats similaires à ceux du chat anesthésié ont été obtenus chez le chat non anesthésié par Wallace, Meredith et Stein (1998).

Ces différentes règles d'intégration suggèrent l'existence de mécanismes neuronaux spécifiques à la perception d'un stimulus multisensoriel ayant une unité spatiale et temporelle et constituent la première description des interactions ayant lieu lors de la perception d'un

évènement audiovisuel proprement dit (voir cependant Bignall & Imbert, 1969). Soulignons cependant qu'elles ont été décrites pour les cellules d'une structure bien particulière, le colliculus supérieur, qui semble sous-tendre directement des comportements moteurs d'orientation. Ainsi de telles interactions ont été mises en évidence dans des cellules du colliculus supérieur projetant directement vers les voies efferentes du tronc cérébral (Meredith & Stein, 1985 ; Meredith, Wallace & Stein, 1992) ou dont la décharge est synchronisée aux saccades oculaires (Peck, 1987 ; voir aussi la partie 4.1.3 page 66).

Ces comportements seraient relativement indépendants de ceux sous-tendus par le cortex. Ainsi la lésion du colliculus supérieur chez le hamster provoque un déficit sélectif des comportements d'orientation vers des stimuli auditifs ou visuels mais pas des capacités de discrimination visuelle, alors qu'une lésion du cortex visuel a l'effet inverse (G. E. Schneider, 1969). Il semble cependant que de telles règles puissent décrire des interactions multisensorielles ayant lieu dans certaines structures corticales (voir la partie 4.1.1 page 64).

1.4.2 Autres structures sous-corticales

La formation réticulée mésencéphalique est depuis longtemps considérée comme une zone de convergence polysensorielle (voir par exemple Amassian & Devito, 1954). On y trouve, chez le chat anesthésié, des cellules répondant à plusieurs modalités sensorielles et le comportement de ces cellules pour des stimulations successives dans différentes modalités a été décrit comme proche de celles des aires corticales associatives (C. Bell, Sierra, Buendia & Segundo, 1964). Il a été proposé que cette structure constitue un relai vers ces aires corticales, qui permet de court-circuiter les aires sensorielles spécifiques. Cependant, il semble qu'une lésion de la formation réticulée chez le chat ne modifie pas les interactions multisensorielles dans ces cortex (Bignall, 1967).

D'autres structures sous corticales présentent des cellules pouvant être activées, ou dont l'activité peut être modulée, par différentes modalités sensorielles : des stimulations auditives et visuelles peuvent ainsi modifier la réponse de cellules somesthésiques dans divers noyaux du thalamus (Hotta & Kameda, 1963) ou dans le bulbe rachidien (Jabbur, Atweh, To'mey & Banna, 1971) du chat anesthésié.

Plus récemment, des cellules répondant à différentes modalités sensorielles ont été identifiées dans la substance noire du singe alerte (Magariños-Ascone, Garcia-Austt & Buno, 1994) et du chat anesthésié (Nagy, Paroczy, Norita & Benedek, 2005), ainsi que dans le noyau caudé du chat (Nagy et coll., 2005). Ces structures seraient impliquées dans l'intégration sensorimotrice. Selon une étude de Nagy, Eordegh, Paroczy, Markus et Benedek (2006), les réponses de ces cellules à un stimulus audiovisuel montreraient les mêmes propriétés multiplicatives que celles observées dans le colliculus supérieur.

Enfin des effets d'interactions audiovisuelles ont récemment été mis en évidence dans des neurones du thalamus : le noyau supragenouillé du chat comprend une proportion faible mais significative de neurones audiovisuels, mais il serait un relai entre deux structures multimodales : le colliculus supérieur et le cortex ectosylvien antérieur (Benedek, Pereny, Kovacs, Fischer-Szatmari & Katoh, 1997). Des noyaux traditionnellement considérés comme modalité-spécifiques peuvent aussi être sensibles à des stimuli d'autres modalités

sensorielles : ainsi chez le rat alerte effectuant une tâche de discrimination auditive, les neurones auditifs du corps genouillé médian qui répondent à la cible peuvent voir leur taux de décharge augmenter de façon très précoce lorsque la cible est accompagnée d'un stimulus visuel accessoire spatialement congruent (Komura, Tamura, Uwano, Nishijo & Ono, 2005). Cette augmentation est associée à une diminution du temps de réaction pour les stimuli audiovisuels congruents par rapport aux stimuli visuels.

1.5 Études anatomiques de la convergence multisensorielle

S'il est un domaine où le modèle de convergence tardive est totalement assumé, c'est celui de l'anatomie cérébrale. Dans une étude relativement exhaustive des connections cortico-corticales du singe rhesus, E. G. Jones et Powell (1970) cherchent à définir les voies de convergence des voies sensorielles auditives, visuelles et somesthésiques par la méthode des lésions. De façon générale, leurs résultats montrent que chaque aire primaire projette vers des aires de même modalité sensorielle dans le cortex temporo-pariétal selon un chemin sériel mais réciproque, et envoie parallèment des projections vers des régions différentes du cortex moteur. La convergence intersensorielle a lieu un peu plus haut dans cette chaîne : les trois systèmes convergent alors vers des zones polysensorielles telles que le sillon temporal supérieur (STS, homologue selon eux des gyrus supramarginal et angulaire chez l'homme), le cortex orbitofrontal, le sillon arqué et l'opercule frontal. Ces aires de convergence projettent à leur tour vers les pôles frontal et temporal. Enfin, tout au long de cette voie ascendante, dans chacun des systèmes sensoriels, on trouve des projections vers le cortex cingulaire et parahippocampique. Ces résultats sont illustrés dans la figure 1.3 page suivante. La méthode est assez grossière par rapport aux études de traceurs qui suivront, mais le message a le mérite d'être simple et clair.

Ces résultats seront largement repris dans une revue de Pandya et Seltzer (1982) selon qui les cortex associatifs unisensoriels ne reçoivent des entrées que du système primaire correspondant, la convergence multisensorielle s'effectuant au niveau des cortex associatifs non spécifiques ou polysensoriels qui seraient au nombre de 5 chez le singe rhésus (voir aussi la figure 1.4 page 19) :

- les cortex polysensoriels (sillon intra-pariétal, IPS et STS) recevant des entrées d'au moins deux cortex associatifs modalité-spécifiques : visuel et somesthésique pour l'IPS, trimodal pour le STS.
- le cortex associatif frontal comprenant les cortex prémoteur et préfrontal
- le cortex associatif paralimbique (gyrus parahippocampique)

Dans une revue de la hiérarchie des aires sensorielles, maintes fois citée pour décrire l'organisation des systèmes sensoriels (Felleman & Van Essen, 1991), les différents systèmes sensoriels sont présentés comme relativement séparés. Les auteurs reconnaissent toutefois que des projections entre systèmes sensoriels existent mais sont peu étudiées. Selon Mesulam (1998), la convergence des voies sensorielles auditive et visuelle n'aurait lieu qu'à partir du 5^{e} relai synaptique cortical dans les zones de convergence hétéromodales définies plus haut.

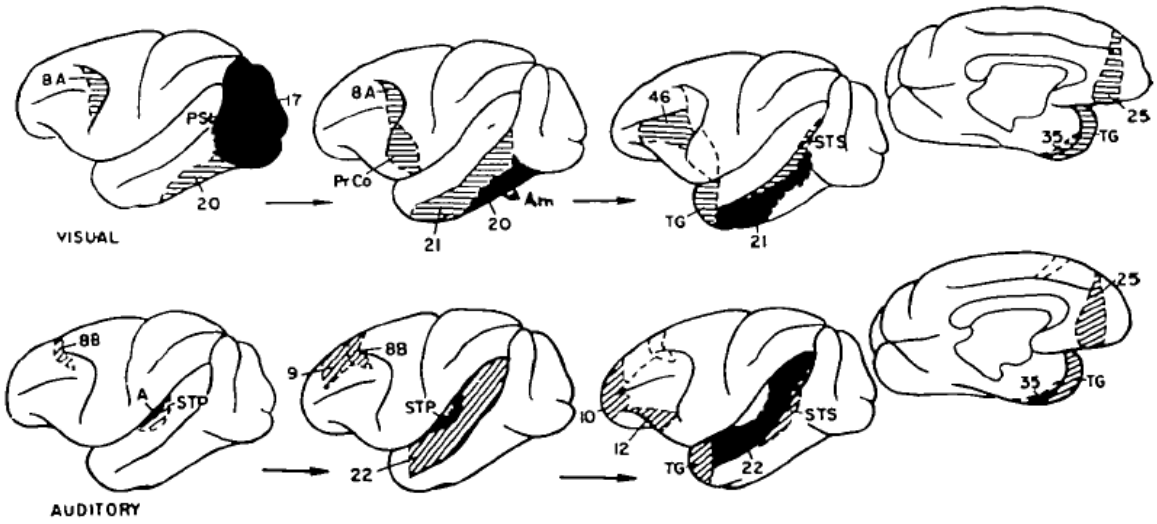

FIG. 1.3 – Schéma récapitulatif des projections cortico-corticales du singe rhésus : sur chaque carte, les zones en noir représentent les zones lésées et les zones hachurées celles où des fibres dégénérées sont trouvées, c'est-à-dire les aires de projection de la zone lésée. Chaque carte représente une étape dans la progression des informations sensorielles auditives et visuelles. D'après E. G. Jones et Powell (1970)

Avec l'utilisation de méthodes anatomiques plus sensibles telles que les traceurs, on a cependant découvert des connexions ne respectant pas cette hiérarchie, en particulier des connexions latérales entre aires sensorielles primaires ou secondaires de modalités différentes. Nous nous limiterons ici aux connexions concernant les aires auditives et visuelles.

En injectant un traceur antérograde dans les différentes aires auditives de la gerbille, Budinger, Heil et Scheich (2000) trouvent un certain nombre de projections vers d'autres aires sensorielles, dont des aires visuelles. Ce résultat sera confirmé chez le macaque où des projections du cortex auditif secondaire vers le cortex visuel strié et extrastrié sont mises en évidence (Rockland & Ojima, 2003). Une autre étude a également montré, par injection d'un traceur rétrograde dans le cortex visuel primaire (strié) du macaque, l'existence de projections des aires auditives primaires et secondaires, ces dernières étant plus nombreuses dans la partie périphérique du cortex visuel primaire que dans sa partie fovéale (Falchier, Clavagnier, Barone & Kennedy, 2002 ; Clavagnier, Falchier & Kennedy, 2004).

Concernant les projections vers les aires auditives pouvant porter des informations visuelles, les résultats actuels suggèrent qu'elles proviennent plutôt d'aires principalement visuelles, mais répondant aussi à des stimuli auditifs. Ainsi, il existe des projections réciproques entre l'aire auditives primaire et une aire visuelle secondaire (qui, par ailleurs, répond également à des stimuli auditifs : Barth, Goldberg, Brett & Di, 1995) chez le rat (Hishida, Hoshino, Kudoh, Norita & Shibuki, 2003). Chez le marmouset, une aire visuelle antérieure au STS (homologue de l'aire polysensorielle temporale supérieure chez le macaque) projette vers le cortex auditif (Cappe & Barone, 2005). Ces données sont compatibles avec le fait que le cortex auditif secondaire chez le macaque montre des potentiels de champ locaux évoqués par un stimulus visuel et que le profil de ces potentiels le long des couches du cortex correspond à des projections de type *feedback* (Schroeder & Foxe,

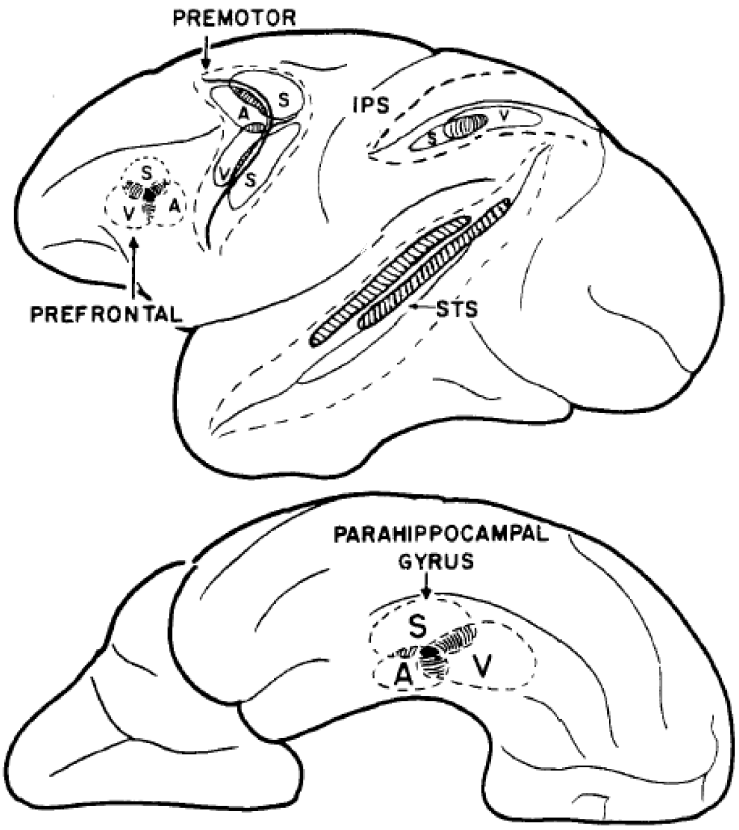

FIG. 1.4 – Aires de convergence définies par la méthode des lésions chez le singe rhésus. A : aires de projection auditive, V : Aires de projection visuelle, S : Aires de projection somesthésique. IPS : sillon intra-pariétal, STS : Sillon temporal supérieur. D'après Pandya et Seltzer (1982)

2002).

1.6 Conclusion

Bien que la question de l'activation multisensorielle ne se pose pas encore en termes d'interactions propres à une stimulation audiovisuelle, une certains nombre d'éléments plaident donc à la fin des années 70 pour une conception complexe de l'interaction des différents systèmes sensoriels. Au moins trois modes de convergence multisensorielle émergent des données présentées :

- convergence sous-corticale
- convergence dans les aires primaires
- convergence dans les aires associatives

Il suffit cependant d'ouvrir n'importe quel ouvrage généraliste sur le système nerveux central pour constater que c'est le modèle de convergence tardive dans les aire associatives qui s'est imposé. L'idée d'une convergence tardive s'entend ici à la fois dans le sens anatomique (les aires associatives correspondent aux aires se situant en bout de chaine des connections

cortico-corticales) et dans les sens fonctionnel et temporel (elles correspondent à des aires dans lesquelles sont enregistrées des réponses non spécifiques à des latences relativement longues par rapport aux latences des réponses sensorielles).

Pourtant dans la plupart des études citées, la convergence multisensorielle a été étudiée avec des stimulations désynchronisées, voire séparément dans les modalités auditive et visuelle. Paradoxalement, l'étude des interactions effectives des informations auditives et visuelles lors d'une véritable stimulation bimodale s'est faite plutôt à travers l'étude de la convergence sous-corticale au niveau du colliculus supérieur. C'est aussi par le biais de ces travaux que semble avoir perduré l'intérêt pour les interactions multisensorielles en neurosciences cognitives, comme nous le verrons dans le chapitre 4.

Chapitre 2

Phénomènes d'interactions audiovisuelles en psychologie expérimentale

Contrairement à la littérature neurophysiologique, la question des interactions entre différentes modalités sensorielles est récurrente en psychologie expérimentale depuis le début du vingtième siècle. Dans une revue sur le sujet, Ryan (1940) cite un nombre non négligeable d'études sur les relations des différents « départements sensoriels », publiées principalement dans les années 30. Toutefois, dans plusieurs revues ayant trait à la perception multimodale, les auteurs déplorent déjà le manque d'intérêt expérimental pour les relations entre les sens, malgré l'intérêt théorique qui leur est porté. Ainsi, Ryan (1940) rapporte que « si les auteurs de discussion générale sur la perception mentionnent occasionnellement le problème de la coopération [intersensorielle] dans la perception, ils donnent rarement les références de résultats expérimentaux ». De même Gilbert (1941) mentionne que, « bien que les preuves d'un interdépendance fonctionnelle des différentes modalités sensorielles soient disponibles depuis plus de 50 ans, elles ont peu attiré l'attention des psychologues jusqu'à très récemment ». Trente plus tard, un constat analogue est fait par Loveless, Brebner et Hamilton (1970) : « l'interaction des systèmes sensoriels en perception est un principe qui a fait l'objet de plus de discours que de recherches systématiques. »

Bien que le nombre des articles directement concernés par les interactions multisensorielles soit sans nul doute infime comparé à la masse des articles consacrés à une modalité particulière de perception, leur nombre absolu est cependant loin d'être négligeable. Une revue exhaustive de cette littérature irait au-delà des objectifs de cette introduction ; d'ailleurs une bonne partie des références n'existe qu'en allemand ou en russe. Des revues plus ou moins complètes existent (Gilbert, 1941 ; London, 1954 ; Loveless et coll., 1970 ; Ryan, 1940 ; Welch & Warren, 1986).

Dans cette partie, je décrirai les différents phénomènes qui suggèrent l'existence d'interactions entre les systèmes auditif et visuel. Nous essaierons de voir quelles ont été les différentes conceptions des relations entre systèmes sensoriels auditif et visuel, soit d'un point de vue fonctionnel soit d'un point de vue anatomique ou physiologique, même si d'une manière générale, cette littérature fait peu référence à des modèles biologiques. Par

ailleurs, nous verrons comment est peu à peu devenue pertinente la notion d'évènement multisensoriel (délimité dans le temps et dans l'espace, dont les dimensions sensorielles sont liées par des relations apprises ou causales). Nous verrons que l'idée d'interactions spécifiques à des traitements auditif et visuel se rapportant à des propriétés communes d'un évènement audiovisuel unique n'a émergé que progressivement.

2.1 Effets intersensoriels sur les capacités perceptives

2.1.1 Effets dynamogéniques

L'un des premiers effets intersensoriels mis en évidence est l'effet dynamogénique, terme emprunté par Ryan (1940) à Johnson (1920) pour qualifier l'effet d'un stimulus accessoire dans une modalité sensorielle sur l'acuité ou le seuil de perception dans une autre modalité.

Dans les années 30, sont mises en évidence aussi bien des modifications du seuil de perception d'un motif visuel (acuité visuelle) par un stimulus auditif accessoire supraliminal (Hartmann, 1933 ; Kravkov, 1934, 1936), que celles du seuil de perception d'un son pur (par exemple Child & Wendt, 1938) ou du seuil de discrimination de différentes intensités et hauteurs de sons purs (Hartmann, 1934) par un stimulus visuel accessoire supraliminal. Dans les années 50 et 60, plusieurs expériences montrent également des effets intersensoriels sur le seuil de perception, soit d'un stimulus auditif supraliminal sur le seuil de perception visuelle (Maruyama, 1959 ; Symons, 1963 ; W. H. Watkins & Feehrer, 1965), soit l'inverse (Gregg & Brogden, 1952 ; O'Hare, 1956 ; Sheridan, Cimbalo, Sills & Alluisi, 1966).

Il faut cependant souligner qu'en général les effets dynanomgéniques sont de faible amplitude (ils correspondent par exemple à une diminution du seuil de 2 dB dans l'étude de Child & Wendt, 1938), qu'ils peuvent correspondre aussi bien à des diminutions du seuil (c'est le cas le plus courant) qu'à des augmentations (voir par exemple E. T. Davis, 1966) et que plusieurs résultats négatifs ont également été rapportés (Serrat & Karwoski, 1936 ; Gulick & Smith, 1959 ; Karlovich, 1969 ; Moore & Karlovich, 1970).

2.1.2 Modèles explicatifs de l'effet dynamogénique

Selon Gilbert (1941), plusieurs facteurs expliquent ces effets contradictoires : il s'agit d'une part de la correspondance des qualités du stimulus accessoire et de la cible et d'autre part leur intensité relative. Le premier facteur est lié à l'idée que certaines qualités, pourtant propres à une modalité sensorielle (telles que la couleur ou la hauteur tonale) sont fondamentalement associées et transcendent les différentes modalités sensorielles (voir la partie 2.2 page 25). Le stimulus accessoire faciliterait d'autant plus la détection du stimulus cible, que leurs qualités correspondent. Le second facteur serait lié à la ségrégation figure/fond : un stimulus accessoire de faible intensité fait partie du fond et faciliterait donc la perception du stimulus cible. À l'inverse, lorsque le stimulus accessoire devient trop intense, il devient la figure et inhibe la détection du stimulus cible.

Ces conceptions d'inspiration gestaltiste co-existent avec des modèles plus biologiques des relations entre systèmes sensoriels. Ainsi, plusieurs auteurs tentent d'exclure une explication périphérique du phénomène (par exemple une propagation incidente d'influx nerveux

entre les voies nerveuses auditives et visuelles London, 1954 ou une action d'un stimulus sur les organes récepteurs de l'autre modalité telle que la pupille ou les muscles de l'oreille interne Child & Wendt, 1938). Ces auteurs privilégient l'hypothèse selon laquelle les interactions audiovisuelles à l'origine du phénomène ont lieu dans le système nerveux central, mais sous une forme qu'on appellerait aujourd'hui non spécifique, puisqu'il s'agirait d'une "irradiation" de l'activité nerveuse, une propagation diffuse entre les systèmes sensoriels (Hartmann, 1933 ; Kravkov, 1934) ou au niveau des centres moteurs (Child & Wendt, 1938).

Dans les années 50-60, l'hypothèse d'irradiation est progressivement remplacée par une autre explication non spécifique des effets dynamogéniques : l'implication de la formation réticulée. En effet, cette structure du tronc cérébral reçoit de multiples entrées sensorielles (voir partie 1.4.2 page 16) et elle est impliquée dans la régulation de l'attention et de l'éveil (*arousal*). La présence d'un stimulus accessoire permettrait donc d'améliorer (ou de dégrader) l'état d'éveil du sujet et faciliterait la détection du stimulus dans l'autre modalité. Si ces interprétations non spécifiques permettent de rendre compte des effets de l'intensité relative des stimuli auditifs et visuels sur l'effet dynamogénique, elles excluent d'emblée l'idée que le stimulus accessoire soit porteur d'informations spatiales ou temporelles qui renseignent sur la présence ou l'absence du stimulus à détecter. Il n'est donc pas étonnant que la plupart des études aient utilisé indifféremment des stimuli accessoires continus ou temporellement définis et que leurs auteurs ne se soient guère soucié de la correspondance spatiale des sources des stimuli auditifs et visuels. Dans le cas d'une stimulation accessoire continue, il n'est pas exclu que les effets dynamogéniques observés soient en grande partie dus à des modifications de l'état d'éveil, la plupart des études utilisant un paradigme par blocs où les conditions unimodales et bimodales duraient suffisamment longtemps pour permettre à de tels effets chroniques de se mettre en place.

Toutefois, certains résultats montrent déjà l'importance de la correspondance temporelle entre le stimulus accessoire et la cible à détecter. Child et Wendt (1938) ont ainsi montré que la diminution du seuil de perception auditive est maximale lorsque le stimulus visuel accessoire précède le son de 500 ms. Mais peut-être du fait qu'un délai semble nécessaire à l'établissement de l'effet, ce résultat reste compatible avec les conceptions non spécifiques d'irradiation ou d'activation réticulaire. D'autres résultats suggèrent cependant que cette explication est insuffisante : Howarth et Treisman (1958) montrent que, si l'on mélange différents délais entre les stimuli auditifs et visuels, l'effet facilitateur disparait et aussi que si le stimulus accessoire est présenté après le stimulus cible, on observe toujours une facilitation. Pour eux, l'effet facilitateur s'explique donc par une réduction de l'incertitude temporelle sur le moment d'apparition de la cible grâce au stimulus accessoire. De leur côté, Loveless et coll. (1970) soulignent que certaines expériences suggèrent des interactions des informations spatiales, non explicables par des facteurs tels que l'éveil. Ainsi, Maruyama (1961) montre qu'une stimulation auditive unilatérale augmente la sensibilité visuelle dans l'hémichamp controlatéral.

2.1.3 Effet dynamogénique et théorie de la détection du signal

Une autre difficulté dans l'interprétation des effets dynamogéniques vient du fait que les études précédemment citées peuvent presque toutes être soupçonnées d'avoir confondu une modification de la sensibilité de la perception avec celle du biais de réponse (Loveless et coll., 1970), tels qu'ils sont définis par la théorie de la détection du signal (TDS : D. M. Green & Swets, 1966). Ainsi, l'augmentation de la performance des sujets pourrait être due, non pas au fait que le seuil de perception diminue (augmentation de la sensibilité), mais au fait que les sujets montrent une plus grande propension à répondre (augmentation du biais) lorsque le stimulus accessoire est présenté. De rares études, telles que celles de Child et Wendt (1938) et Howarth et Treisman (1958), avaient cependant utilisé des essais pièges (*catch trials*) leur permettant de contrôler les fausses alarmes et montré que la variation de la propension des sujets à détecter un signal (qu'il soit réel ou non) ne pouvait rendre compte de l'augmentation du nombre de vraies détections en condition bimodale.

L'application de la TDS n'a toutefois pas permis de trancher entre biais et sensibilité : Loveless et coll. (1970, expérience 4) montre en effet que la présence d'un stimulus auditif synchrone supraliminal dans une tâche de détection visuelle augmente à la fois la sensibilité et le biais par rapport à une situation unimodale. En ce qui concerne l'effet d'un stimulus visuel synchrone sur le seuil de perception auditif, Bothe et Marks (1970) ne trouvent un effet facilitateur que chez 1 sujet sur 4, tandis qu'un autre sujet montre une diminution de la sensibilité.

Des études récentes ont cependant réussi à mettre en évidence un effet intersensoriel sur la sensibilité dans les deux cas visuo-auditif (Lovelace, Stein & Wallace, 2003) et auditivo-visuel (Bolognini, Frassinetti, Serino & Ladavas, 2005 ; Frassinetti, Bolognini & Ladavas, 2002). Dans ces deux dernières expériences, l'effet disparaissait lorsque l'origine spatiale des stimulations unimodales était différente. Comme le font remarquer Lovelace et coll. (2003), l'absence d'effets intersensoriels dans les premières études pourrait être dû au manque de correspondance spatiale des stimuli auditifs et visuels

2.1.4 Modèles de détection d'un stimulus bimodal au seuil

La TDS a également été utilisée pour modéliser la diminution intersensorielle du seuil. Toutefois, elle n'est pas adaptée pour modéliser l'action d'une stimulus supraliminal sur la détection au seuil, car c'est un modèle dans lequel la détection supraliminale n'est pas formalisée. La modélisation a donc concerné le cas particulier où l'on mesure le seuil de détection d'un stimulus liminal présenté dans deux modalités à la fois, la question sous-jacente étant de savoir si l'on peut améliorer le seuil de détection d'un signal en fournissant la même information dans différentes modalités (voir par exemple Osborn, Sheldon & Baker, 1963).

Fidell (1970) définit deux types de modèles selon que les interactions entre les systèmes ont lieu plutôt au niveau "sensoriel" ou "décisionnel" dans le modèle de décision perceptuelle postulé par la TDS (voir aussi Mulligan & Shaw, 1980).

- dans les modèles d'interaction décisionnelle, chaque système sensoriel déciderait de la probabilité de la présence ou de l'absence d'un signal en fonction de sa sensibilité

et de son biais propres. La présence d'un signal bimodal est détectée si l'un ou l'autre des deux systèmes l'a détecté ("ou" inclusif). La décision bimodale est donc basée sur le résultat des décisions unimodales sans qu'il soit besoin de postuler une influence entre systèmes sensoriels au niveau de la détection de chaque stimulus.

– dans les modèles d'intégration sensorielle les probabilités de détection des deux systèmes de détection auditif et visuel s'additionnent, ce qui implique un échange d'informations entre les systèmes au niveau physiologique (d'où le nom de sommation physiologique donnée par Loveless et coll., 1970), et le biais est commun aux deux modalités. Ces modèles permettent de rendre compte de diminutions de la sensibilité supérieures à celles prédites par les modèles de convergence décisionnelle.

Chacun de ces deux types de modèles peut être, à son tour, décliné en plusieurs versions selon la corrélation pouvant exister entre la probabilité de détecter un stimulus dans l'une et l'autre des modalités (voir Mulligan & Shaw, 1980, pour les modèles décisionnels et Fidell, 1970, pour les modèles d'intégration sensorielle).

Chacun de ces modèles a été soutenu par des résultats expérimentaux : les données d'une expérience de détection bimodale menée par (Brown & Hopkins, 1967) favorisent un modèle décisionnel (voir cependant Morton, 1967, pour une critique) tandis que les données de Fidell (1970) sont plutôt compatibles avec un modèle d'intégration à corrélation nulle, voire négative (qui pourrait correspondre à une compétition pour les ressources attentionnelles : voir J. O. Miller, 1982, et la partie 7.1.1 page 101). Toutefois en comparant directement les prédictions des modèles d'intégration et de sommation statistique décisionnelle, plusieurs études trouvent des données mieux expliquées par un modèle décisionnel (Loveless et coll., 1970, expérience 1 ; Mulligan & Shaw, 1980).

Les modèles inspirés de la TDS favorisent donc plutôt un modèle de convergence décisionnelle (qui fut peut-être rapidement assimilé à un modèle de convergence tardive au niveau biologique et a pu contribuer au renforcement de cette hypothèse) et semblent exclure la sommation physiologique. Notons cependant qu'assimiler la distinction sensibilité/biais à une distinction en termes de niveau de traitement sensoriel et décisionnel suppose d'accepter la TDS comme modèle sériel du fonctionnement cognitif dans une tâche de détection. Remarquons également que dans toutes les expériences de détection bimodale (excepté Mulligan & Shaw, 1980), la source du signal dans les modalités auditive et visuelle était différente, l'expérience type consistant à dériver un même signal vers un oscilloscope pour la modalité visuelle et un casque pour la modalité auditive. Il est donc possible qu'elles aient sous-estimé l'amélioration bimodale du seuil, si celle-ci ne dépend pas uniquement de la congruence temporelle mais également de la congruence spatiale du stimulus audiovisuel.

2.2 Correspondance des dimensions synesthésiques

Nous avons vu que l'un des déterminants de l'effet dynamogénique était la correspondance supposée de certaines qualités ou dimensions entre différentes modalités sensorielles. Cette correspondance est assez intuitive concernant les dimensions telles que l'étendue spatiale et temporelle car elles peuvent être connues à la fois par les biais des informations visuelles et auditives. En effet, dans ce cas, les informations auditives et visuelles spatiales

ou temporelles se réfèrent à un même événement du monde extérieur et on peut donc imaginer aisément que la connaissance des unes peut faciliter le traitement des autres.

Cette correspondance est cependant loin d'être évidente concernant les dimensions d'un objet ou d'un évènement qui ne sont accessibles que par le biais d'une modalité sensorielle, comme la couleur pour la vision ou la hauteur tonale pour l'audition, et qui ne renvoient a priori pas à la même réalité. Dans les années 30, plusieurs théories proposent pourtant que des correspondances intersensorielles puissent exister entre ce second type de dimensions. Ainsi, selon les théories de la consonance (par exemple Werner, 1934), un mode de perception indifférencié existerait dans lequel le stimulus est ressenti comme un tout, indépendamment de la modalité sensorielle dans laquelle il est perçu.

Ces correspondances ont souvent été discutées dans le contexte de la synesthésie, un état assez rare dans lequel certaines personnes font l'expérience d'une sensation dans une modalité sensorielle alors qu'elles sont stimulées dans une autre modalité, l'exemple le plus connu étant celui de personnes qui voient des couleurs en entendant un mot ou un phonème particulier (revues dans Marks, 1975 ; Grossenbacher & Lovelace, 2001 ; Rich & Mattingley, 2002 ; Mulvenna & Walsh, 2006). Selon Marks (1975), ce phénomène aurait son pendant dans la population des non-synesthètes et des sujets normaux associeraient de manière consistante certaines dimensions auditives et visuelles, appelées alors dimensions synesthésiques.

2.2.1 Établissement des dimensions synesthésiques

La littérature psychologique des années 30 est riche d'études qui vont chercher à démontrer la correspondance entre différentes dimensions sensorielles. Ces études visent, d'une part, à découvrir quelles sont ces correspondances, c'est-à-dire identifer les qualités d'une modalité qui correspondent avec celles d'autres modalités sensorielles et, d'autre part, à étudier l'effet des qualités d'un stimulus sur la perception des qualités d'un stimulus d'une autre modalité, avec l'idée que différentes qualités secondes ne s'influencent pas au hasard mais reflèterait la structure d'un espace sensoriel commun à toutes les modalités.

Certains auteurs ont ainsi tenté de montrer une correspondance entre couleur et hauteur tonale : la hauteur tonale influencerait la perception des couleurs, le rouge tendant vers le violet ou le jaune selon qu'il est accompagné d'un son grave ou aigu (Zietz, 1931 cité par Gilbert, 1941), un son aigu augmenterait la vivacité du vert/bleu et diminuerait celle de l'orange/rouge (Kravkov, 1936).

Une autre dimension censée être commune aux différents sens est la brillance (*brightness*) : von Schiller (1935) montre par exemple que la brillance d'un stimulus visuel influence le perception de celle d'un stimulus auditif et réciproquement. Hornbostel (1931, cité par Ryan, 1940) prétend dériver ainsi une correspondance consistante entre la brillance de stimuli auditifs, visuels et olfactifs (!) sur la base de jugements de ressemblance intersensorielle d'un grand nombre de sujets. Il semble que la brillance d'un stimulus visuel dépende en grande partie de sa couleur et de sa luminosité, et que la brillance d'un son dépende principalement de sa hauteur. Notons que Cohen (1934) ne parvient pas à reproduire cette correspondance (ni même une quelconque correspondance consistante entre les sujets). De

même Pratt (1936) rapporte qu'il n'y a pas de modulation de la perception de la brillance d'un stimulus visuel par une stimulation auditive simultanée, qu'elle soit aigüe ou grave.

Des analogues de la rugosité ont été trouvés dans les domaines auditif (dissonance tonale) et visuel (scintillement) et ont été objectivés par von Schiller (1935) : des accords dissonants ou consonants influencent la fréquence critique à laquelle un stimulus visuel oscillant en intensité (*flicker*) est perçu comme continu. Selon Moul (1930), il existerait aussi une dimension commune et directement comparable d'épaisseur entre des sons purs et des couleurs, correspondant à leur intensité pour les premiers et à leur couleur et leur luminosité pour les seconds.

L'étude de ces correspondances a connu un certain renouveau à partir des années 60-70. Marks (1974) montre par exemple que des sujets normaux associent spontanément des sons aigus à des stimuli visuels brillants et des sons graves à des stimuli visuels ternes, alors qu'ils sont en désaccord sur l'appariement entre sonie (*loudness*) et brillance (*brightness*). Il existerait également une correspondance entre hauteur tonale et clarté (*lightness*), les sons les plus aigus ressemblant plus aux stimuli les plus clairs (Hubbard, 1996). Une correspondance également très étudiée est celle existant entre la hauteur tonale et la hauteur d'un stimulus visuel sur un axe vertical : un son plus aigu est spontanément associé à une position verticale plus haute qu'un son grave (Mudd, 1963). Roffler et Butler (1967) montrent également que des sujets localisent spontanément des sons aigus plus haut dans l'espace que des sons graves, même si leurs sources sont identiques.

2.2.2 Réalité des correspondances synesthésiques

Plusieurs études ont tenté d'objectiver ces correspondances en étudiant leur effet sur le temps de discrimination de l'une des dimensions, dans un paradigme de Garner (Garner, 1976) : dans ce paradigme expérimental, le sujet doit réaliser une tâche de discrimination entre deux stimuli audiovisuels variant sur une des deux dimensions (dimension pertinente), par exemple entre un son aigu et un son grave. Cette tâche est réalisée dans quatre conditions qui dépendent de la variation du stimulus dans l'autre dimension (dimension non pertinente) :

- dans la condition de base, le trait visuel ne varie pas.
- dans la condition d'interférence, le trait visuel varie indépendamment du trait auditif.
- dans la condition de corrélation positive (ou condition congruente), le trait visuel varie de façon consistance avec le trait auditif dans le sens prédit par la correspondance synesthésique (un son aigu est par exemple toujours associé à un stimulus visuel brillant).
- dans la condition de corrélation négative (incongruente) le trait visuel varie en sens inverse

Ce type de paradigme expérimental a pour but de mettre en évidence des effets d'interférence et des effets de congruence entre les deux dimensions manipulées : les premiers désignent le fait que les temps de réactions (TR) sont plus longs en condition d'interférence que dans la condition de base. Ils montrent que le traitement de la dimension non pertinente est automatique (ou que l'attention se partage nécessairement entre les deux

modalités). Les effets de congruence correspondent au fait que les TR sont plus courts en condition congruente qu'en condition de base, ce qui suggère que les traitements des deux dimensions interagissent.

Des effets d'interférence et de congruence ont effectivement été trouvés notamment pour les correspondances entre hauteur tonale du stimulus auditif et hauteur du stimulus visuel sur l'axe vertical (Melara & O'Brien, 1987), brillance et hauteur tonale (Marks, 1987 ; Melara, 1989), hauteur tonale et forme (anguleuse ou arrondie : Marks, 1987, expérience 4), brillance et sonie (Marks, 1987, expérience 3). Une asymétrie entre dimensions auditives et visuelles a souvent été rapportée, la dimension auditive non pertinente n'exerçant souvent qu'un effet faible, voire inexistant, sur la classification visuelle et ce même si la discriminabilité des traits auditifs et visuels est égalisée (par exemple Ben-Artzi & Marks, 1995).

L'effet d'interférence en lui-même ne permet pas de conclure à l'existence d'une dimension synesthésique qui transcenderait les modalités sensorielles puisqu'il peut s'expliquer par un partage d'attention obligatoire entre les modalités sensorielles, sans que les informations portées par les stimuli auditifs et visuels n'interagissent. L'effet de congruence en revanche pourrait refléter l'existence d'une telle dimension.

Toutefois, si l'effet de congruence existe effectivement entre condition de base et condition congruente (donc dans des blocs différents), on ne le retrouve pas si l'on compare les TR aux paires audiovisuelles congruentes et incongruentes au sein d'un même bloc (dans la condition de base ; par exemple : Melara & O'Brien, 1987 ; Patching & Quinlan, 2002 ; voir aussi Marks, 1987). Donc l'effet de congruence n'est observé que s'il est susceptible d'aider le sujet à répondre plus rapidement. Ces résultats suggèrent que la correspondance des dimensions n'est pas due à une correspondance sensorielle absolue de certains traits auditifs et visuels mais plutôt à une interaction au niveau de la sélection de la réponse, les sujets exploitant au maximum la différence sur la dimension non pertinente, en fonction du contexte. Dans le même ordre d'idée, Marks (1989) montre que les appariements subjectifs réalisés entre une hauteur tonale donnée et une luminosité donnée changent pour un même sujet en fonction de la gamme de hauteurs et de luminosité qu'il a à apparier dans un bloc expérimental (voir aussi Hubbard, 1996).

Que ces effets d'interférence et de congruence ne soient pas dus à une véritable correspondance sensorielle est corroboré par le fait que les effets d'interférence et de congruence peuvent être obtenus si l'une des dimensions sensorielles est remplacée par un stimulus verbal : le TR dans une tâche de classification des mots "haut" et "bas" est influencé par la hauteur tonale d'un son ou la hauteur d'un stimulus visuel (Melara & O'Brien, 1990 ; P. Walker & Smith, 1986) et inversement, la classification d'un son ou d'un stimulus visuel le long de ces dimensions interagit avec un stimulus verbal non pertinent pour la tâche (Melara & Marks, 1990 ; Melara & O'Brien, 1990). Ces résultats suggèrent que les interactions entre dimensions synesthésiques pourraient en partie avoir lieu à un niveau sémantique.

Cependant une partie des correspondances synesthésiques concerne des dimensions qui ne partagent a priori pas d'étiquettes verbales (par exemple la hauteur tonale et la brillance), ce qui oblige à postuler l'existence d'un lien sémantique d'un autre ordre que

simplement lexical. Le niveau sémantique des interactions n'implique pourtant pas qu'elles ne peuvent avoir lieu de manière automatique : Melara et O'Brien (1990) montrent en effet que l'effet de congruence ne dépend ni du délai séparant le stimulus auditif du stimulus visuel, ni de la probabilité que les deux traits soient congruents.

Les résultats les plus récents sur la correspondance des qualités secondes entre modalités auditive et visuelle suggèrent donc qu'elles sont largement induites par la tâche et dépendent plus de la réponse demandée que des liens physiques entretenus par les stimuli auditifs et visuels. Cependant, la direction des correspondances trouvées montre une certaine consistance, qui pourrait s'expliquer par des liens sémantiques entre dimensions auditives et visuelles, ces liens sémantiques pouvant s'exprimer de façon automatique dans un paradigme de Garner.

2.2.3 Correspondance des intensités

Il est cependant des dimensions dont il est plus difficile de dire a priori si elles renvoient à la même réalité alors qu'elles sont perçues dans les modalités auditive et visuelle. Ainsi les intensités auditive ou visuelle d'un stimulus peuvent ou non renvoyer à une caractéristique commune de l'évènement audiovisuel. Dans le cas, par exemple, d'un objet bruyant s'approchant, l'augmentation du volume sonore correspond à une augmentation de la taille du stimulus et donc à une plus grande énergie des stimuli auditif et visuel. Mais dans le cas d'un stimulus plus complexe, tel qu'une action produisant un bruit, il n'existe pas de lien direct entre l'énergie visuelle et l'intensité auditive. Il a pourtant paru naturel à de nombreux expérimentateurs d'étudier les effets d'une correspondance entre intensité auditive et visuelle (qu'il s'agisse de son étendue spatiale, de sa luminosité, de sa saturation en couleur).

Selon Ryan (1940), la correspondance entre intensité auditive et visuelle n'est en fait pratiquement pas étudiée, tellement elle est évidente. Par la suite, Dorfman et Miller (1966 cités par L. K. Morrell, 1968b) montrent qu'un stimulus visuel accessoire modifie le jugement d'intensité d'un son et Karlovich (1968) montre que lors de l'appariement d'intensité d'un son seul avec un son accompagné d'un flash, l'égalité est perçue pour des sons seuls plus intenses que les sons accompagnés, ce qui suggère que cet effet n'est pas dû à un biais de réponse. Cet effet a été répliqué par Odgaard, Arieh et Marks (2004), qui montrent également que l'effet persiste lorsqu'on varie la proportion relative des stimuli unimodaux et bimodaux. Il semble donc qu'il existe une véritable influence automatique de l'intensité visuelle sur le traitement de l'intensité sonore. D'autres études ont étudié l'effet inverse d'un stimulus auditif sur l'intensité perçue d'un flash : Stein, London, Wilkinson et Price (1996) montrent que des sujets jugent plus intense un stimulus visuel accompagné d'un bruit qu'un stimulus visuel présenté seul. Que ces effets reflètent des interactions sensorielles automatiques est cependant remis en cause par Odgaard, Arieh et Marks (2003) qui montrent que l'effet disparait lorsque l'on diminue la proportion des essais bimodaux, ou lorsqu'on utilise une variable dépendante moins sensible au biais de réponse (comparaison appariée d'intensité entre un stimulus unimodal et un stimulus bimodal).

Comment expliquer ces effets sensoriels (qui existent au moins dans le sens visuo-auditif) en tenant compte du fait que les intensités auditives et visuelles ne renvoient pas en général au même aspect d'un évènement audiovisuel ? Stein et coll. (1996) se réfèrent en fait explicitement à un modèle de sommation énergétique (qui n'est pas sans rappeler l'hypothèse d'irradiation (voir la partie 2.1.2 page 22) : la luminance d'un flash et l'amplitude du son correspondent tous deux à une certaine quantité d'énergie qui est censée déterminer la force de l'activité neuronale résultante : plus le nombre de photons atteignant la rétine, ou plus l'amplitude des ondes acoustiques est grande, plus les neurorécepteurs déchargent. La perception de l'intensité est censée découler directement de cette quantité d'activation et la modulation de la perception de l'intensité reflèterait la sommation énergétique des systèmes auditif et visuel et donc leurs interactions sensorielles précoces.

Cependant, l'asymétrie trouvée entre les systèmes auditif et visuel ne peut s'expliquer par une simple sommation d'énergie, sauf à rendre compte d'une moindre perméabilité du système visuel à l'énergie auditive (voir cependant la partie 1.2 page 11). Une piste alternative pourrait venir d'une étude de Rosenblum et Fowler (1991) qui montre que des jugements d'intensité de syllabes et de claquements de mains sont influencés par la présentation vidéo concomitante de l'effort apparent de l'auteur des sons (et non par des caractéristiques physiques, au sens quantité d'énergie, du stimulus visuel). Les auteurs excluent un simple biais de réponse car l'effet n'existe que lorsque les sujets sont incapables de détecter un conflit entre l'intensité auditive et l'effort visuel. Une telle interaction sensorielle s'explique selon les auteurs par le fait que les systèmes sensoriels ont internalisé les régles d'occurrence conjointe des évènements auditifs et visuels dans l'environnement (théorie directe-réaliste : Fowler & Rosenblum, 1991). Ce modèle pourrait également expliquer l'asymétrie si on admet qu'un stimulus visuel est plus souvent perçu comme la cause d'un stimulus auditif que l'inverse.

2.2.4 Résumé

Il a semblé à une époque que certaines formes d'interaction entre traitement auditif et traitement visuel pouvaient s'expliquer par un lien synesthésique existant entre certaines dimensions auditives et visuelles ne renvoyant pas à une même réalité. Dans le cadre des théories de la consonance ou des dimensions synesthésiques, on comprend que l'information à propos d'une dimension sensorielle puisse faciliter le traitement de l'information correspondante dans une autre modalité, par analogie à des dimensions telles que l'étendue spatiale ou temporelle, qui renvoient de façon claire à un objet unique. Cependant, on peine à comprendre le rapport de ces dimensions synesthésiques avec la réalité d'un évènement audiovisuel. Le manque de réalisme de ces études était déjà relevé par Ryan (1940), qui soulignait la nécessité d'utiliser des situations plus écologiques et des stimuli plus complexes pour mettre véritablement en évidence une coopération entre les sens. Bien que l'existence de telles correspondances puisse être mise en évidence dans des paradigmes expérimentaux objectifs, une partie des résultats pourrait bien s'expliquer par des liens d'ordre sémantique mais automatique, et non par un échange d'information entre des traitements sensoriels auditifs et visuels. On retrouve cette idée de correspondance dans des résultats plus récents concernant la perception de l'intensité, mais de façon non ambigüe uniquement pour l'influence d'informations visuelles sur la perception de l'intensité sonore.

2.3 Temps de réaction audiovisuels

L'utilisation de la chronométrie mentale va permettre d'affiner les modèles décrivant les interactions entre traitements auditif et visuel grâce à une mesure objective et supraliminale. Ces recherches vont donner naissance à des modèles formels et des méthodes permettant de mettre en évidence, dans une certaine mesure, des interactions entre traitements auditif et visuel. Ces études ont également favorisé l'émergence de la notion d'évènement audiovisuel bien défini dans le temps.

2.3.1 Premières études

Hershenson (1962) montre que le temps de réaction (TR) pour détecter un stimulus audiovisuel est inférieur au TR pour détecter le même stimulus présenté séparément dans l'une ou l'autre des modalités auditive ou visuelle (résultat déjà montré par Todd, 1912). La présence de cette facilitation comportementale dépend du délai séparant le stimulus visuel du stimulus auditif (celui-ci arrivant toujours simultanément ou après le stimulus visuel). Afin d'estimer la facilitation pour les différents délais en tenant compte du fait que le TR auditif est inférieur au TR visuel, il confronte ses données à un modèle d'indépendance selon lequel le TR bimodal est déterminé par le TR au premier des deux stimuli détecté (voir la figure 2.1).

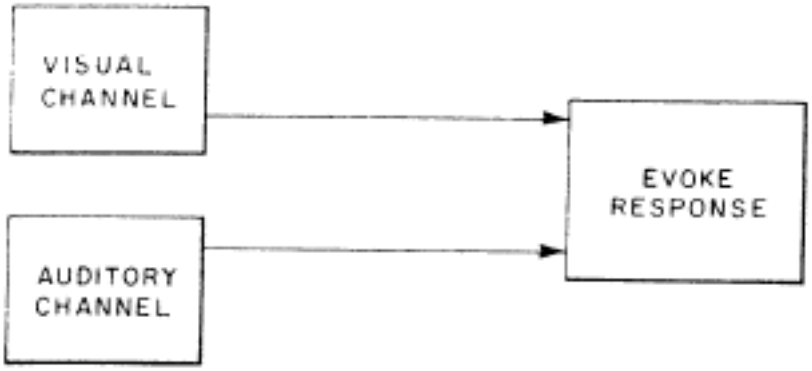

FIG. 2.1 – Illustration du modèle d'indépendance de Hershenson (1962). D'après Nickerson (1973).

Dans ce modèle d'indépendance, le TR bimodal devrait être déterminé par l'un ou l'autre des TR unimodaux selon le délai séparant le stimulus auditif du stimulus visuel : dans les données de Hershenson (1962), le TR auditif moyen est inférieur d'environ 50 ms au TR visuel. Donc pour des délais inférieurs à la différence des TR unimodaux (50 ms), le TR bimodal devrait être égal au TR auditif puisque le stimulus auditif est détecté plus vite. Pour les délais supérieurs, le TR devrait être égal au TR visuel puisque le stimulus visuel est détecté avant le stimulus auditif. La figure 2.2 page suivante présente les gains de TR pour la condition bimodale par rapport à chacune des deux conditions unimodales, en fonction du délai. On peut constater que pour les valeurs de délai autour de 50 ms, les deux gains sont positifs (zone hachurée), ce qui signifie que le TR bimodal ne peut s'expliquer

ni par le TR auditif, ni par le TR visuel. Ces données semblent donc impliquer l'existence d'interactions entre traitements auditif et visuel en ce qu'elles ne semblent pas explicables par des traitements unimodaux indépendants.

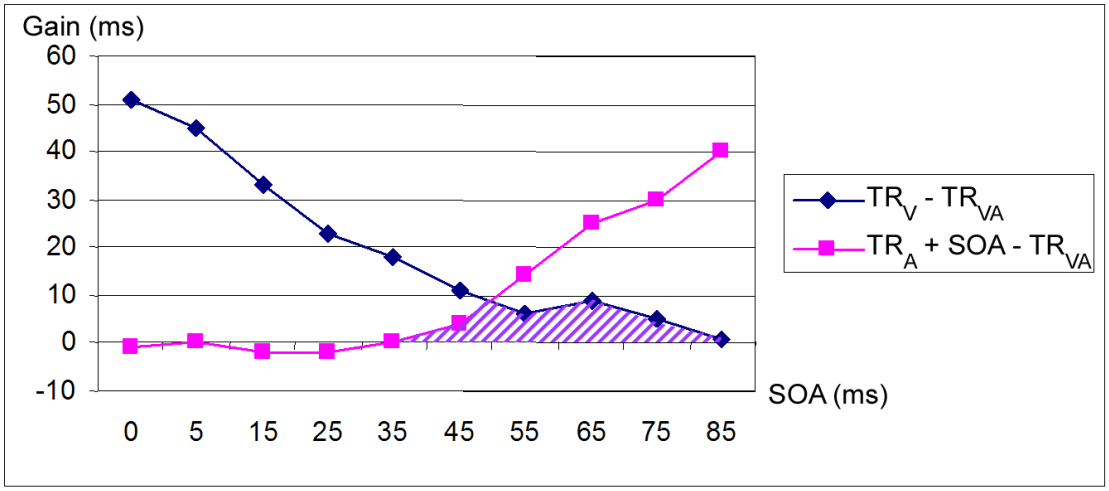

FIG. 2.2 – Facilitation par rapport aux TR unimodaux en fonction du délai séparant le stimulus visuel du stimulus auditif (SOA), sous l'hypothèse que le sujet répond au premier signal traité. La courbe bleue présente le gain de TR en condition bimodale par rapport à la condition visuelle. La courbe rose représente le gain de TR en condition bimodale par rapport à la condition auditive, en tenant compte du fait que le TR bimodal est mesuré à partir du début du stimulus visuel. La partie hachurée correspond à la plage de délais pour laquelle un facilitation bimodale est observée par rapport aux deux TR unimodaux. Figure réalisée à partir des données de Hershenson (1962).

Il faut toutefois garder à l'esprit que le calcul de la facilitation dépend du modèle d'indépendance choisi. Or l'une des caractéristiques du modèle d'indépendance de Hershenson (1962), comme l'a souligné Nickerson (1973), est qu'il suppose l'invariance des temps de traitement d'un essai à l'autre pour une condition donnée : les TR unimodaux et bimodaux sont estimés uniquement par leur moyenne.

Lorsque cette variabilité est prise en compte, elle peut produire ce qu'on appelle une facilitation statistique (Raab, 1962), même dans un modèle d'indépendance : comme le temps de traitement dans chacun des deux canaux unisensoriels présente une certaine variabilité, il en résulte qu'à chaque essai, la détection du stimulus peut être déterminée par le plus court des TR auditif ou visuel. Dans un modèle d'indépendance, la moyenne des temps de traitement audiovisuel sera donc déterminée par la distribution des minima des temps de traitements unimodaux à chaque essai. Or on peut montrer que la moyenne d'une distribution des minima de deux distributions est inférieure à la plus petite des moyennes de ces deux distributions. Raab (1962) montre qu'un modèle d'indépendance prenant en compte la variabilité des temps de traitement peut expliquer le gain bimodal de TR trouvé par Hershenson (1962), et donc que ce gain ne démontre pas l'existence d'interactions entre traitements auditifs et visuels.

Le modèle d'indépendance suppose que le sujet partage son attention entre les modalités auditive et visuelle pour pouvoir répondre à la première des deux cibles. Or deux autres études montrent que la présentation d'un stimulus auditif diminue le TR dans une tâche de détection visuelle, alors qu'il n'apporte aucune information pour la réalisation de la tâche et pourrait donc être ignoré (John, 1964 cité par L. K. Morrell, 1967 ;L. K. Morrell, 1967). De plus cette facilitation peut avoir lieu même si le stimulus auditif suit le stimulus visuel cible (L. K. Morrell, 1967), ce qui semble exclure un pur effet d'alerte. Ces deux études suggèrent que le phénomène de facilitation statistique est insuffisant pour expliquer le gain comportemental apporté par la double modalité et vont donner lieu à une série d'expériences avec ce paradigme, dans lequel un des deux stimulus sera accessoire.

2.3.2 Paradigme du stimulus accessoire

Mise en évidence des interactions dans le paradigme du stimulus accessoire

Les résultats de John (1964) et L. K. Morrell (1967) restent explicables par une facilitation statistique dans le modèle d'indépendance si l'on suppose que le sujet ne respecte pas la consigne et répond indifféremment au stimulus auditif ou visuel. Afin de montrer que de véritables interactions audiovisuelles ont lieu, L. K. Morrell (1968c) introduit des essais pièges auditifs auxquels le sujet doit se garder de répondre. La tâche devient donc une tâche de choix dans laquelle les stimuli visuels et bimodaux demandent une réponse mais non les stimuli auditifs. Bien que les sujets parviennent à effectuer correctement la tâche, une facilitation intersensorielle est toujours observée. Le nombre limité de fausses alertes montre que les sujets ne répondent pas au stimulus auditif et donc que le modèle d'indépendance doit être rejeté, au moins dans le cas d'une tâche visuelle où le stimulus auditif n'est pas informatif.

Ce résultat est confirmé par I. H. Bernstein, Clark et Edelstein (1969a) dans le même paradigme, avec un plus grand nombre de valeurs de délai entre le stimulus visuel et le stimulus auditif (l'auditif suit toujours le visuel) mais aussi une tâche de discrimination spatiale visuelle dans laquelle la présence ou l'absence d'un stimulus auditif n'est pas pertinente (I. H. Bernstein, Clark & Edelstein, 1969b ; I. H. Bernstein & Edelstein, 1971 ; Simon & Craft, 1970). Dans le même ordre d'idée, Taylor et Campbell (1976) ; Taylor (1974) montrent qu'un stimulus auditif, présenté au cours d'une tâche de comparaison d'un stimulus visuel test à un stimulus présenté précédemment, facilite le TR de reconnaissance.

Notons que deux études seulement ont étudié l'effet inverse d'un stimulus visuel accessoire sur le TR auditif de choix (avec essais visuels pièges ; L. K. Morrell, 1968a ; I. H. Bernstein, Chu, Briggs & Schurman, 1973, expérience 2) et ont trouvé des effets de facilitation analogues, quoique moins importants. Posner, Nissen et Klein (1976) trouvent un effet d'un stimulus visuel accessoire beaucoup moins fort que l'effet d'un stimulus auditif accessoire et le met sur le compte d'un pouvoir alertant moins important du stimulus visuel.

Modèles des interactions dans le paradigme du stimulus accessoire

Tous ces résultats indiquent non seulement que des interactions audiovisuelles ont lieu mais aussi que l'influence du stimulus auditif n'est pas spécifique car elle ne peut s'expliquer par sa contribution à la décision visuelle : ce ne sont pas les informations portées par le stimulus accessoire qui sont responsables de la facilitation, mais sa simple présence (et donc le moment de son occurrence). Deux types de mécanismes sont proposés pour rendre compte des effets de facilitation : un mécanisme de sommation énergétique et un mécanisme d'amélioration de la préparation.

- dans le premier, l'énergie portée par les stimuli détermine la vitesse de la réponse. Lorsque deux stimuli sont présentés ensemble, les énergies s'additionnent, ce qui a pour effet de diminuer le TR. La sommation d'énergie a lieu entre les modalités sensorielles, que le stimulus soit pertinent ou non, ce qui n'est pas sans rappeler les théories de l'irradiation (voir partie 2.1.2 page 22).
- dans le second mécanisme, le stimulus accessoire améliore la préparation du sujet à effectuer sa réponse motrice, ce qui dans le cas de la facilitation auditive du traitement visuel est possible parce que le stimulus auditif est traité plus rapidement.

Selon I. H. Bernstein (1970), les deux mécanismes sont également nécessaires pour rendre compte de tous les effets observés. D'une part, la sommation d'énergie permet de rendre compte de l'effet de l'intensité relative des stimuli : l'augmentation de l'intensité du stimulus accessoire auditif augmente la facilitation alors que celle de l'intensité du stimulus cible visuel la décroit car le TR approche un seuil et ne peut plus diminuer (I. H. Bernstein, Rose & Ashe, 1970a, expérience 1). D'autre part, I. H. Bernstein, Rose et Ashe (1970b) montrent que l'efficacité du stimulus accessoire dépend de l'état de préparation du sujet. Dans cette expérience, un signal d'alerte au début de chaque essai induit un certain état de préparation qui varie selon le délai séparant le stimulus d'alerte des stimuli cible et accessoire (*fore period*). Plus le niveau de préparation diminue (le TR visuel augmente) et plus le stimulus accessoire facilite le temps de réaction. I. H. Bernstein et coll. (1970b) en concluent que le stimulus accessoire a un pouvoir préparatoire.

D'un point de vue neurophysiologique, puisque le stimulus auditif ne semble pas influencer la justesse de la réponse visuelle, I. H. Bernstein (1970) considère que ces mécanismes d'interaction doivent nécessairement être parallèles à la voie principale et classique d'analyse du stimulus (la voie géniculo-striée pour la vision). Selon I. H. Bernstein et coll. (1970a) une structure nerveuse candidate pour la sommation d'énergie serait la formation réticulée. De son côté, L. K. Morrell (1968b) a montré que l'amplitude des potentiels évoqués enregistrés en montage bipolaire en regard du cortex moteur controlatéral à la main de réponse entre 120 et 240 ms de traitement est corrélée à la facilitation intersensorielle du TR pour différents délais entre le stimulus accessoire et la cible, ce qui suggère que l'amélioration de la préparation pourrait avoir lieu au niveau du cortex moteur.

Selon Nickerson (1973) néanmoins, on peut se passer de la sommation énergétique. D'une part, ce mécanisme présente des difficultés d'ordre logique : comment, en effet, expliquer qu'un processus parallèle de sommation d'énergie diminue le TR alors que ce dernier dépend avant tout de l'analyse du stimulus dans la mesure où la réponse donnée par le sujet est

généralement juste (nombre de faux positifs limité) : si la sommation d'énergie a lieu avant la fin de l'analyse, la facilitation est impossible sans un nombre important de faux positifs ; si elle a lieu après, elle ne peut plus influencer le TR, sauf à agir au niveau de la préparation de la réponse, ce qui revient à une explication en termes d'amélioration de la préparation.

D'autre part, la sommation énergétique est facilement réductible à l'amélioration de la préparation car l'effet de l'intensité peut a le même effet dans les deux cas : plus le stimulus accessoire est intense, plus il augmente l'état de préparation ; à l'inverse, plus le stimulus cible est intense, plus le TR est rapide et moins le stimulus accessoire peut le diminuer car la réponse est efficace indépendamment de la préparation du sujet.

Un autre argument contre la sommation énergétique est que la facilitation a lieu également pour des stimulus auditifs accessoires qui sont des extinctions de sons continus, ce qui exclut un lien direct entre intensité et énergie (I. H. Bernstein & Eason, 1970, cités par Nickerson, 1973), lien qui peut cependant facilement être remplacé par un lien variation d'intensité/énergie.

Afin de tenter de rendre compte de tous ces résultats, Nickerson (1973) propose un modèle dans lequel les traitements auditifs et visuels peuvent être dirigés soit vers un processus de préparation (stimulus accessoire) soit vers un processus d'évocation de la réponse (stimulus cible) de type énergétique (voir la figure 2.3). Le problème de ce modèle est que le sujet doit choisir a priori de diriger le traitement du stimulus vers l'un ou l'autre des mécanismes.

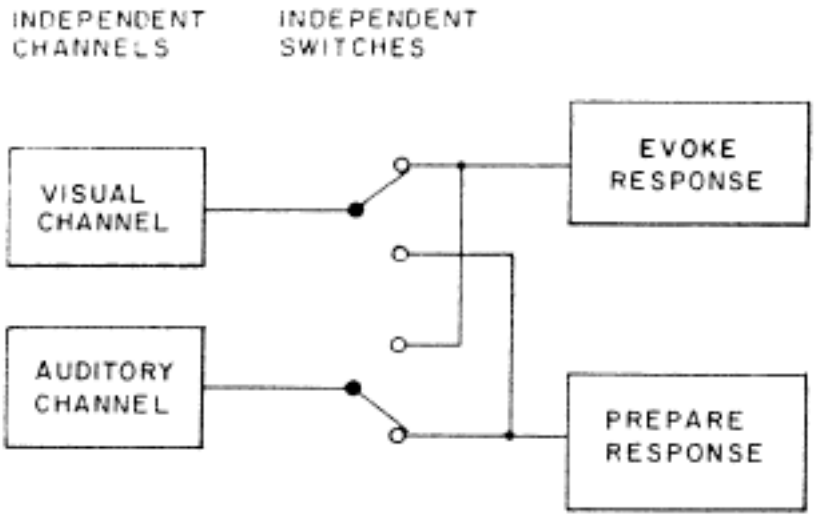

FIG. 2.3 – D'après Nickerson (1973).

Cependant I. H. Bernstein et coll. (1973) montrent que les facteurs d'intensité et de durée de l'avant-période, censés agir respectivement sur des mécanismes énergétiques et de préparation, n'interagissent pas, ce qui suggère qu'ils agissent à des niveaux de traitement différents. Ils trouvent en outre que le nombre de faux positifs augmente avec la facilitation lorsque celle-ci dépend de facteurs d'intensité, mais diminue avec la facilitation lorsqu'elle dépend de la durée de l'avant période, ce qui confirme l'existence de deux mécanismes

indépendants.

Spécificité des interactions dans le paradigme du stimulus accessoire

Les mécanismes proposés pour rendre compte de l'effet de facilitation intersensorielle dans le paradigme du stimulus accessoire préservent un modèle de convergence tardive des voies sensorielles car l'effet de facilitation intersensorielle est attribué à des voies parallèles et non spécifiques. Cette orientation non spécifique est très influencée par le choix du paradigme expérimental utilisé pour étudier la facilitation intersensorielle (l'utilisation d'un stimulus auditif accessoire non pertinent), mis en place à l'origine pour contrer le modèle de facilitation statistique de Raab (1962).

Dans certains protocoles, cependant, la possibilité que le stimulus auditif fournisse des informations pertinentes pour l'analyse du stimulus visuel a été envisagée, même si la portée des résultats obtenus semble avoir échappé aux théoriciens des modèles d'interaction audiovisuelle (partie précédente). Il s'agit d'expériences ayant étudié l'effet de la compatibilité entre les informations spatiales portées pas le stimulus accessoire et le stimulus cible : Simon et Craft (1970) montrent ainsi qu'un stimulus auditif accessoire présenté du même côté que le stimulus visuel cible augmente la facilitation et que cet effet diminue avec le délai séparant les stimuli cible et accessoire. Si des tentatives sont faites pour préserver des modèles d'interactions non spécifiques (I. H. Bernstein & Edelstein, 1971 ; Nickerson, 1973), par exemple en invoquant une spécificité hémisphérique de la sommation énergétique ou de la préparation, elles reviennent en réalité à considérer que ces mécanismes parallèles participent à l'analyse du stimulus.

De plus, ce type de résultat ne se limite pas à la dimension spatiale puisque I. H. Bernstein et Edelstein (1971) montrent un effet analogue de la hauteur tonale sur la rapidité de jugement de hauteur spatiale d'un stimulus visuel (dimensions censées être synesthésiques). Un autre résultat suggérant qu'un stimulus auditif agit directement sur l'analyse visuelle est que la facilitation intersensorielle est plus importante pour l'analyse de stimuli visuels familiers que non familiers (présentés en miroir, Taylor & Campbell, 1976). Aucun modèle convaincant n'est proposé à l'époque pour rendre compte de ces résultats.

Notons que certains auteurs ont ultérieurement attribué ces effets de congruence à des effets de compatibilité stimulus/réponse, et donc à un niveau décisionnel plutôt que sensoriel (Simon, 1982 ; Stoffels, van der Molen & Keuss, 1985 ; Stoffels & van der Molen, 1988 ; Stoffels, van der Molen & Keuss, 1989).

2.3.3 Paradigme d'attention partagée

Falsification du modèle d'activations séparées

Au début des années 80 s'opère un tournant dans l'étude de la facilitation intersensorielle du temps de réaction : le paradigme du stimulus accessoire est presque totalement abandonné au profit du paradigme d'attention partagée, c'est-à-dire celui utilisé originellement par Hershenson (1962, voir la partie 2.3.1 page 31). Deux études (J. O. Miller, 1982 ; Gielen, Schmidt & Van den Heuvel, 1983) montrent que la diminution du temps de réaction, lorsque les sujets doivent détecter un stimulus audiovisuel synchrone, ne peut

s'expliquer par la facilitation statistique dans un modèle d'indépendance. Ces deux études montrent que les TR bimodaux ne peuvent s'expliquer en considérant qu'ils sont déterminés, à chaque essai, par le plus court des traitements auditif ou visuel. Cette démonstration s'appuie sur un modèle d'activations séparées (équivalent au modèle d'indépendance proposé par Hershenson, 1962 et Raab, 1962) : les stimuli auditifs et visuels seraient évalués indépendamment et la première de ces évaluations terminée déclencherait des processus de réponse communs aux deux modalités (voir la figure 2.4).

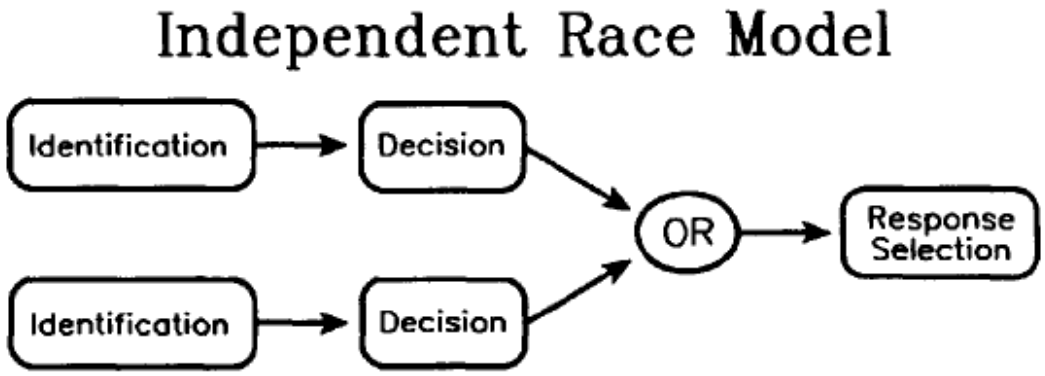

FIG. 2.4 – Modèle d'activations séparées, encore appelé modèle de compétition. Chaque cible auditive ou visuelle est traitée et évaluée indépendamment l'une de l'autre. La première évaluation terminée déclenche la sélection de la réponse et détermine donc le TR bimodal. D'après Mordkoff et Yantis (1991).

Les deux études utilisent des méthodes très proches pour exclure le modèle d'activations séparées, consistant à montrer que la distribution des TR audiovisuels ne peut être prédite par le modèle à partir des distributions des TR unimodaux. C'est la méthode de J. O. Miller (1982, connue sous le nom d'inégalité de Miller) qui va connaitre le plus grand succès puisqu'elle remplace désormais souvent la simple comparaison de la moyenne des TR bimodaux avec le plus court des TR unimodaux pour déclarer que de "véritables" interactions entre modalités sensorielles ont lieu. Le test de l'inégalité de Miller sera décrit en détails dans la partie 7.1 page 99. Contentons nous simplement ici de souligner que la formalisation du test à partir du modèle repose sur un certain nombre de postulats, dont celui d'indépendance au contexte (Colonius, 1990 ; Townsend, 1997), selon lequel il est possible d'estimer la distribution des temps de traitement unimodaux en condition de détection bimodale par la distribution des TR en condition de détection unimodale.

Notons également que le test de l'inégalité de Miller a été appliqué aussi bien à des situations de détection bimodale qu'à des situations de détection unimodale avec plusieurs cibles visuelles, pour tester ce qu'il est convenu d'appeler l'effet du signal redondant (*Redundant Signal Effect, RSE*). Alors que la violation de l'inégalité de Miller semble être quasiment systématique dans le RSE bimodal et a été reproduite à de multiples reprises par la suite, elle est beaucoup moins courante dans le cas unimodal (voir par exemple Eriksen, Goettl, St James & Fournier, 1989).

Modèles de coactivation

Plusieurs modèles alternatifs au modèle d'activations séparées ont été proposés pour rendre compte de la violation de l'inégalité de Miller. La première classe de modèles proposée est celle des modèles de coactivation, dont une version est illustrée dans la figure 2.5.

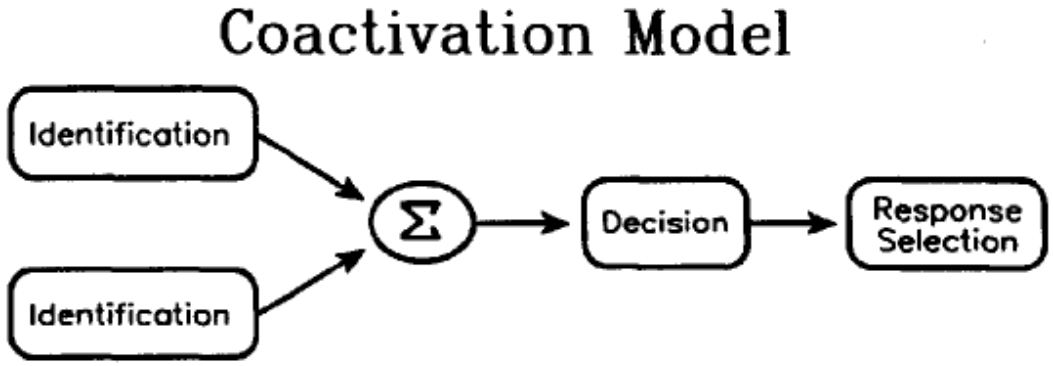

FIG. 2.5 – Modèle de coactivation. D'après Mordkoff et Yantis (1991).

Selon J. O. Miller (1982), la coactivation désigne le fait que les deux sources d'informations, auditive et visuelle, participent à l'accumulation des éléments (*evidence*) permettant le déclenchement des processus de réponse communs aux deux modalités. Cette accumulation est plus rapide si deux sources y participent, ce qui explique l'accélération du TR. Dans la perspective de J. O. Miller (1982), la possibilité d'une coactivation n'est cependant pas limitée au stade de la décision (cas illustré dans la figure 2.5), elle peut aussi avoir lieu au niveau de l'analyse du stimulus ou de la préparation de la réponse. Ainsi, selon lui, le modèle de préparation de Nickerson (1973, voir la partie 2.3.2 page 34) est un modèle de coactivation.

Cette définition est assez large et plusieurs études vont tenter de départager différentes versions du modèle de coactivation. D'abord, la coactivation se distingue de la sommation énergétique en ce qu'elle opère sur les stimuli identifiés comme des cibles. Ainsi, J. O. Miller (1982, expérience 3) montre que la falsification de son inégalité est toujours observée si le sujet doit distinguer une cible d'un distracteur dans les deux modalités. Dans cette expérience tous les stimuli sont bimodaux et le sujet doit répondre si au moins l'une des composantes du stimulus (auditive ou visuelle) est une cible, mais pas si les stimuli auditifs et visuels sont tous les deux des distracteurs. Cette facilitation, que l'on appelle souvent effet de la cible redondante (*Redundant Target Effect, RTE*) confirme que ce n'est pas la simple présence d'un stimulus d'une autre modalité, mais sa signification pour la tâche demandée qui accélère le traitement. Selon J. O. Miller (1982), ce résultat suggère également que la coactivation a lieu au niveau de la décision. À cet égard, ce type de modèle de coactivation rend difficilement compte de la facilitation dans un paradigme de type stimulus accessoire puisque celui-ci n'est pas censé participer à la décision. Mais les explications en termes

de coactivation et de sommation énergétique ne sont pas mutuellement exclusives. En effet, la coactivation est censée avoir lieu entre traitements auditifs et visuels alors que la sommation d'énergie aurait lieu par le biais de mécanismes parallèles. Une expérience de Gondan, Niederhaus, Rösler et Röder (2005) combinant les deux effets suggère que le RTE et le RSE peuvent coexister, le second étant d'amplitude plus importante.

Ensuite, J. O. Miller (1986) tente de distinguer entre des modèles de coactivation accumulative et exponentielle : la coactivation est dite accumulative si les éléments déclenchant une réponse s'accumulent dans le temps, exponentielle si c'est la simple présence simultanée de deux signaux à un instant donné qui permet un TR plus rapide. Les études qui ont fait varier le délai entre les stimuli auditifs et visuels ont montré que la violation de l'inégalité de Miller est maximale lorsque le stimulus auditif suit le stimulus visuel avec un délai comparable à la différence de TR en conditions auditives et visuelles seules (Diederich & Colonius, 1987 ; Giray & Ulrich, 1993 ; J. O. Miller, 1986). Ce résultat est compatible avec les deux type de modèle de coactivation, mais le modèle exponentiel permet des prédictions formelles sur les distributions des TR qui sont falsifiées par les résultats de J. O. Miller (1986). Le modèle de coactivation accumulatif est donc retenu par défaut.

Enfin, J. O. Miller (1991) distingue entre modèles de coactivation dépendant et indépendant. Le modèle représenté dans la figure 2.5 page ci-contre est un modèle indépendant en ce que les canaux n'échangent pas d'information avant leur convergence et l'accumulation de preuves. J. O. Miller (1991, expérience 1) montre que le RSE est plus important si les stimuli auditifs et visuels sont congruents (sur les dimensions synesthésiques de hauteur tonale et hauteur spatiale) et ce, dans une simple tâche de détection dans laquelle ces dimensions ne sont pas pertinentes. Ce résultat n'est pas compatible avec un modèle de coactivation indépendante dans lequel les éléments s'accumulent de façon indépendante et requiert que les canaux sensoriels soient perméables aux informations extraites par l'autre canal sensoriel. La même conclusion s'impose dans l'étude de Gondan et coll. (2005) qui montre que le RTE est plus important pour des cibles spatialement congruentes. Il s'agit donc ici d'une interdépendance informationnelle entre les traitements auditifs et visuels puisque l'informations portée par un stimulus peut modifier le traitement de l'information dans l'autre canal sensoriel.

Plusieurs tentatives de caractérisation mathématique de modèles de coactivation vont être proposées. La caractéristique commune de ces modèles formels est qu'ils nécessitent une discrétisation du processus de coactivation afin d'être appréhendables en termes mathématiques. Dans le modèle de superposition (Schwarz, 1989), l'accumulation d'éléments de preuve par chaque canal sensoriel correspond à un décompte qui doit atteindre un certain critère pour déclencher la réponse pertinente. La superposition des décomptes des deux canaux accélère la vitesse à laquelle ce critère est atteint. Selon Diederich et Colonius (1991), ce modèle explique correctement le RSE trouvé par J. O. Miller (1986) aux différentes valeurs de délai audiovisuel.

Selon J. O. Miller et Ulrich (2003) la coactivation serait équivalente à une facilitation statistique dans un modèle d'activations séparées massivement parallèles : chaque stimulus active un grand nombre de canaux, appelés grains, correspondant chacun à une caractéristique particulière ou codant une partie de l'espace contenant ce stimulus (c'est une

analogie à la fois avec la coexistence d'aires spécialisées parallèles dans le système visuel et leur caractère spatiotopique). Les processus communs de réponse sont déclenchés lorsqu'un nombre défini de grains atteint un certain seuil. Dans ce modèle tous les grains sont activés indépendamment et participent indépendamment à l'apport d'éléments de preuve. Une facilitation apparait parce que le nombre de grains nécessaire au déclenchement sera atteint plus rapidement lorsque le stimulus est redondant puisque le nombre de grains activés est plus grand. Une dérivation mathématique de ce modèle montre qu'il peut rendre compte du RSE dans une tâche de détection intersensorielle.

Autres modèles

Le fait que les différents modèles de coactivations expliquent certaines données ne constitue bien entendu pas la preuve de leur véracité. La falsification de l'inégalité de Miller n'implique en effet pas logiquement un modèle de coactivation, mais seulement le rejet des modèles d'activations séparées. De ce fait, les modèles de coactivations ont été essentiellement définis par défaut, comme ceux susceptibles d'expliquer le RSE.

D'autres modèles ont par la suite été proposés pour rendre compte du RSE et du RTE audiovisuels : Mordkoff et Yantis (1991) reprennent à leur compte la notion d'interdépendance informationnelle des canaux sensoriels, tout en l'appliquant à un modèle d'activations séparées : les canaux sensoriels échangent des informations, mais fournissent des éléments de preuve à des processus de décision séparés : donc bien que des interactions soient possibles à un premier niveau, c'est bien la compétition entre les temps de traitement qui détermine le temps de réaction final.

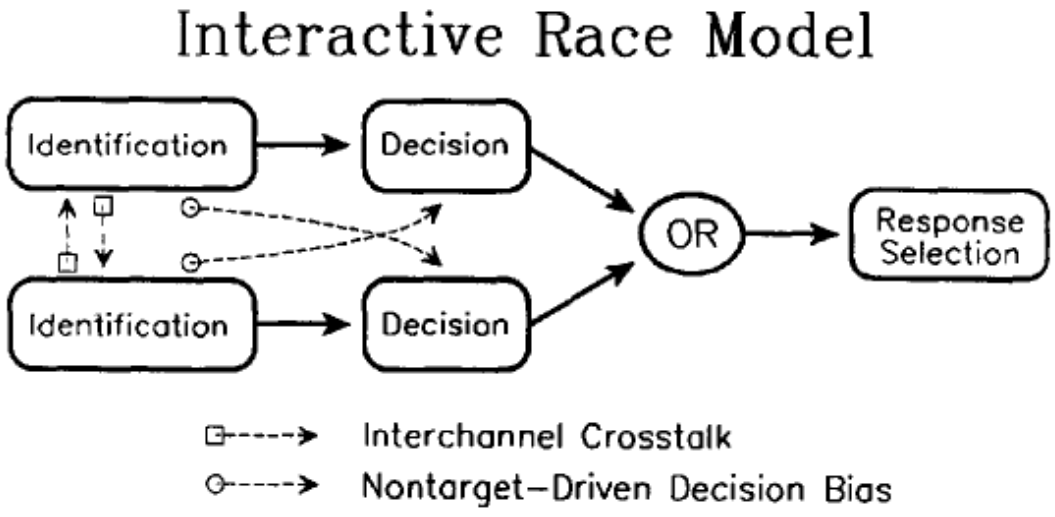

FIG. 2.6 – Modèle de compétition interactif. D'après Mordkoff et Yantis (1991).

Selon ces auteurs, les échanges d'information prennent la forme d'une évaluation de la contingence des stimuli dans les deux canaux : ils montrent qu'au cours des expériences qui ont montré un RSE, certains stimuli étaient associés avec une probabilité plus grande

à certains autres. Ces contingences auraient biaisé l'identification d'un stimulus en fonction de l'identification dans l'autre canal. En supprimant totalement ce biais, Mordkoff et Yantis (1991, expérience 1) parviennent à supprimer le RSE unimodal dans un protocole visuel d'attention partagée. Mais Giray et Ulrich (1993), ainsi que Hughes, Reuter-Lorenz, Nozawa et Fendrich (1994), obtiennent une violation de l'inégalité de Miller dans un protocole audiovisuel alors que le biais était nul ou négatif (un biais négatif devrait ralentir la réponse). Ce résultat montre que l'explication en termes d'évaluation des contingences est insuffisante pour rendre compte du RSE audiovisuel. Il n'empêche que c'est une facteur important, qui plus est, compatible avec les modèles de coactivation : J. O. Miller (1991, expérience 2) montre que la probabilité d'association de paires audiovisuelles de stimuli influence l'amplitude du RSE : les associations les plus fréquentes induisent un RSE plus important que les associations moins fréquentes. Ce mécanisme pourrait aussi expliquer l'influence de la compatibilité entre les stimuli auditifs et visuel (compatibilité spatiale : Gondan et coll., 2005, compatibilité synesthésique : J. O. Miller, 1991, expérience 1) si la perception des contingences audiovisuelles est biaisée par l'expérience préalable des sujets, ce qui n'est pas sans rappeler la théorie directe-réaliste (Rosenblum & Fowler, 1991, voir aussi la partie 2.2.3 page 30). .

Tous les modèles présentés jusqu'à présent ont en commun de considérer les traitements spécifiquement unisensoriels comme parallèles. Townsend (1997) propose un modèle radicalement différent susceptible d'expliquer une violation de l'inégalité de Miller. Ce modèle a trois caractéristiques : il est sériel, en ce que les traitements unisensoriels se succèdent, l'un devant attendre que l'autre soit terminé pour commencer ; il est exhaustif, en ce que l'analyse de chaque stimulus prend un temps déterminé et incompressible ; enfin il pose que le traitement des stimuli distracteurs doit être plus long que celui d'une cible. Du propre aveu de l'auteur, ce modèle est peu plausible, mais la démarche souligne bien le fait que les modèles de coactivation ne sont pas les seuls à pouvoir rendre compte des violations de l'inégalité de Miller.

Niveau de traitement des interactions audiovisuelles

Le point commun de la majorité des modèles alternatifs à l'activation séparée est qu'ils présentent, à une étape donnée, un processus de mise en commun des informations auditives et visuelles. Une question récurrente est savoir à quel niveau de traitement ces interactions ont lieu, la réponse à cette question ayant des conséquences sur le type de modèle pouvant rendre compte de la facilitation.

En général trois niveaux possibles d'interaction ont été envisagés : le niveau sensoriel, le niveau décisionnel et le niveau moteur (préparation et exécution de la réponse). La coexistence du RSE et du RTE (Gondan et coll., 2005) semble indiquer que les interactions peuvent avoir lieu aux deux premiers niveaux. D'autres études suggèrent en revanche que la composante motrice pourrait être affectée par la redondance intersensorielle : Diederich et Colonius (1987) montrent, par exemple, dans un paradigme de double réponse, que la différence de TR pour répondre avec la main droite et la main gauche est affectée par la redondance audiovisuelle, ce qui ne devrait pas être le cas si la composante motrice était une étape totalement indépendante des interactions audiovisuelles. De leur côté Giray et

Ulrich (1993) montrent que la force exercée par le sujet pour effectuer sa réponse motrice est supérieure dans les essais bimodaux.

Notons que les modèles proposés pour rendre compte de la facilitation intermodale en attention partagée ne sont, en général, pas biologiquement contraints, dans la tradition des modèles du fonctionnement cognitif des années 80-90. D'ailleurs ces modèles sont souvent conçus pour rendre compte aussi bien d'effets de redondance intrasensorielle (visuelle) qu'intersensorielle, ce qui en dit long sur l'intérêt porté aux données de la neurophysiologie dans la psychologie expérimentale de l'époque. De plus certaines notions sont définies parfois pour rendre compte uniquement du RSE visuel sans qu'il ne soit envisagé qu'elles soient applicable aux interactions multisensorielles.

Ces modèles peuvent-ils cependant apporter des informations quant à l'architecture des relations entre systèmes auditif et visuel dans le système nerveux central ? Il semble que la plupart des modèles décrits font référence à une représentation au moins implicite du système nerveux. Ainsi la plupart impliquent un point de convergence unique entre les différents canaux sensoriels qui n'est pas sans rappeler ce qu'on a désigné comme le modèle classique de la convergence fondé sur les données neuroanatomiques chez l'animal, même s'ils sont en désaccord sur les mécanismes de cette convergence (compétition contre coactivation). Par ailleurs, l'existence d'une facilitation intersensorielle, qui ne s'explique apparemment pas par cette simple convergence, fait émerger l'idée d'une interdépendance informationnelle qu'on a du mal à ne pas associer à des projections, éventuellement directes, entre systèmes sensoriels. À partir de la fin des années 90, il devient difficile de trouver une étude de l'effet de redondance audiovisuelle qui ne fasse référence à des résultats neurophysiologiques, en particulier aux neurones bimodaux du colliculus supérieur.

2.4 Conflit des indices spatiaux auditifs et visuels

Parallèlement aux études de la facilitation du temps de réaction s'est développé un autre grand courant de recherches qui concerne la résolution de conflit entre indices spatiaux provenant de modalités sensorielles différentes. Les études du conflit spatial intersensoriel permettent d'étudier comment le traitement d'une information spatiale perçue dans une modalité sensorielle (la localisation d'un stimulus) peut influencer les traitements dans une autre modalité d'une information de même type. Il existe cependant plusieurs façons de mesurer l'effet du conflit intersensoriel et ces différentes procédures peuvent aboutir à des mesures reflétant des mécanismes différents. Il est donc important de bien les distinguer. Si, traditionnellement, ces études ont surtout concerné les conflits entre les indices visuels et proprioceptifs, un certain nombre s'est intéressé au conflit entre des indices spatiaux auditifs et visuels (ces résultats sont passés en revue dans : Bertelson, 1998 ; Radeau, 1976, 1994a ; Welch & Warren, 1980, 1986).

Dans une situation de conflit visuo-proprioceptif classique, le sujet porte des lunettes prismatiques qui déplacent le champ visuel, en général d'une dizaine de degrés d'angle visuel. Les indices visuels sont donc en contradiction avec les indices proprioceptifs s'il est permis au sujet d'apercevoir une partie de son corps, en général son bras. Trois effets

liés à différentes phases de la résolution de ce conflit peuvent être mis en évidence lorsque l'on demande au sujet de pointer, soit avec l'autre main (cachée) soit grâce à un dispositif adéquat, vers un stimulus proprioceptif et/ou un stimulus visuel :

- le biais immédiat : c'est l'erreur de pointage vers la partie visuelle ou proprioceptive d'un stimulus bimodal (en général sa propre main visible) commise par le sujet par rapport à une condition contrôle où chaque modalité est présentée seule. Suivant la terminologie introduite par Welch et Warren (1980), on désignera par V(P) l'erreur de pointage vers un stimulus proprioceptif causée par des informations visuelles conflictuelles, et P(V) l'erreur de pointage vers une cible visuelle causée par des informations proprioceptives conflictuelles. Dans toutes les études, V(P) est supérieur à P(V), qui est lui-même proche de zéro : le sujet pointe vers la position apparente de la cible visuelle en étant peu influencé par les informations proprioceptives ; en revanche, la position de la cible proprioceptive est biaisée vers sa position apparente.
- l'adaptation : c'est la réduction de l'erreur de pointage vers l'emplacement réel d'un stimulus au cours du port des prismes, lorsque le sujet s'aperçoit de l'erreur qu'il commet. Elle peut être mesurée comme la différence entre l'erreur de pointage vers une cible, en général visuelle, après une certaine durée du port des prismes et l'erreur de pointage vers cette même cible au début du port des prismes. L'adaptation augmente avec la durée du port des prismes et permet à terme au sujet d'agir de manière efficace sur son environnement conflictuel. Elle n'a d'ailleurs lieu que s'il est permis au sujet d'agir sur cet environnement.
- les effets consécutifs (*after effects*) : c'est la différence entre l'erreur de pointage vers la source réelle mesurée après retrait des prismes et l'erreur de pointage (en général nulle) mesurée avant port des prismes. Les effets consécutifs sont observés dans la direction opposée au déplacement créé par les prismes et sont supposés refléter l'adaptation.

Les premières études d'adaptation au conflit audiovisuel ont été menées pour tester des hypothèses spécifiques issues de l'adaptation au conflit visuo-proprioceptif (Radeau & Bertelson, 1974). Ainsi une explication classique de l'adaptation visuo-proprioceptive est que le sujet, en contrôlant visuellement son bras ou le dispositif de pointage peut comparer les réafférences issues de ce contrôle visuel aux informations efférentes issues des commandes motrices. Cette comparaison permet une recalibration des rapports entre espace visuel et espace proprioceptif (au moins du bras concerné). Or Canon (1970, 1971), puis Radeau et Bertelson (1974) montrent l'existence d'effets consécutifs à la présentation conflictuelle d'indices auditifs et visuels, alors que le sujet effectue ses pointages à l'aveugle, donc en l'absence de réafférences visuo-proprioceptives. Ces résultats montrent donc qu'une adaptation peut avoir lieu pour un conflit spatial audiovisuel, c'est-à-dire purement sensoriel, comme si les sujets cherchaient à faire correspondre leurs espaces auditif et visuel.

2.4.1 Ventriloquie

L'adaptation audiovisuelle résulte cependant d'un effet d'apprentissage qui s'exprime très progressivement et il est donc difficile d'en tirer des conclusions sur les interactions

entre informations spatiales auditives et visuelles lors de la perception d'un évènement audiovisuel. Selon Welch et Warren (1980), la mesure du biais immédiat serait plus informative sur les relations entre les différentes modalités sensorielles dans une situation normale de perception car elle serait exempte d'apprentissage et de stratégies. Le biais immédiat V(A) a été le plus étudié et correspond au phénomène bien connu de ventriloquie, mis en évidence dès 1909 par Klemm (1909, cité par ; Bertelson & Radeau, 1981) puis par un grand nombre d'autres auteurs : la localisation d'un stimulus auditif est biaisée vers sa source visuelle apparente, lorsque celle-ci est déplacée à l'aide de prismes ou par séparation effective des sources, et bien que le sujet doive ignorer les informations visuelles. Cet effet a été mis en évidence dans différentes situations expérimentales :

- en demandant au sujet de pointer vers la source auditive (par exemple : Bermant & Welch, 1976 ; Pick, Warren & Hay, 1969 ; Radeau, 1985 ; Warren, 1979 ; Warren, Welch & McCarthy, 1981, expérience 2) ou de donner une estimation de son excentricité (Warren et coll., 1981, expériences 1 et 3) et en mesurant le biais V(A).
- en demandant au sujet un jugement droite/gauche sur la source auditive : Thomas (1941) puis Warren et coll. (1981, expérience 4) et Radeau et Bertelson (1987) montrent ainsi qu'un stimulus auditif proche du plan médian est jugé plus souvent à gauche s'il est accompagné d'un stimulus visuel à sa gauche et plus souvent à droite s'il est accompagné d'un stimulus visuel à sa droite. Cette mesure est supposée être moins biaisée par des facteurs cognitifs.
- en demandant au sujet si les stimuli proviennent de la même source ou de sources différentes, ou encore s'il fait l'expérience d'une fusion des sources auditives et visuelles (par exemple : Choe, Welch, Guilford & Juola, 1975 ; Jack & Thurlow, 1973 ; Radeau & Bertelson, 1977 ; Thurlow & Jack, 1973 ; Witkin, Wapner & Leventhal, 1952). Cette dernière mesure ne permet pas de quantifier précisément le biais ni de différencier l'influence de la position du stimulus visuel sur la localisation du stimulus auditif V(A) de l'influence de la position du stimulus auditif sur la localisation visuelle A(V), au contraire des deux autres procédures.

Une supériorité de l'effet de ventriloquie V(A) sur le biais inverse A(V) a été obtenue de manière récurrente par tous les expérimentateurs. Le biais A(V) a en fait été beaucoup moins étudié, sans doute à cause de sa faiblesse : lorsqu'il existe, il est beaucoup moins fort que le biais V(A) (Bertelson & Radeau, 1981 ; Warren et coll., 1981). Cet avantage de la capture visuelle a été mis sur le compte, soit de la supériorité de la vision dans les tâches de localisation, soit du fait que les sujets portent naturellement plus leur attention sur la modalité visuelle (Welch & Warren, 1986).

L'effet de ventriloquie et son effet réciproque suggèrent donc que les informations spatiales visuelles peuvent influencer la localisation auditive (et inversement) et donc que les systèmes sensoriels auditif et visuel interagissent. Mais pour aboutir à cette conclusion, encore faut-il montrer que ces biais sont dus à des véritables interactions sensorielles, et non à une propension des sujets à vouloir faire correspondre les sources auditives et visuelles.

2.4.2 Facteurs influençant l'effet de ventriloquie

L'effet de nombreux autres facteurs concernant, soit les stimuli, soit les connaissances du sujet à propos des stimuli, a été étudié, principalement sur le biais V(A), le biais A(V) étant souvent trop faible pour qu'une modulation puisse être mise en évidence. Parmi les facteurs propres aux stimuli (appelés parfois facteurs sensoriels), on trouve :

- la séparation spatiale : les biais V(A) et A(V) augmentent moins vite que la séparation effective des sources, c'est-à-dire qu'exprimée en pourcentage, elle diminue (Bermant & Welch, 1976 ; Bertelson & Radeau, 1981 ; Jackson, 1953 ; Witkin et coll., 1952). Selon certains auteurs, elle disparaitrait presque totalement au-delà de 30° (Jack & Thurlow, 1973 ; Thurlow & Jack, 1973) alors que d'autres l'obtiennent jusqu'à 90° de séparation (Jackson, 1953 ; Witkin et coll., 1952).
- la contigüité temporelle : l'importance de ce facteur a été montrée dans des études qui ont utilisé comme stimuli des flux sonores et visuels. Un décalage de 150 ms (Warren et coll., 1981) ou 200 ms (Jack & Thurlow, 1973 ; Thurlow & Jack, 1973) entre une bande son et la vidéo d'un locuteur diminue le biais. Thomas (1941), puis Radeau et Bertelson (1987), utilisant des flux plus simples de type son pur et flash, montrent que l'effet de ventriloquie est plus important lorsque les flux auditif et visuel sont tous les deux continus, ou tous les deux intermittents, à condition que leur rythme soit identique.
- la saillance : un flux visuel intermittent est capable de capturer un flux auditif continu, mais non l'inverse : cet effet a été mis sur le compte de la saillance du stimulus par Radeau et Bertelson (1987).
- l'intensité relative des stimuli : l'augmentation de l'intensité du stimulus visuel augmente la capture visuelle, alors que l'augmentation du stimulus auditif la diminue (Radeau, 1985).

Parmi les facteurs liés aux connaissances du sujet (appelés parfois facteurs cognitifs), on trouve

- la consigne : Warren et coll. (1981) montrent que les informations concernant la source des stimuli influencent les biais V(A) et A(V) : le biais est plus important si les sujets pensent que la source est la même, que s'ils connaissent le mécanisme destiné à produire le conflit audiovisuel. Lorsqu'une source commune est explicitement suggérée, la somme des biais V(A) et A(V) atteint d'ailleurs presque 100%, ce qui n'est pas le cas lorsqu'aucune consigne de ce type n'est donnée.
- la vraisemblance (*compellingness*) de la situation : Jackson (1953) montre que le biais V(A) est plus grand pour des stimuli naturels (une bouilloire qui siffle) que pour des associations artificielles de flashs et de sons de cloche. De la même façon, Radeau et Bertelson (1977) montrent que l'expérience de fusion audiovisuelle dure plus longtemps pour des sons de percussions accompagnés des mouvements qui les produisent que pour les mêmes sons accompagnés de flashs synchronisés.

Parmi ces facteurs, on peut distinguer ceux qui peuvent influencer l'attention que le sujet va porter à chacune des modalités sensorielles, telle que l'intensité, la saillance, le pouvoir localisateur d'un stimulus par rapport à l'autre, et ceux qui influencent la probabilité que

les stimuli proviennent de la même source, tels que la proximité spatiale, la proximité temporelle, la vraisemblance de la situation et, bien sûr, la présomption d'une source unique. Ce second type de facteurs serait lié à ce que Welch et Warren (1980) appellent le postulat d'unité (*unity assumption*) : selon eux, tous les facteurs, qu'ils soient sensoriels ou cognitifs, qui favorisent le postulat d'unité, augmentent le biais.

2.4.3 Niveau des interactions dans l'effet de la ventriloquie

Si tous les facteurs influençant le phénomène de ventriloquie se ramènent à des phénomènes d'attention et au postulat d'unité, on n'a pas besoin de supposer l'existence d'interactions sensorielles de bas niveau entre traitements spatiaux auditif et visuel. Cependant le fait que le biais immédiat puisse avoir lieu dans des situations très simplifiées avec des flashs et des bips semble suggérer le contraire (Bertelson, 1998 ; Radeau, 1994a), même si ces auteurs admettent que le phénomène puisse être facilité par les croyances du sujet. Une partie importante de leur argumentation est toutefois basée sur des résultats d'adaptation audiovisuelle, qui semble en effet moins sensible aux manipulations purement cognitives (Radeau & Bertelson, 1977, 1978) et aux stratégies délibérées des sujets.

Toutefois, bien qu'à première vue, biais immédiat et adaptation semblent refléter le même phénomène, il est probable qu'ils reflètent des processus ne se recouvrant que partiellement. Selon Welch et Warren (1980) en effet, le biais immédiat est mesuré dans une situation bimodale effective sans que le sujet ne s'aperçoive nécessairement du conflit (Bertelson & Radeau, 1981), alors que l'adaptation mesure la façon dont le sujet apprend à pointer vers la source réelle d'un stimulus unimodal dans une situation où la détection du conflit est nécessaire. L'adaptation et le biais immédiat reflèteraient donc deux processus opposés. La situation ne semble toutefois pas si simple : Radeau (1994b) montre que l'adaptation (mesurée en termes d'effets consécutifs) et le biais immédiat dans une même expérience ne sont pas corrélés et donc, que s'ils représentent effectivement des processus en partie différents, ils ne sont pas antithétiques. Par ailleurs, Bertelson et Radeau (1981, expérience 2) montrent que le biais intersensoriel peut exister même lorsque le conflit entre les indices auditifs et visuels est perçu par le sujet. La situation se complique encore lorsque l'on constate que la plupart des expériences sur la ventriloquie ont confondu le biais immédiat et l'adaptation en utilisant une séparation constante entre indices auditifs et visuels : au fur et à mesure de l'expérience, il est en effet probable que le sujet s'adapte à ce conflit. Toutefois Bertelson et Radeau (1981, expérience 1) montrent qu'en changeant la taille de la séparation à chaque essai, on obtenait toujours un biais V(A).

Quoiqu'il en soit, l'existence d'un biais immédiat purement sensoriel et automatique a été mise en évidence plus récemment, de façon plus convaincante par une expérience de Bertelson et Aschersleben (1998) : dans cette expérience les sujets doivent juger si un stimulus auditif, dont la source est cachée, se situe à droite ou à gauche du plan médian (matérialisé par un trait). Le stimulus auditif est rapproché du centre par une procédure en escalier et on mesure le point où le jugement droite/gauche s'inverse. Ce point arrive plus tôt (est donc plus loin du plan médian) si un stimulus visuel central est présenté en même temps que le son, que si le stimulus visuel est toujours présent ou toujours absent. Selon ces auteurs, l'intégration des informations spatiales auditives et visuelles serait donc automatique et aurait lieu à un niveau très bas de l'analyse des stimuli.

D'autres résultats suggèrent cependant qu'un biais immédiat d'origine purement cognitive peut également exister : dans les expériences de Pick et coll. (1969), Morais (1975) et Weerts et Thurlow (1971), l'effet de ventriloquie est obtenu par la simple suggestion que le stimulus auditif puisse venir d'un haut-parleur factice, le véritable haut-parleur étant caché, et découle donc de la simple connaissance sémantique d'un lien causal entre le haut-parleur et la production de sons. Une autre étude a échoué à reproduire ce résultat (Radeau, 1992).

2.5 Conflit des indices temporels

Alors que la modalité visuelle semble dominer lors de conflits spatiaux, c'est l'inverse qui semble se produire lorsque le conflit met en jeu le traitement d'informations temporelles. Ce biais en faveur des indices temporels auditifs a été le plus souvent étudié en utilisant comme stimuli des flux modulés périodiquement en amplitude, soit dans le domaine lumineux (*flicker*) soit dans le domaine sonore (*flutter*). À la suite de von Schiller (1935, voir partie 2.2 page 25), plusieurs auteurs vont essayer de montrer que la présentation d'un flux sonore influence la fréquence à partir de laquelle le flux lumineux périodique est perçu comme continu (seuil critique ou seuil de fusion). Mais les résultats sont contradictoires, certains n'obtenant aucun effet (Knox, 1945 ; Regan & Spekreijse, 1977), d'autres montrant qu'un changement du seuil de fusion dépend à la fois de la couleur du stimulus utilisé et des caractéristiques du flux sonore (Maier, Bevan & Behar, 1961).

En étudiant la capacité de sujets à faire correspondre la fréquence d'un flux sonore à celle d'un flux lumineux, Gebhard et Mowbray (1959) constatent que les erreurs sont supérieures d'un facteur dix, par rapport à une tâche où les flux à synchroniser appartiennent à la même modalité sensorielle. Les sujets indiquent qu'ils ont l'impression que la variation de la fréquence sonore entraine celle du flux lumineux, alors que celle-ci reste en réalité constante. Mais les auteurs ne parviennent pas à mesurer le phénomène.

Shipley (1964) parvient à mesurer l'amplitude du phénomène en demandant à ses sujets, à partir de deux flux sonores et visuels de même fréquence présentés en synchronie, d'augmenter ou de diminuer la fréquence du flux sonore jusqu'à détecter une asynchronie. La capture auditive de la fréquence visuelle est mise en évidence pour des fréquences supérieures à 4 Hz. Pour une fréquence de départ de 10 Hz, certains sujets peuvent augmenter la fréquence sonore jusqu'à plus de 20 Hz sans détecter de conflit.

Ces résultats sont répliqués par Regan et Spekreijse (1977), puis Myers, Cotton et Hilp (1981). Les données de ces auteurs indiquent que l'illusion visuelle reste stable tant que les sujets fixent le flux lumineux, même si le flux sonore est arrêté. La capture auditive semble plus importante si les stimuli sont présentés en périphérie et ne semble pas dépendre de la séparation spatiale des sources auditives et visuelles (voir aussi : Noesselt, Fendrich, Bonath, Tyll & Heinze, 2005 ; Welch, DuttonHurt & Warren, 1986). L'illusion inverse semble ne pas exister : au contraire, lorsque le sujet modifie la fréquence du flux lumineux, celui-ci semble rester constant et en synchronie avec le flux sonore. Welch et coll. (1986) montrent tout de même qu'il peut exister un faible biais V(A) de la fréquence lumineuse sur la fréquence sonore lorsque l'on compare les jugements de magnitude de la fréquence visuelle dans une condition visuelle seule et une condition audiovisuelle (c'est-à-dire un

paradigme ressemblant plus au paradigme de biais spatial immédiat). Il est cependant beaucoup moins important que le biais A(V).

Un autre cas de dominance auditive dans le domaine de la perception temporelle est rapporté par J. T. Walker, Irion et Gordon (1981) : un stimulus visuel est jugé plus long s'il est accompagné par un stimulus auditif long et plus court s'il est accompagné d'un stimulus auditif court. Par contre la durée d'un stimulus visuel n'influence pas la durée perçue d'un stimulus auditif.

Ces résultats ont essentiellement été interprétés dans le cadre de la théorie de l'appropriation modalitaire (*modality appropriateness*) selon laquelle c'est la modalité sensorielle la plus appropriée pour traiter un type d'information qui domine l'intégration de ces informations entre plusieurs modalités : information spatiale pour la vision, temporelle pour l'audition. Notons que les phénomènes qui ont permis de mettre en evidence ces asymétries dépendent différemment de la correspondance spatiale et temporelle, puisque le phénomène de ventriloquie semble nécessiter une certaine correspondance temporelle des stimuli, alors que la capture auditive de la fréquence d'un flux lumineux semble indépendante de la correspondance spatiale (mais pas de l'excentricité).

2.6 Conclusion

Nous venons de montrer que de nombreux effets résultant de la confrontation d'informations auditives et visuelles pouvaient être mis en évidence dans des paradigmes comportementaux. Chacun de ces résultats correspond à une situation expérimentale particulière et les processus d'intégration audiovisuelle mis en jeu dans chacune de ces situations sont probablement très différents. Certains effets intersensoriels pourraient impliquer des voies parallèles aux voies principales de traitement des stimuli auditifs et visuels, ainsi que des informations de nature peu spécifique, telle que la simple présence et absence d'un stimulus. Mais d'autres semblent impliquer l'existence d'échanges d'informations (spatiales, temporelles, etc...) entre des traitements sensoriels auditifs et visuels. D'autres, enfin, sont liés à des facteurs sémantiques ou cognitifs et pourraient correspondre à une convergence des informations après extraction indépendante des informations auditives et visuelles dans les cortex sensori-spécifiques.

Chapitre 3

Perception audiovisuelle de la parole

La perception de la parole a donné lieu à un nombre particulièrement important d'études concernant les interactions audiovisuelles. En effet, bien que la modalité sensorielle principale de la communication langagière soit l'audition, la vue du locuteur fournit au sujet percevant un nombre non négligeable d'informations susceptibles de participer au décodage du message.

3.1 Contribution des indices visuels à l'intelligibilité de la parole

La première démonstration d'une contribution des indices visuels à la perception de la parole est sans doute celle de Cotton (1935). Dans son expérience, un locuteur se trouve dans une cabine munie d'un double vitrage qui l'isole acoustiquement des sujets. Le son de sa voix est transmis aux sujets par un haut-parleur situé à l'extérieur de la cabine. Le locuteur peut être rendu visible ou invisible au sujet en éclairant ou pas l'intérieur de la cabine. Le message est rendu inintelligible par adjonction d'un bruit intense, si bien que lorsque la lumière est éteinte, les sujets n'en comprennent que quelques mots. Dès que la lumière s'allume cependant, les sujets sont capables de rapporter la quasi intégralité du message, bien que le niveau de bruit reste identique. Malgré l'absence de données chiffrées, l'effet semble particulièrement frappant. Cette amélioration de l'intelligibilité sera quantifiée par (Sumby & Pollack, 1954) en comparant le nombre de mots correctement reconnus dans le bruit en condition auditive et en condition audiovisuelle : excepté pour les conditions les moins bruitées, où la performance atteint un plafond, l'intelligibilité est systématiquement meilleure en condition audiovisuelle. Cette contribution des informations visuelles à la performance augmente avec le niveau de bruit et peut atteindre l'équivalent d'une amélioration du rapport signal sur bruit de 20 dB. L'effet sera répliqué de nombreuses fois (par exemple : Erber, 1969, 1975 ; Neely, 1956 ; MacLeod & Summerfield, 1987).

Si pour des niveaux de bruit où la performance auditive est nulle, l'effet s'explique évidemment par la capacité des sujets à lire sur les lèvres, pour des niveaux de bruit intermédiaires, où le sujet est capable d'extraire à la fois des informations auditives et visuelles, les performances en condition audiovisuelle sont systématiquement supérieures à

celles de l'une ou l'autre des conditions unisensorielles, ce qui montre que les deux types d'information sont utilisés dans le décodage du message. Selon Sumby et Pollack (1954), l'information visuelle fournie serait relativement constante à tous les niveaux de bruit. Beaucoup plus récemment, certains auteurs ont proposé qu'il existe un niveau de rapport signal/bruit (environ -12 dB) pour lequel le gain d'intelligibilité serait maximal (Ross, Saint-Amour, Leavitt, Javitt & Foxe, sous presse) et pour lequel l'intégration audiovisuelle dans la perception de la parole serait donc plus efficace.

Plusieurs causes ou mécanismes de l'amélioration de l'intelligibilité de la parole par les informations visuelles ont été proposés.

3.1.1 Complémentarité des informations auditives et visuelles de parole

La première explication tient à la complémentarité des informations fournies par les modalités auditive et visuelle, en particulier dans les situations où la qualité des stimuli auditifs est dégradée. Cette explication a été avancée essentiellement pour la perception des consonnes : le voisement et la nasalité sont les traits phonétiques des consonnes qui résistent le mieux au bruit. Or ces deux traits phonétiques sont également impossibles à distinguer visuellement. À l'inverse, le lieu d'articulation est un trait phonétique dont la discrimination diminue très rapidement avec le bruit, mais c'est aussi le trait le plus visible (Binnie, Montgomery & Jackson, 1974). Dans une situation de perception audiovisuelle dans le bruit, toutes les informations nécessaires seraient donc présentes, dans une modalité ou une autre, alors que sans bruit, la perception auditive suffit à accéder à toutes ces informations. (Les autres traits phonétiques tels que le mode d'articulation occuperaient une position intermédiaire, visibles dans une certaine mesure, moins dégradés par le bruit que le lieu d'articulation.)

Les traits acoustiques de voisement et de nasalité sont portés essentiellement par des variations d'énergie dans la bande de fréquence du premier formant, alors que le lieu d'articulation correspond à des variations dans la fréquence des deuxième et troisième formants. Lorsque le signal de parole est filtré de manière à ne conserver que la première bande de fréquence, la contribution des informations visuelles à l'identification des consonnes dans le bruit est plus importante que lorsque seule la seconde bande de fréquence est conservée, à intelligibilité équivalente (Grant & Walden, 1996). Ce résultat suggère que lorsque la complémentarité des informations auditives et visuelles est conservée (dans le premier cas, les trois traits phonétiques sont présents), l'amélioration audiovisuelle de l'intelligibilité est plus importante, et donc que cette complémentarité est essentielle dans la perception audiovisuelle de la parole. Toutefois lorsque l'intelligibilité est mesurée sur des phrases entières, l'amélioration audiovisuelle pour des bandes de fréquence d'intelligibilités équivalentes ne varie pas (Grant & Braida, 1991), ce qui suggère que le phénomène n'est pas réductible à la complémentarité des informations.

En ce qui concerne la perception des voyelles, une complémentarité spécifique semble exister puisque les voyelles les plus difficiles à discriminer dans le bruit sont celles qui se lisent le mieux sur les lèvres (Benoit, Mohamadi & Kandel, 1994). Cette complémentarité se retrouve au niveau des traits articulatoires définissant l'espace des voyelles (Robert-Ribes,

Schwartz, Lallouache & Escudier, 1998). Notons également que le contexte voyellique a une influence sur la résistance des consonnes au bruit et sur l'amélioration de l'intelligibilité des consonnes par la modalité visuelle (Benoit et coll., 1994).

3.1.2 Redondance des informations auditives et visuelles de parole

Un autre mécanisme pouvant expliquer en partie l'amélioration de l'intelligibilité par les informations visuelles a été identifié plus récemment : il s'agit d'une diminution du seuil de détection de la parole en condition audiovisuelle par rapport à une condition auditive seule (Grant, 2001 ; Grant & Seitz, 2000), mise en évidence au-dessous du seuil d'intelligibilité. L'hypothèse est que c'est l'amélioration de la détection du signal de parole qui permet l'amélioration de l'identification. Grant et Seitz (2000) montrent que cette amélioration de la détection est d'autant plus importante qu'il existe une corrélation entre la variation dans le temps de l'ouverture de la bouche et le signal acoustique. Cette corrélation est, de façon générale, maximale dans la bande de fréquence des 2ème et 3ème formants et il a été par la suite montré que la diminution du seuil est plus importante dans cette bande de fréquence que dans celle du premier formant (Grant, 2001). Il est donc probable que cette corrélation temporelle soit à l'origine de l'amélioration de la détection. Kim et Davis (2003) montrent que la diminution du seuil peut avoir lieu même lorsque le signal à détecter est prononcé dans une langue inconnue des sujets, ce qui suggère que cette corrélation est en partie suffisante pour expliquer la diminution du seuil de détection.

Plusieurs aspects de cette corrélation temporelle peuvent expliquer la diminution du seuil : les signaux pourraient se renforcer mutuellement et dépasser ainsi le niveau de bruit, ou le sujet pourrait exploiter le fait que les moments d'ouverture maximale de la bouche précèdent de quelques dizaines de millisecondes les pics d'énergie dans la bande de fréquence des 2ème et 3ème formants afin d'augmenter la probabilité de détection d'un signal. Kim et Davis (2004) montrent que l'inversion dans le temps des signaux auditifs et visuels, qui supprime notamment l'avance temporelle de l'ouverture de la bouche sur les pics d'énergie, tout en conservant la corrélation globale, empêche la diminution du seuil. Cependant l'explication en termes d'avance temporelle seule est insuffisante parce que si le signal visuel est décalé de façon à devancer à nouveau le signal auditif dans ces stimuli inversés, la diminution du seuil ne réapparait pas.

Est-ce que cette diminution audiovisuelle du seuil de détection rend réellement compte de l'amélioration audiovisuelle de l'intelligibilité dans le bruit ? Il se pourrait en effet, qu'au seuil d'intelligibilité, les facteurs expliquant l'amélioration du seuil de détection ne jouent plus. Les résultats d'une étude de Schwartz, Berthommier et Savariaux (2004) suggèrent pourtant que les deux sont liés : dans une situation où les indices visuels n'apportent aucune information phonétique permettant l'identification d'une syllabe (en l'absence donc de complémentarité entre les indices visuels et auditifs), la corrélation temporelle audiovisuelle suffit à augmenter l'intelligibilité du voisement dans le bruit.

3.1.3 Facteurs liés à la connaissance de la langue

D'autres facteurs que la complémentarité et la redondance des informations rendent compte d'une partie de l'amélioration audiovisuelle de l'intelligibilité. Il s'agit de facteurs

liés à la connaissance des contraintes linguistiques, notamment phonologiques et/ou lexicales, du signal de parole : la diminution du seuil de détection en condition audiovisuelle est ainsi plus importante pour des sujets ayant une connaissance de la langue que pour ceux à qui elle est inconnue (Kim & Davis, 2003). Et elle peut également être obtenue si les sujets connaissent la phrase à détecter, même si, dans ce cas, la diminution est beaucoup moins importante qu'avec les indices articulatoires visuels (Grant & Seitz, 2000). Ces effets peuvent avoir lieu soit parce que les informations visuelles interagissent directement avec des niveaux de traitement lexicaux ou sémantiques permettant des effets descendants sur les mécanismes de détection auditifs, soit parce que la connaissance des contraintes potentialise le gain audiovisuel à bas niveau. Cette dernière possibilité est suggérée par le fait que la réduction du nombre de réponses possibles augmente l'amélioration de l'intelligibilité en condition audiovisuelle (Sumby & Pollack, 1954).

Certaines études ont montré qu'une facilitation audiovisuelle du traitement de la parole pouvait se manifester en l'absence de dégradation du signal auditif, c'est-à-dire lorsque les indices visuels ne contribuent a priori ni à la détection, ni à l'intelligibilité du message. Ainsi les performances dans la compréhension d'un texte complexe d'un point de vue sémantique ou syntaxique, lu dans des conditions acoustiques garantissant une intelligibilité parfaite, sont meilleures lorsque les sujets voient le visage du locuteur (Arnold & Hill, 2001 ; Reisberg, McLean & Goldfield, 1987). Ces résultats suggèrent que les indices visuels peuvent être pris en compte à tous les niveaux de traitement d'un stimulus de parole.

3.2 Effet McGurk

La première démonstration d'une influence des indices articulatoires visuels sur la perception d'un signal de parole parfaitement distinct a en fait été celle de McGurk et McDonald (1976) : dans leur expérience, une syllabe auditive commençant par une consonne bilabiale (par exemple /ba/) présentée de manière synchrone avec les mouvements articulatoires d'une syllabe vélaire (par exemple /ga/) est perçue dans une proportion importante des essais comme commençant par une consonne alvéolaire (/da/). Cet "effet McGurk", obtenu en dépit du fait que les sujets sont informés de l'incongruence, est devenue emblématique de la perception audiovisuelle de la parole car il montre que les informations auditives et visuelles sont naturellement intégrées.

L'aspect le plus marquant de l'illusion McGurk est le fait que le phonème perçu diffère de ceux spécifiés respectivement par l'une ou l'autre des modalités sensorielles (phénomène de fusion). Cela ne doit pas faire oublier que dans un nombre non négligeable d'essais, le sujet entend l'une des syllabes unimodales et que l'association inverse d'une bilabiale auditive et d'une vélaire visuelle est le plus souvent perçue comme une combinaison des consonnes auditives et visuelles (/bga/).

Le phénomène de fusion se généralise à un certain nombre d'autres associations de consonnes que celle découverte par McGurk et McDonald (1976) :

- l'association d'une bilabiale auditive (/b/, /p/ ou /m/) et d'une vélaire visuelle (par exemple /g/ ou /k/) est perçue comme une alvéolaire (/d/, /t/ ou /n/) ou comme une vélaire (par exemple McGurk & McDonald, 1976).

- une bilabiale auditive associée à une alvéolaire visuelle peut être perçue comme alvéolaire ou linguodentale (/ð/)(par exemple Massaro & Cohen, 1983).
- une bilabiale auditive et une labiodentale visuelle (/v/ ou /f/) peuvent être perçues comme une labiodentale (par exemple Rosenblum & Saldaña, 1992).

Toutes ces paires audiovisuelles ont en commun d'associer des consonnes différant sur leur lieu d'articulation : la syllabe auditive correspond à une articulation bilabiale et la syllabe visuelle à un lieu d'articulation en arrière des lèvres. Le lieu d'articulation entendu lors de la fusion correspond soit à un lieu d'articulation intermédiaire entre ceux spécifiés par les indices auditifs et visuels, soit au lieu d'articulation spécifié par les indices visuels (dans ce dernier cas, on ne peut pas véritablement parler de fusion, mais il a souvent été utilisé pour étudier des variables affectant l'effet McGurk : J. A. Jones & Jarick, 2006 ; Rosenblum & Saldaña, 1992, 1996, etc...).

3.2.1 L'hypothèse VPAM

Dans toutes les illusions de type McGurk rapportées dans littérature, le lieu d'articulation semble donc jouer un rôle important. La première hypothèse avancée pour rendre compte de cet effet (McGurk & McDonald, 1976 puis MacDonald & McGurk, 1978) est connue sous le nom de VPAM (*Visual : Place, Auditory : Manner*). Cette hypothèse part du constat que la vision permet principalement de distinguer un lieu d'articulation antérieur (bilabial) d'un lieu d'articulation plus postérieur (alvéolaire ou vélaire), alors que le lieu d'articulation est justement le trait acoustique le moins discriminable (dans le bruit : voir par exemple Binnie et coll., 1974). Tous les autres traits phonétiques sont mieux spécifiés par l'audition (la manière désigne en fait ici à la fois le mode, la nasalité, le voisement, etc...). Dans cette hypothèse, l'effet McGurk s'expliquerait par le fait que dans le cas de la perception audiovisuelle de la parole, la vision spécifie le lieu d'articulation et l'audition tous les autres traits phonétiques. Mais cette théorie, qui est plutôt une première hypothèse de travail, ne rend pas compte d'un certain nombre de caractéristiques de l'illusion, notamment l'existence des combinaisons, comme le constatent les auteurs de cette hypothèse eux-mêmes (MacDonald & McGurk, 1978).

Ainsi que le souligne Summerfield (1987), même si le lieu d'articulation est difficile à discriminer dans le bruit, il reste intelligible dans de bonnes conditions acoustiques. Par ailleurs, la parole est compréhensible sans la vision. Il n'y a donc pas de raison que les sujets n'exploitent pas l'information auditive disponible sur le lieu d'articulation, à moins de considérer la perception audiovisuelle comme un mode particulier de perception de la parole. Plusieurs expériences montrent d'ailleurs que les lieux d'articulation auditifs et visuels sont pris en compte dans la perception de syllabes audiovisuelles incongruentes (Summerfield, 1979, expérience 2 ; Massaro & Cohen, 1983). Il semble en fait que les sujets tirent parti de toutes les informations auditives et visuelles disponibles, mais qu'ils le fassent en exploitant également les connaissances (implicites) qu'il ont des contraintes articulatoires de l'appareil phonatoire (Summerfield, 1979), comme cela avait déjà été suggéré par McGurk et McDonald (1976) : ainsi le lieu d'articulation perçu (entendu) doit être compatible avec les lieux d'articulation spécifiés par les indices auditifs et visuels et ceci se fait souvent au détriment des indices auditifs du lieu d'articulation, car la présence ou l'absence d'une articulation bilabiale visuelle impose de fortes contraintes sur les sons qu'il est possible de

produire (voir aussi Massaro, 1993).

3.2.2 Intégration audiovisuelle pré-phonologique

Une caractéristique de l'hypothèse VPAM (et d'autres modèles, voir partie 3.4 page 58) est que le processus d'intégration a lieu après que les traits phonétiques aient été catégorisés, c'est-à-dire qu'une segmentation phonologique aurait lieu indépendamment dans les modalités auditive et visuelle, avant convergence audiovisuelle. Toutefois, plusieurs expériences montrent que des indices visuels peuvent influencer le processus de catégorisation phonémique auditive, et donc que l'intégration des informations auditives et visuelles doit avoir lieu avant cette catégorisation.

La première démonstration d'intégration audiovisuelle pré-catégorielle (K. P. Green & Miller, 1985, répliqué par Brancazio & Miller, 2005) n'utilisait pas l'effet McGurk : elle consistait à montrer que la vitesse d'articulation d'une syllabe visuelle influençait la catégorisation de syllabes auditives ambigües sur leur voisement (appartenant à un continuum /ba/-/pa/). Dans une expérience utilisant l'effet McGurk, K. P. Green et Kuhl (1989) montrent qu'une vélaire visuelle (/igi/) associée à des syllabes auditives ambigües sur leur voisement (/ibi/-/ipi/) non seulement donne l'illusion aux sujets de percevoir des consonnes alvéolaires (effet McGurk), mais déplace également la frontière de catégorisation du voisement. Brancazio, Miller et Paré (2003) reproduisent ce résultat et montrent que, non seulement la frontière, mais également le meilleur représentant des non-voisées, se déplacent le long du continuum sous l'influence des indices visuels. Dans le même ordre d'idée, K. P. Green et Kuhl (1991) montrent que le lieu d'articulation visuel influence la vitesse de discrimination du voisement (auditif) et réciproquement dans un paradigme d'interférence de Garner (voir partie 2.2.2 page 27).

Une autre façon de montrer que les segmentations phonétiques auditive et visuelle ne sont pas indépendantes est d'étudier l'effet de la coarticulation sur l'intégration audiovisuelle, en l'occurrence, l'effet McGurk : si l'intégration des consonnes auditives et visuelles est post-catégorielle, la nature de la voyelle qui précède ou qui suit la consonne ne devrait pas modifier l'effet McGurk. Or K. P. Green, Kuhl, Meltzoff et Stevens (1991) montrent qu'une syllabe McGurk classique génère significativement plus de réponses linguodentales (/ð/) dans un contexte voyellique /a/ et plus de réponses alvéolaires (/d/) avec un contexte voyellique /i/. De même, l'incompatibilité des voyelles suivant les consonnes dans une syllabe McGurk (par exemple /da/ associé à /gi/) diminue le nombre de fusions (K. P. Green & Gerdeman, 1995 ; Munhall, Gribble, Sacco & Ward, 1996, expérience 1). Une analyse de la variation d'ouverture de la bouche (Munhall et coll., 1996) montre que l'amplitude d'ouverture est plus faible en contexte /i/ qu'en contexte /a/, ce qui pourrait en partie expliquer cette différence.

Si les informations visuelles peuvent pénétrer le processus de catégorisation, d'autre processus auditifs semblent cependant imperméables à l'effet McGurk, et donc, par extension, à l'intégration audiovisuelle : après exposition prolongée à l'une des consonnes extrêmes d'un continuum phonétique (par exemple /ba/-/da/), la frontière catégorielle se déplace

vers cet extrême : c'est le phénomène d'adaptation sélective. Si on expose les sujets à une syllabe McGurk ayant un /b/ auditif et un /g/ visuel, donc perçue comme /d/, la frontière se déplace vers le phonème spécifié par les indices acoustiques (/b/) et non vers celui perçu (/d/) et l'effet est de même amplitude qu'en condition auditive seule (Roberts, 1987 ; Roberts & Summerfield, 1981 ; Saldaña & Rosenblum, 1994, avec /b/ et /v/). L'absence d'adaptation sélective à un percept illusoire McGurk suggère que l'intégration audiovisuelle a lieu après le stade de traitement correspondant au phénomène d'adaptation, qui serait d'assez bas niveau (Schwartz, Robert-Ribes & Escudier, 1998, p 96).

Cependant, certaines données suggèrent que cette absence d'effet pourrait être due à un contre-effet de recalibration auditive : tout comme l'exposition à des stimuli audiovisuels spatiaux conflictuels (voir la partie 2.4 page 43), l'exposition à une syllabe McGurk pourrait déplacer la frontière catégorielle dans un sens opposé à l'adaptation sélective (Bertelson, Vroomen & de Gelder, 2003). Une étude récente (Vroomen, Linden, de Gelder & Bertelson, 2007) a cherché à séparer ces deux effets et suggère qu'une adaptation sélective à l'illusion McGurk pourrait émerger, plus lentement cependant que les effets de recalibration (voir aussi Vroomen, Linden, Keetels, de Gelder & Bertelson, 2004). À l'appui de cette hypothèse, dans une étude de l'effet d'ancrage (qui ressemble fort au phénomène d'adaptation sélective) de syllabes McGurk audiovisuelles, le déplacement de la frontière catégorielle était plus important dans la condition audiovisuelle que dans la condition auditive seule (Shigeno, 2002).

3.2.3 Influence des facteurs linguistiques et cognitifs

Tous ces résultats concourent à montrer que l'intégration des indices auditifs et visuels de parole dans l'effet McGurk peut avoir lieu avant toute catégorisation en un code linguistique (phonétique), et pourrait éventuellement influencer des processus acoustiques de bas niveau (adaptation sélective). Néanmoins, cela ne signifie nullement que l'intégration doive se limiter à ce niveau pré-linguistique.

Si une première étude a semblé montrer que l'effet McGurk était plus difficile à obtenir lorsque la consonne faisait partie d'un mot (Easton & Basala, 1982), ce qui suggérait une influence du traitement lexical sur l'intégration audiovisuelle, d'autres ont obtenu un effet McGurk robuste dans des mots en choisissant plus judicieusement leurs stimuli (Dekle, Fowler & Funnell, 1992). Une étude de Sams, Manninen, Surakka, Helin et Kättö (1998) échoua à montrer un effet de la lexicalité ou du contexte sémantique en comparant des mots audiovisuels incongruents donnant soit un mot soit un pseudo-mot par effet McGurk : le nombre de fusions est aussi grand que le mot existe ou non, et s'il existe, qu'il soit induit par le contexte ou non.

Cependant des études plus récentes ont montré des effets significatifs de ces deux variables : Windmann (2004) a montré que des pseudo-mots auditifs et visuels, mais dont la fusion donne un mot, sont plus souvent fusionnés lorsqu'ils sont induits par le contexte sémantique. Brancazio (2004) a montré que les indices auditifs et visuels avaient d'autant plus de chance d'influencer la perception d'un mot qu'il font respectivement partie d'un mot plutôt que d'un pseudo-mot. Ces résultats montrent que l'effet McGurk, et donc l'intégration audiovisuelle de la parole, ne sont pas impénétrables par les traitements lexicaux et sémantiques.

D'autres facteurs traditionnellement considérés comme cognitifs peuvent également influencer l'effet McGurk, par exemple l'attention endogène. Tiippana et Andersen (2004) montrent que le fait de porter son attention sur un objet traversant le visage réduit la contribution des indices visuels à l'illusion, alors que la performance en lecture labiale ne varie pas. Alsius, Navarra, Campbell et Soto-Faraco (2005) montrent que la réalisation d'une tâche concurrente auditive ou visuelle diminue le nombre de fusions McGurk.

Répétons tout de même que l'effet McGurk, en dépit du fait qu'il est rarement obtenu dans 100% des essais, reste un phénomène relativement automatique qui se manifeste même si les sujets sont informés de l'incongruence. Le fait que cet effet soit robuste sur le plan phénoménologique n'a d'ailleurs pas favorisé la vérification expérimentale de cette automaticité. Quelques études se sont cependant attachées à montrer que l'effet McGurk pouvait être obtenu avec des méthodes excluant un biais de réponse : Rosenblum et Saldaña (1992, expérience 1) montrent ainsi que le percept illusoire McGurk (auditif /ba/-visuel /fa) est jugé acoustiquement plus ressemblant à la syllabe auditive correspondant à la syllabe illusoire (/va/) qu'à la syllabe correspondant acoustiquement à sa dimension auditive (/ba/). Soto-Faraco, Navarra et Alsius (2004) montrent, avec d'autres syllabes, que cela reste vrai même si les sujets ne jugent pas directement la ressemblance, mais qu'elle intervient dans un paradigme d'interférence de Garner (2.2.2 page 27) en tant que dimension non pertinente pour la tâche à réaliser. Ces deux expériences suggèrent que la dimension audiovisuelle intégrée prend automatiquement le pas sur la dimension auditive dans l'expérience subjective du sujet. Par ailleurs, certains facteurs diminuant la probabilité que les indices auditifs et visuels proviennent du même locuteur (une syllabe prononcée par une voix féminine associée à la vidéo d'un visage masculin) ne diminuent pas l'effet McGurk (K. P. Green et coll., 1991, voir cependant S. Walker, Bruce & Omalley, 1995 pour l'effet d'une autre variable cognitive sur l'effet McGurk).

Un certain nombre de caractéristiques de l'intégration audiovisuelle de la parole peuvent donc être déduits des études de l'effet McGurk. Il faut toutefois garder à l'esprit que cette illusion ne représente qu'un aspect de l'intégration audiovisuelle de la parole : celui de la perception des consonnes, et uniquement de celles qui présentent un lieu d'articulation externe et donc visible. Il n'y a pas a priori de raison de penser que les facteurs affectant l'intégration audiovisuelle aux abords de telles consonnes soit différents de ceux affectant l'intégration audiovisuelle de la parole en général. Quelques études ont montré une influence des indices visuels sur la catégorisation de voyelles dont l'identité visuelle est relativement bien identifiable (Lisker & Rossi, 1992 ; Summerfield & MacGrath, 1984) ; mais le phénomène est beaucoup plus faible que la fusion dans l'illusion McGurk (voir aussi Massaro, 1993).

3.3 Facteurs spatiaux et temporels

Une différence entre l'intégration audiovisuelle de la parole et les autres domaines décrits dans le chapitre 2 est la résistance apparente de l'intégration des indices auditifs et visuels de parole aux conflits spatiaux et temporels.

C'est l'effet de la séparation temporelle qui a été étudié le plus tôt, d'abord pour étudier

l'éventuel effet délétère de l'introduction d'un délai dans des prothèses acoustiques sur l'aide apportée par la lecture labiale aux personnes malentendantes (McGrath & Summerfield, 1985 ; Pandey, Kunov & Abel, 1986). La première évaluation du seuil auquel l'asynchronie entre indices auditifs et visuels de parole est détectée (Dixon & Spitz, 1980) montre que les sujets sont insensibles à un retard du signal visuel d'environ -130 ms et un retard du signal auditif d'environ 260 ms pour le discours continu, alors que ces valeurs sont de -75 et 190 ms pour le film d'un marteau frappant un clou. Ces valeurs sont bien supérieures à celles trouvées pour des stimuli auditifs et visuels simplifiés dont le temps d'attaque est relativement abrupt, qui sont de l'ordre de 20 ms (Hirsh & Sherrick, 1961). Cependant certaines études ont trouvé une tolérance équivalente pour les sons de parole et les stimuli non langagiers (Conrey & Pisoni, 2006 ; Vatakis & Spence, 2006b) ou une tolérance plus faible pour l'asynchronie des sons de parole, surtout pour des syllabes isolées (Vatakis & Spence, 2006b), ou des stimuli de parole simplifiés (McGrath & Summerfield, 1985, expérience 2). Il semble en fait que la tolérance à la désynchronisation dépende non seulement de la nature du signal audiovisuel (avec une tolérance plus grande pour la musique par exemple) mais également de la complexité et de la durée des stimuli, avec des tolérances plus faibles pour les stimuli les plus simples (Vatakis & Spence, 2006a). Un autre facteur pouvant expliquer les différences tient aux différentes techniques d'estimation du seuil utilisées (estimation directe : Dixon & Spitz, 1980 ; méthodes des limites : McGrath & Summerfield, 1985 ; expérience 2, méthode des stimuli constants avec jugement d'asynchronie : Vatakis & Spence, 2006a ou d'ordre temporel : Vatakis & Spence, 2006b).

Selon plusieurs études, il existerait une certaine correspondance entre les seuils de détection de l'asynchronie et la fenêtre temporelle dans laquelle l'effet McGurk (J. A. Jones & Jarick, 2006) ou l'amélioration audiovisuelle de l'intelligibilité de la parole (Grant, van Wassenhove & Poeppel, 2004) sont maximums. Les estimations des bornes de cette fenêtre d'intégration varient entre 0 et -60 ms pour le retard visuel et 120 et 240 ms pour le retard auditif (J. A. Jones & Jarick, 2006, expérience 1 ; amélioration de l'intelligibilité : McGrath & Summerfield, 1985 ; effet McGurk : Munhall et coll., 1996 ; Pandey et coll., 1986 ; van Wassenhove, Grant & Poeppel, 2007). Toutefois, les indices visuels peuvent encore être exploités au moins jusqu'à 300 ms de désynchronisation pour augmenter l'intelligibilité de la parole dans le bruit (Pandey et coll., 1986), et une certains nombre de fusions ou de combinaisons McGurk ont lieu pour des désynchronisation pouvant aller jusqu'à 360 ms (Munhall et coll., 1996) et même 500 ms (Massaro, Cohen & Smeele, 1996 ; van Wassenhove et coll., 2007).

L'asymétrie entre la tolérance aux retards auditifs et visuels a été régulièrement retrouvée et pourrait être due au fait que les indices visuels précèdent naturellement les indices auditifs pour un phonème donné : le fait d'avancer le son par rapport à l'image briserait la correspondance phonétique plus rapidement que l'inverse (Cathiard & Tiberghien, 1994). Selon ces auteurs, et d'autres (McGrath & Summerfield, 1985 ; Pandey et coll., 1986), la durée de la fenêtre de tolérance ou d'intégration correspondrait grosso modo à la durée moyenne d'une syllabe (voir cependant Munhall et coll., 1996). Soulignons toutefois qu'une telle asymétrie peut exister aussi pour des stimuli non langagiers, bien que la direction de l'asymétrie varie d'un stimulus à l'autre (Vatakis & Spence, 2006a, 2006b)

Les données sur l'effet de la séparation spatiale des stimuli auditifs et visuels sur l'intégration audiovisuelle de la parole sont plus éparses et ont uniquement concerné l'effet McGurk. L'illusion semble résister à des séparations allant jusqu'à 180° lorsque le stimulus visuel est présenté au centre du champ visuel (le stimulus auditif est donc présenté derrière le sujet J. A. Jones & Jarick, 2006 ; J. A. Jones & Munhall, 1997). Lorsque le stimulus auditif est présenté devant le sujet et que c'est l'excentricité du visage qui augmente, le nombre de fusions diminue sans toutefois s'annuler jusqu'à 60°. Mais cette diminution est probablement liée à la perte de résolution du système visuel avec l'excentricité (Paré, Richler, ten Hove & Munhall, 2003).

Peut-on en conclure pour autant que l'intégration audiovisuelle de la parole est fondamentalement différente des autres formes d'intégration audiovisuelle ? La taille de la fenêtre temporelle d'intégration semble dépendre au moins autant de la structure temporelle des stimuli auditifs et visuels que du fait qu'il s'agisse de parole ou non. Concernant la largeur de la fenêtre spatiale, elle pourrait s'expliquer par un effet de ventriloquie particulièrement fort dans le cas de la parole. En effet la corrélation temporelle importante existant entre les indices auditifs et visuels de la parole semble pouvoir donner lieu à des effets de ventriloquie particulièrement robustes qui peuvent même structurer l'espace auditif dans lequel s'exprimeront des mécanismes auditifs spécifique tels que l'attention spatiale auditive (Driver, 1996, réplication partielle par Rudmann, McCarley & Kramer, 2003).

3.4 Modèles de perception de la parole audiovisuelle

De nombreux modèles qualitatifs ou quantitatifs ont été proposés pour rendre compte de l'intégration des informations visuelles dans la perception de la parole. La plupart sont des extensions audiovisuelles de modèles existant en perception auditive de la parole. Les deux principales questions auxquelles tentent de répondre ces modèles sont :

1. À quel niveau de traitement a lieu l'intégration des informations auditives et visuelles ?
2. Quelle est la nature des informations au moment de leur intégration ? La question subsidiaire étant : les informations d'une modalité sont-elles converties dans une métrique propre à l'autre modalité, ou existe-t-il une métrique commune qui permette l'intégration audiovisuelle ?

3.4.1 Modèles post-catégoriels

La première question s'est souvent ramenée au problème de savoir si l'intégration était pré-catégorielle ou post-catégorielle. Dans le cas post-catégoriel, la nature des représentations au moment de la convergence est commune aux informations fournies par les modalités auditive et visuelle puisqu'il s'agit d'un code linguistique (phonétique, phonologique ou lexical).

Dans l'un des tous premiers modèles, l'hypothèse VPAM proposée par MacDonald et McGurk (1978), l'intégration a lieu après la catégorisation en un code phonétique puisque cette hypothèse suppose que les indices visuels spécifient un lieu d'articulation et que les indices acoustiques spécifient les autres traits phonétiques. La convergence de ces catégories

phonétiques, établies indépendamment pour la vision et l'audition permet alors l'identification du phonème. Ce modèle n'a jamais été soutenu par aucune donnée. Un modèle d'intégration post-phonologique a été évalué par (Braida, 1991) : dans ce *post-labeling model*, une catégorisation phonologique a lieu dans chaque modalité : un phonème est spécifié par les informations auditives et un autre par des informations visuelles. Chaque combinaison d'un phonème auditif et d'un phonème visuel est associée à un phonème perçu donné avec une certaine probabilité. Ce modèle sous-estime les performances en perception audiovisuelle.

Un autre modèle, souvent considéré comme post-catégoriel (Schwartz et coll., 1998), est le *Fuzzy Logical Model of Perception* (FLMP : Massaro & Cohen, 1983 ; Massaro, 1987). Ce modèle comprend 2 niveaux de prototypes linguistiques (ou représentations en mémoire à long terme). Le premier niveau est unimodal : les parties auditives et visuelles d'un stimulus bimodal supportent à divers degrés différents prototypes unimodaux appelés traits perceptifs. L'évaluation de ce soutien se fait sur une échelle de valeurs de vérité continue, d'où le nom de logique floue, et a lieu de manière indépendante dans chaque modalité sensorielle. Le second niveau de prototype est bimodal et correspond au niveau des phonèmes : le prototype d'un phonème consiste en une combinaison de traits perceptifs auditifs et visuels. L'intégration audiovisuelle est une étape de classification qui consiste à calculer la probabilité de chaque phonème en fonction des valeurs de vérité attribuées à chaque trait perceptif durant l'étape d'évaluation unimodale. L'étape d'évaluation unimodale peut être considérée comme catégorielle puisqu'il s'agit de comparer des informations continues à des prototypes. Dans ce sens, il s'agit donc bien d'un modèle post-catégoriel. Cependant l'évaluation unimodale se fait de manière continue et non exclusive et l'intégration audiovisuelle a donc lieu sur des représentations qui ne sont pas totalement catégorisées. Les auteurs du modèle eux-mêmes contestent que la catégorisation phonétique, au sens d'une classification en deux entités mutuellement exclusives, soit un mécanisme fondamental de la perception (Massaro & Cohen, 1983). Le FLMP a été testé par ses auteurs sur un grand nombre de données expérimentales, notamment dans des paradigmes d'effet McGurk. L'adéquation entre le modèle et les données est généralement excellente mais le test consiste uniquement à trouver des paramètres qui permettent l'adéquation du modèle aux données unimodales et bimodales et non à prédire les performances bimodales à partir des données unimodales. Cette démarche a été contestée sur le principe (Vroomen & de Gelder, 2000, voir cependant Braida, 1991 pour une application prédictive du FLMP). D'autres auteurs, sans en contester le principe, mettent en doute la validité mathématique du calcul d'adéquation du FLMP avec les données de type McGurk (Schwartz, 2003).

Un dernier type de modèle post-catégoriel, récemment proposé par (L. E. Bernstein, Auer & Moore, 2004) repousse l'intégration à un niveau post-perceptif. Dans ce modèle « modalité-spécifique », un décodage complet de la parole est réalisé dans chaque modalité sensorielle, sans convergence des informations auditives et visuelles. Tout effet d'interaction entre informations auditives et visuelles relèverait nécessairement d'un niveau décisionnel ou associatif.

3.4.2 Modèles pré-catégoriels

Mis à part ce cas extrême, il existe un consensus apparent sur la vraisemblance d'une convergence pré-catégorielle des informations auditives et visuelles de parole, c'est-à-dire avant tout accès à un code linguistique. Si ce n'est pas le code linguistique qui permet la combinaison des informations auditives et visuelles, sous quelle forme ces informations convergent-elles ? Summerfield (1987) propose que les informations visuelles pourraient être converties sous une forme propre à la perception auditive. Un argument pour ce type de métrique est que l'expérience d'une amélioration de l'intelligibilité ou d'une illusion audiovisuelle est apparemment vécue dans la modalité auditive. Une métrique auditive pré-phonétique possible est l'estimation de la fonction filtre de l'appareil phonatoire qui peut être réalisée indépendamment sur la base des indices auditifs et visuels (Summerfield, 1987, 2ème métrique).

Une autre possibilité est qu'il existe une représentation pré-phonétique qui ne soit propre ni à la modalité auditive ni à la modalité visuelle. Cette métrique commune pourrait être la représentation des gestes articulatoires du locuteur soit par le biais de représentations motrices (théorie motrice de la perception de la parole : Liberman & Mattingly, 1985), soit par le biais de représentations des évènements « distaux » (c'est-à-dire hors du sujet) qui ont produit les stimulations auditives et visuelles (théorie directe-réaliste : Fowler & Rosenblum, 1991, voir aussi la partie 2.2.3 page 30). Dans les deux cas, les objets de la perception de la parole ne sont plus les variations du signal acoustique, mais le geste articulatoire intentionnel qui peut être retrouvé aussi bien à partir des indices auditifs que des indices visuels.

Un dernier type de modèle propose de supprimer l'étape de segmentation phonétique (Summerfield, 1987, 3ème métrique). Comme cette étape n'existe plus, ces modèles ne peuvent pas véritablement être qualifiés de pré-catégoriels, au sens phonologique. Ces modèles sont des extensions audiovisuelles de modèles qui postulent un codage direct du spectre auditif en représentations lexicales, sans niveau de représentation intermédiaire. L'intégration audiovisuelle dans ce type de modèles consiste essentiellement à juxtaposer des indices visuels (par exemples des paramètres d'ouverture de la bouche) aux informations spectrales auditives. Cet ensemble de paramètres auditifs et visuels est alors comparé à des prototypes lexicaux. Notons que Braida (1991) propose un modèle de ce type (le *pre-labeling model*) pour rendre compte de l'identification des consonnes audiovisuelles dans le bruit, mais, dans son cas, les prototypes sont des phonèmes et non des mots. Son modèle est donc plutôt à rapprocher d'un modèle d'intégration pré-phonologique.

Dans tous les modèles cités, l'intégration des informations se fait à une étape unique du traitement des stimuli. Il n'y a pas de raison a priori de limiter le nombre d'étapes auxquelles les indices auditifs et visuels peuvent converger, excepté le principe de parcimonie, et il semble qu'une étape unique d'intégration ne puisse rendre compte de la variété des effets des indices visuels sur la perception de la parole.

3.5 Conclusion

Les deux principaux effets intersensoriels dans la perception de la parole, l'amélioration de l'intelligibilité dans le bruit et l'effet McGurk montrent sans ambigüité l'existence d'interactions entre traitement des informations auditives et visuelles, au moins sous la forme d'une influence des informations visuelles sur le traitement auditif. Les études sur la perception de la parole bimodale ont montré que cette intégration pouvait concerner non seulement des informations complémentaires à propos du même évènement linguistique, mais aussi des informations redondantes, sous la forme d'une corrélation temporelle entre les signaux acoustiques et visuels. C'est principalement l'intégration des informations complémentaires qui a été étudiée et a donné naissance à des modèles qui pour beaucoup d'entre eux situent le stade d'intégration à un niveau pré-phonétique. Des effets d'intégration audiovisuelle à des niveaux de traitement linguistiques plus élevés suggèrent cependant que l'intégration n'a pas lieu une fois pour toutes à un niveau pré-phonologique et qu'il existe soit des effets descendants influençant l'intégration audiovisuelle ou soit des apports d'informations visuelles à plusieurs niveaux du traitement linguistique.

Comment situer ces différents niveaux d'intégration dans une architecture générale des systèmes sensoriels ? Si des échanges d'informations auditives et visuelles ont lieu avant la catégorisation en phonèmes, ont-ils lieu pour autant selon les mêmes mécanismes que ceux qui sont à l'œuvre dans d'autres cas d'intégration audiovisuelle ? La réponse dépend du modèle de perception de la parole dans lequel on se place et comment celui-ci considère les traitements de la parole par rapport aux autres traitements auditifs.

Si l'on se place dans le cadre de la théorie motrice de la parole, la perception de la parole est réalisée par des structures corticales dédiées, différentes des structures auditives traitant les autres types de stimuli auditifs, et ce à un niveau de traitement assez précoce. L'intégration audiovisuelle des informations de parole ne signifie donc pas qu'il y ait des échanges d'informations entre les systèmes sensoriels auditifs et visuels qui ne sont pas impliqués dans l'analyse de la parole. À l'appui de cette théorie, Tuomainen, Andersen, Tiippana et Sams (2005) ont montré que l'amplitude de l'effet McGurk pour des syllabes dont les formants avaient été remplacés par des sons purs de même fréquence (*sinewave speech*) était supérieur lorsque ces sons étaient perçus comme de la parole que lorsqu'ils ne l'étaient pas.

Pour les tenants du FLMP ou de la théorie directe réaliste, à l'inverse, les mécanismes d'intégration à l'œuvre dans les effets audiovisuels ne sont pas propres au traitement de la parole : Saldaña et Rosenblum (1993) ont ainsi mis en évidence une illusion analogue à l'effet McGurk hors du domaine de la parole : le fait de voir le frottement ou le pincement d'un corde de violoncelle influence la catégorisation d'un continuum acoustique entre les deux sons produits par l'une ou l'autre de ces actions. Dans ce cas, les résultats concernant l'intégration audiovisuelle dans la perception de la parole pourraient être généralisés à la perception d'un évènement audiovisuel en général.

Chapitre 4

Intégration audiovisuelle en neurosciences cognitives

De nombreuses études, essentiellement à partir de la fin des années 90 avec l'avènement des nouvelles techniques de neuroimagerie non invasives chez l'homme, ont tenté de faire le lien entre les résultats de la neurophysiologie et ceux de la psychologie expérimentale ou cognitive. Les techniques d'imagerie cérébrale plus (ou moins) récentes ont permis d'étudier plus directement les mécanismes cérébraux, ou du moins les aires cérébrales, impliqués dans l'intégration des informations auditives et visuelles chez l'homme. Même si une partie de ces études a repris des paradigmes issus de la psychologie expérimentale, elles ont également permis d'étudier les mécanismes cérébraux impliqués dans le traitement d'un véritable évènement audiovisuel et un certain nombre de paradigmes originaux ont été développés. En effet, l'utilisation de techniques de neuroimagerie permet d'étudier les réponses à une combinaison d'informations auditives et visuelles congruentes de façon plus directe, sans avoir recourir à des artifices expérimentaux tels que des conflits intersensoriels ou la variation du délai entre les informations auditives et visuelles. L'identification des aires cérébrales impliquées dans cette intégration a nécessité également d'établir des critères d'intégration audiovisuelle, qui dépendent de la technique utilisée. Les problèmes méthodologiques relatifs à l'utilisation de certains de ces critères seront discutés plus en détail dans la partie 7.2 page 106 et seront simplement évoqués dans ce chapitre, le cas échéant.

4.1 Comportements d'orientation et colliculus supérieur

L'un des premiers ensembles d'études dans lequel émerge une volonté de faire un lien entre comportement et processus neurophysiologiques ne provient cependant pas des études chez l'homme mais de celles sur l'animal. Il s'agit de l'étude des comportements d'intégration multisensorielle liés au colliculus supérieur (voir aussi la partie 1.4.1 page 13).

4.1.1 Orientation vers un stimulus audiovisuel chez l'animal

Partant du constat que cette structure sous-tend des comportements d'orientation vers un stimulus (par exemple G. E. Schneider, 1969), les auteurs qui ont mis en évidence les règles d'intégration multisensorielle de certaines cellules nerveuses du colliculus supérieur (voir la partie 1.4.1 page 15) montrent que le comportement d'orientation vers un évènement audiovisuel suit les même règles d'intégration (Stein, Huneycutt & Meredith, 1988 ; Stein, Meredith, Huneycutt & McDade, 1989). Des chat sont entrainés à se diriger vers un stimulus visuel ou un stimulus auditif, qui peut être présenté à différentes excentricités. Pour des intensités liminaires, la performance des animaux dans l'orientation vers un stimulus bimodal est meilleure que celle prédite sur la base des performances unimodales, sous l'hypothèse d'une indépendance de traitement des stimuli unimodaux, ce qui n'est pas le cas pour des stimuli supraliminaires. Ce résultat semble imiter la règle d'efficacité inverse. En outre, le fait de présenter un stimulus auditif à une excentricité différente du stimulus visuel (dans ce cas, la tâche du chat est de se diriger vers le stimulus visuel tout en ignorant le stimulus auditif) diminue la performance, ce qui rappelle la règle de proximité spatiale, mais uniquement si le stimulus auditif est plus central que le stimulus visuel (et pas l'inverse).

D'autres arguments plus récents confortent l'hypothèse d'une implication des neurones bimodaux du colliculus supérieur dans l'amélioration multisensorielle du comportement d'orientation spatiale : le colliculus supérieur (chez le chat) reçoit des entrées de plusieurs aires corticales telles que le sillon ectosylvien antérieur (AES)[1] et le sillon suprasylvien rostral (rLS). Or d'une part la "déactivation" transitoire des aires corticales AES et rLS chez le chat anesthésié supprime le caractère multiplicatif des réponses des cellules bimodales du colliculus supérieur, sans toutefois supprimer leurs réponses aux stimuli bimodaux (Jiang & Stein, 2003 ; Jiang, Wallace, Jiang, Vaughan & Stein, 2001). D'autre part, la déactivation transitoire de ces mêmes aires compromet la facilitation multisensorielle de l'orientation vers un stimulus bimodal tout en préservant les performances dans l'orientation vers un stimulus unimodal auditif ou visuel (Jiang, Jiang & Stein, 2002 ; Wallace & Stein, 1994 ; Wilkinson, Meredith & Stein, 1996). C'est donc précisément le caractère multiplicatif des cellules du colliculus supérieur qui semble fondamental pour l'exploitation du caractère bimodal des stimuli, caractère qui semble être conféré au colliculus supérieur par ces deux aires corticales (notons que les aires corticales ne semblent pas suffire puisqu'un lésion excito-toxique du colliculus supérieur affecte également spécifiquement la facilitation d'un chat à s'orienter vers un stimulus bimodal : Burnett, Stein, Chaponis & Wallace, 2004). Bien qu'indirects, ces résultats suggèrent l'existence d'un lien entre l'augmentation du taux de décharges observé dans certaines cellules bimodales du colliculus supérieur et l'amélioration comportementale de l'orientation vers un stimulus audiovisuel.

[1]Mais bien que l'aire AES du chat soit aussi une aire montrant une certain proportion de cellules bimodales au comportement intégratif analogue à celui des cellules multimodales du colliculus supérieur (Wallace, Meredith & Stein, 1992 ; Benedek, Fischer-Szatmari, Kovacs, Pereny & Katoh, 1996 ; Benedek, Eordegh, Chadaide & Nagy, 2004), les systèmes multimodaux du colliculus supérieur et de l'aire AES semblent constituer deux systèmes indépendants car les cellules de l'aire AES qui projettent vers le SC sont uniquement les cellules unimodales (Wallace, Meredith & Stein, 1993)

4.1.2 Saccades oculaires vers un stimulus audiovisuel, chez l'homme

Un aspect particulièrement étudié du comportement d'orientation spatiale est la réalisation de saccades oculaires, qui est en partie sous la dépendance du colliculus supérieur, dont certains neurones présentent des décharges synchronisées aux saccades (voir par exemple Peck, 1987). Chez l'homme, plusieurs études ont montré que la présentation concomitante de stimuli auditifs et visuels influence les saccades oculaires, par rapport à des saccades vers un stimulus unimodal. Contrairement aux comportements d'orientation chez le chat, la performance n'est pas affectée par la bimodalité mais l'exécution des saccades oculaires vers un stimulus visuel (Frens, Van Opstal & Willigen, 1995) ou vers un stimulus auditif (Lueck, Crawford, Savage & Kennard, 1990) est plus rapide en présence d'un stimulus accessoire dans l'autre modalité. Cette différence pourrait être attribuée au fait que les études chez l'homme ont pour la plupart utilisé des stimuli supraliminaires. Cette diminution de latence s'observe également dans un paradigme d'attention partagée (voir la partie 2.3.3 page 36) dans lequel le sujet doit effectuer une saccade indifféremment vers un stimulus auditif ou visuel (Arndt & Colonius, 2003 ; Harrington & Peck, 1998 ; Hughes, Nelson & Aronchick, 1998 ; Hughes et coll., 1994). Dans ce cas, l'application de l'inégalité de Miller permet de rejeter une explication en termes de facilitation statistique et a été interprétée comme la preuve de l'existence d'une convergence des traitements auditif et visuel, éventuellement au niveau du colliculus supérieur. Une comparaison directe entre l'amplitude de la violation de l'inégalité de Miller dans un paradigme de saccade et un paradigme de TR manuel (RSE ou RTE) suggère que les mécanismes neuronaux qui sous-tendent ces deux tâches sont très différents (Hughes et coll., 1994). Concernant les aspects dynamiques de la saccade, cette dernière est essentiellement contrôlée par le stimulus visuel (Frens et coll., 1995), et l'influence d'un stimulus auditif sur la trajectoire ou la vitesse sont assez faibles (Hughes et coll., 1998).

En revanche, les effets de la proximité temporelle et spatiale ainsi que ceux de l'intensité des stimuli ne sont pas directement prédictibles à partir des règles d'intégration décrites au niveau neuronal dans le colliculus supérieur du chat anesthésié par Stein et Meredith (1993). D'abord, si le gain bimodal saccadique diminue effectivement avec la séparation spatiale des deux stimuli (Arndt & Colonius, 2003 ; Frens et coll., 1995), une violation de l'inégalité de Miller peut exister pour des séparations allant jusqu'à 30° d'angle visuel (Harrington & Peck, 1998). De plus, la facilitation maximale n'est pas obtenue pour des stimuli auditifs et visuels strictement alignés lorsqu'ils sont périphériques (Hughes et coll., 1998). Ensuite, les effets de la séparation temporelle sont plus variables et dépendent de la tâche (attention partagée, modalité accessoire auditive ou visuelle : Frens et coll., 1995 ; Hughes et coll., 1998 ; Kirchner & Colonius, 2005).

Enfin, l'intensité des stimuli, soit n'a pas d'effet sur l'amplitude de la facilitation (Frens et coll., 1995 ; Hughes et coll., 1994), soit a un effet qui peut être totalement expliqué par un modèle d'activations séparées (Arndt & Colonius, 2003). Ces résultats semblent s'opposer au principe de l'efficacité inverse qui s'applique au niveau neuronal dans le colliculus supérieur et à la performance comportementale des chats dans des tâches d'orientation. Cette disparité pourrait s'expliquer par l'absence d'effets de seuil sur les TR lorsque l'intensité des stimuli diminue, contrairement à ce qui est le cas pour les performances et pour

le taux de décharges des neurones.

Une comparaison stricte de ces résultats avec les réponses bimodales multiplicatives des neurones du colliculus supérieur est hasardeuse étant donné la différence entre les paradigmes expérimentaux, les stimuli et les mesures utilisés. En revanche, un certain nombre d'études ont tenté d'établir un lien entre interactions neuronales audiovisuelles et facilitation comportementale, en enregistrant les réponses unitaires de neurones du colliculus supérieur chez l'animal alerte et conditionné à effectuer une saccade vers un stimulus auditif ou visuel.

4.1.3 Expériences chez l'animal alerte et actif

La première étude à s'être intéressée spécifiquement à cette question est sans doute celle de Peck (1987), chez le chat, qui montre une augmentation de l'activité pré-saccadique de certains neurones du colliculus supérieur lorsque les saccades sont évoquées par des stimuli bimodaux plutôt qu'unimodaux. Des études plus récentes chez le macaque ont montré que la diminution de la latence des saccades vers un stimulus audiovisuel était plutôt corrélée à une augmentation de la réponse prémotrice de ces neurones, qui précède de peu la saccade, qu'à des interactions au niveau de leur réponse sensorielle au stimulus vers lequel la saccade doit être faite (A. H. Bell, Meredith, Van Opstal & Munoz, 2005 ; Frens & Van Opstal, 1998).

Bien que des effets multiplicatifs similaires à ceux montrés sur la réponse sensorielle de neurones bimodaux d'animaux anesthésiés aient été montrés chez l'animal alerte, mais passif (A. H. Bell, Corneil, Meredith & Munoz, 2001 ; Wallace et coll., 1998), il semble que, lorsque l'animal est actif, ces effets soient plus rares et que l'on observe plus souvent des diminutions du taux de décharge en réponse aux stimuli bimodaux (Frens & Van Opstal, 1998 ; Populin & Yin, 2002). Si une partie de ces différences peut être attribuée à l'utilisation d'indices différents pour le calcul des interactions multisensorielles, ou à l'utilisation de stimuli supraliminaires plutôt que liminaires (voir Perrault, Vaughan, Stein & Wallace, 2003, 2005 ; Stanford, Quessy & Stein, 2005), l'anesthésie pourrait avoir des effets non négligeables sur le comportement intégratif des neurones bimodaux du colliculus supérieur (voir la partie 1.1.4 page 9), si bien qu'on peut s'interroger sur le rôle des interactions multiplicatives dans le comportement puisqu'elles ont essentiellement été trouvées chez des animaux anesthésiés ou passifs (voir cependant Cooper, Miya & Mizumori, 1998). Paradoxalement, la proportion de neurones bimodaux chez des singes ayant une tâche de saccades oculaires à réaliser semble beaucoup plus importante que chez le singe anesthésié (Frens & Van Opstal, 1998).

Bien que le colliculus supérieur soit sans nul doute impliqué dans des comportements d'orientation, en particulier les saccades oculaires, il reste à prouver que les réponses multiplicatives de certaines neurones du colliculus supérieur sont directement liés aux gains observés au niveau comportemental. Il n'en reste pas moins vrai qu'une intégration des informations spatiales auditives et visuelles a sans doute lieu dans cette structure, sans doute par des mécanismes neuronaux complexes, en interaction avec d'autres structures corticales telles que l'aire AES, le sillon suprasylvien ou encore le champ oculaire frontal (Meredith, 1999). Chez l'homme, il semble en revanche qu'il n'y ait pas eu d'études

avec des techniques de neuroimagerie des corrélats neurophysiologiques de la facilitation du comportement d'orientation par un stimulus bimodal.

4.2 Effet du stimulus redondant

Les bases neurophysiologiques de l'effet de redondance du stimulus sur le TR manuel (voir la partie 2.3 page 31) ont été beaucoup plus étudiées chez l'homme. Quelques études relativement anciennes ont mesuré les potentiels évoqués dans des tâches de détection d'un stimulus bimodal, soit dans le paradigme du stimulus accessoire (L. K. Morrell, 1968b), soit dans un paradigme d'attention partagée (Andreassi & Greco, 1975 ; Squires, Donchin, Squires & Grossberg, 1977).

4.2.1 Premières études

Ainsi L. K. Morrell (1968b), en comparant les potentiels évoqués par une cible audiovisuelle à la somme des potentiels évoqués par une cible visuelle et par un stimulus auditif accessoire non-cible (moyennés sur une fenêtre temporelle entre 140 et 256 ms post-stimulus), montre un effet compatible avec une activation ou une modulation d'activité des aires motrices, qui de plus est corrélé au gain de TR pour traiter un cible audiovisuelle par rapport à une cible visuelle. Andreassi et Greco (1975), puis Squires et coll. (1977) montrent que les latences des composantes N2 et P3 enregistrées au vertex se comportent comme le TR : leurs latences en condition bimodale sont inférieures ou égales à la plus courte des latences en conditions unimodales, ce qui suggère que l'intégration des stimuli auditifs et visuels a lieu avant les stades de traitement correspondant à ces deux ondes. Le problème pour ces deux études est qu'elles ne prennent pas en compte la superposition spatiale des champs de potentiel électrique (voir la partie 6.2.2 page 89) : la réponse évoquée n'est enregistrée qu'à une (ou quelques) électrodes sur le scalp et les réponses des trois conditions de stimulation sont comparées directement sans tenir compte du fait que la réponse bimodale peut être composée de différentes activités modalité-spécifiques superposées, ce qui rend le potentiel électrique au vertex ininterprétable. A posteriori, cette approche peut se justifier pour l'onde P3, qui n'est pas spécifique à une modalité sensorielle et dont la latence est relativement tardive et l'amplitude suffisamment grande pour être préservée des effets de diffusion d'éventuelles activités modalité-spécifiques concomitantes. Mais l'interprétation reste plus spéculative concernant l'onde N2, dont au moins une partie des générateurs est modalité-spécifique.

4.2.2 Tâches de discrimination

Après ces premières expériences, je n'ai trouvée aucune étude d'imagerie cérébrale de l'effet de facilitation audiovisuelle du TR avant la fin des années 90. Toutes les études récentes ont été réalisées en potentiels évoqués, enregistrés sur l'ensemble du scalp, dans des paradigmes d'attention partagée permettant de mettre en évidence un effet du stimulus redondant dans une tâche de détection simple ou dans une tâche de discrimination de deux stimuli. Dans toutes ces études la réponse au stimulus bimodal était comparée à

la somme des réponses à leurs composantes unimodales (modèle additif, voir partie 7.2.1 page 107). Dans l'étude de Giard et Peronnet (1999), les sujets devaient discriminer deux objets, définis chacun soit uniquement par un trait dynamique visuel (déformation d'un cercle dans la direction horizontale ou verticale), soit uniquement par un trait auditif (son pur grave ou aigu), soit par la combinaison congruente et simultanée de leurs traits auditifs et visuels. Le TR en condition bimodale était inférieur aux TR auditif ou visuel, comme on l'attendait (bien que l'inégalité de Miller n'ait pas été testée). L'application du modèle additif a montré l'existence d'activités occipitales très précoces (entre 40 et 140 ms) qui ne s'expliquent ni par la réponse unimodale au stimulus visuel seul, ni a fortiori par la réponse au stimulus auditif seul. D'autres activités ou modulations d'activité propres à la stimulation audiovisuelle ont été trouvées dans cette expérience entre 100 et 200 ms, dans les aires sensorielles unimodales, ainsi que dans les régions fronto-temporales. Dans une variante de ce paradigme expérimental, Fort, Delpuech, Pernier et Giard (2002b) ont montré que les interactions audiovisuelles étaient partiellement différentes lorsque le traitement des informations auditives et des informations visuelles étaient tous deux nécessaires pour discriminer les cibles audiovisuelles (c'est-à-dire lorsque les traits auditifs et visuels définissant un objet audiovisuel n'étaient pas redondants). On n'observait notamment pas d'activités occipitales précoces dans ce cas.

De façon intéressante, dans les deux études précédentes, les interactions audiovisuelles dans les cortex sensoriels spécifiques étaient différents selon la modalité dominante du sujet pour la tâche (identifiée par la modalité dans laquelle le TR unimodale était le plus court) : l'amplitude des interactions était plus grande dans le cortex de la modalité non-dominante.

L'existence d'interactions audiovisuelles précoces dans le cortex occipital a fait l'objet de controverses : dans un paradigme consistant pour le sujet à détecter des cibles auditives, visuelles et audiovisuelles rares (15% des essais) différant des stimuli standards sur leur intensité (paradigme différent du précédent mais impliquant lui aussi la discrimination de stimuli unimodaux et bimodaux), Teder-Sälejärvi, McDonald, Di Russo et Hillyard (2002) trouvent effectivement des interactions occipitales précoces entre 40 et 100 ms, mais les attribuent à des effets pervers de l'application du modèle additif, due à des activités anticipatoires communes aux trois conditions de présentation (voir partie 7.2.1 page 108). Dans cette expérience, la diminution du TR pour la détection des cibles audiovisuelles est associée à des interactions débutant vers 130 ms dans le cortex occipital, et suivies par des interactions d'origine vraisemblablement supra-temporales entre 170 et 250 ms. La différence de paradigme expérimental et de stimuli rend cependant difficile la comparaison des résultats.

4.2.3 Tâche de détection

Un autre ensemble de résultats concerne les interactions audiovisuelles observées dans des paradigmes de simple détection de stimuli auditifs, visuels et audiovisuels. En utilisant exactement les mêmes stimuli que Giard et Peronnet (1999), mais en demandant aux sujets de répondre le plus rapidement possible quelle que soit l'identité de l'objet présenté, Fort, Delpuech, Pernier et Giard (2002a) observent des interactions partiellement différentes,

ce qui montre que les mécanismes d'intégration multisensorielle peuvent être influencées par la tâche réalisée, et que ces interactions ne reflètent pas simplement la rencontre des informations auditives et visuelles selon un schéma rigide de convergence. Les résultats montrent les mêmes interactions occipitales précoces que celles de Giard et Peronnet (1999). De plus, elles résistent aux contrôles proposés par Teder-Sälejärvi et coll. (2002) pour éliminer les effets pervers de l'application du modèle additif. Ces interactions précoces sont suivies d'interactions vers 100 ms, compatibles avec l'activation du colliculus supérieure et d'interactions fronto-temporales vers 170 ms, analogues à celles trouvées par Giard et Peronnet (1999) à la même latence.

Molholm et coll. (2002), dans un paradigme similaire, trouve des interactions audio-visuelles fort ressemblantes ainsi qu'une modulation de l'onde N1 visuelle, curieusement observée par Giard et Peronnet (1999) dans leur paradigme de discrimination, mais pas par Fort et coll. (2002a) dans leur paradigme de détection simple. Dans cette étude, la diminution du TR bimodal est inférieure à celle prédite par un modèle d'activations séparées (sous l'hypothèse d'indépendance des distributions unimodales des TR, voir la partie 7.1.1). Notons qu'une étude de potentiels évoqués intracérébraux récente chez 3 patients épileptiques, utilisant le même protocole, montre des interactions au niveau du cortex pariétal à partir de 120 ms de traitement (Molholm et coll., 2006).

Les interactions audiovisuelles identifiées grâce aux modèle additif appliqué aux potentiels évoqués semblent donc varier aussi bien en fonction de la tâche, du paradigme, des stimuli utilisés et des sujets. Malgré cette variabilité, certaines ont été reproduites par plusieurs équipes : une activité occipitale précoce observée à partir de 40 ms de traitement (Fort et coll., 2002b ; Giard & Peronnet, 1999 ; Molholm et coll., 2002), une modulation de l'amplitude de l'onde N1 visuelle dans la condition audiovisuelle par rapport à la condition visuelle seule autour de 170 ms (Fort et coll., 2002b ; Giard & Peronnet, 1999 ; Teder-Sälejärvi et coll., 2002), une activité fronto-temporale autour de 170 ms de traitement (Fort et coll., 2002a ; Giard & Peronnet, 1999 ; Molholm et coll., 2002). Toutes ces interactions semblent avoir lieu avant les activités motrices liées à la réponse. Une étude des réponses unitaires de neurones du cortex moteur chez le macaque, dans une tâche de détection simple (J. O. Miller, Ulrich & Lamarre, 2001) a montré que la latence de décharge de ces neurones était diminuée en condition bimodale de façon parallèle à la diminution bimodale de TR, le délai entre ces latences et le TR de détection étant constant quelle que soit la condition. Tous ces résultats sont compatible avec un modèle de coactivation audiovisuelle ayant lieu avant l'étape motrice et pouvant prendre place au niveau des cortex sensoriels spécifiques dès les première étapes de traitement cortical. Elles suggèrent en revanche que les stades de coactivation sont multiples et modulées par le contexte expérimental.

4.3 Interactions audiovisuelles dans la perception des émotions

L'utilisation de techniques d'exploration de l'activité cérébrale chez l'homme a également coïncidé avec l'utilisation de stimuli plus écologiques et donc plus complexes que

ceux utilisés dans les paradigmes d'attention partagée, tels que des stimuli émotionnels, des objets existants (voir la partie 4.4) ou la parole (traité en partie 4.7 page 74))

La perception des émotions peut donner lieu à une influence réciproque des indices auditifs et visuels et à un certain nombre de phénomènes typiques des interactions intermodales. Un protocole expérimental souvent utilisé consiste à présenter des mots ou des phrases dont l'intonation exprime l'une des émotions primaires (joie, peur, colère...), associés à des visage portant des expressions émotionnelles congruentes ou incongruentes avec ces intonations. Plusieurs études ont ainsi mis en évidence un biais perceptif audiovisuel dans la catégorisation émotionnelle des voix ou des visages (de Gelder & Vroomen, 2000 ; de Gelder, Vroomen & Bertelson, 1998 ; Massaro & Egan, 1996 ; Vroomen, Driver & de Gelder, 2001). D'autres ont montré une amélioration des performances (Hietanen, Leppänen & Illi, 2004) ou une diminution du TR (Dolan, Morris & de Gelder, 2001) pour des visages et des voix congruents, par rapport à une condition incongruente, dans des tâches de reconnaissance auditive ou visuelle d'émotions.

Pourtois, de Gelder, Vroomen, Rossion et Crommelinck (2000) ont étudié les activités cérébrales potentiellement associées à ces effets intersensoriels : dans leur expérience, un visage et un voix émotionnellement congruente ou incongruente étaient présentés à des délais variable de façons à pouvoir calculer indépendamment les potentiels évoqués par la voix et le visage. Le traitement de la voix était modulé par la congruence émotionnelle du visage à un niveau relativement précoce du traitement auditif (onde N1 auditive, vers 100 ms) mais uniquement si le visage était présenté à l'endroit. Dans un protocole légèrement différent, Pourtois, Debatisse, Despland et de Gelder (2002) montrent que la congruence des émotions exprimées par une voix et un visage module l'amplitude d'une onde pariétale plus tardive (vers 220 ms), qui pourrait refléter une activité dans le cortex cingulaire antérieur ; mais selon les auteurs, cet effet serait plutôt liée à la détection de l'incongruence qu'à des interactions spécifiques au traitement des émotions.

Dolan et coll. (2001) ont montré dans un protocole d'imagerie par résonance magnétique fonctionnel (IRMf) évènementiel que des activités dans l'amygdale gauche et le gyrus fusiforme, spécifiques du traitement de la peur exprimée par un visage étaient modulées par la présentation d'une voix exprimant la peur comparativement à une voix exprimant la joie. Cette interaction était accompagnée d'une diminution du TR pour catégoriser les émotions faciales, et semble spécifique au traitement de la peur car aucune modulation n'a été observée dans ces structures dans le cas de la joie. La technique utilisée ne permet évidemment pas d'avoir une idée de la latence de ces effets.

4.4 Objets écologiques audiovisuels

Récemment, diverses expériences de neuroimagerie ont utilisé un autre type de stimuli écologiques comme des images ou des photos d'objets fabriqués (par exemple des outils) ou naturels (par exemple des animaux), associées aux sons qu'ils produisent. Les activités cérébrales propres aux interactions audiovisuelles dans la perception de tels stimuli ont essentiellement été étudiées en IRMf, avec des résultats qui, ici encore, sont très variables et dépendent sans doute tout à la fois des protocoles utilisés, des tâches demandées aux sujets

et des analyses effectuées. Dans un protocole d'IRMf par blocs, comparant les réponses à des stimuli audiovisuels congruents et incongruents durant une tâche portant sur la modalité visuelle, Laurienti et coll. (2003) montrent une implication des cortex cingulaire antérieur et préfrontal médian, associée — mais non corrélée — à une diminution du TR pour traiter les stimuli visuels lorsque ceux-ci sont congruents avec les stimuli auditifs. Dans une expérience, dans laquelle les sujets sont passivement exposés à des objets audiovisuels congruents et incongruents, Olivetti Belardinelli et coll. (2004) montrent que les gyrus para-hippocampique et lingual sont plus activés par les stimuli congruents que par des stimuli incongruents.

Dans une série d'expériences d'IRMf, Beauchamp, Lee, Argall et Martin (2004) montrent qu'une aire bordant le sillon temporal supérieur et débordant sur le gyrus temporal médian (STS/GTM), et le cortex temporal ventral pourraient constituer des aires de convergence des informations auditives et visuelles relatives aux objets : elles sont plus activées par des objets auditifs ou visuels que par des stimuli ne correspondant à aucun objet, et par des stimuli bimodaux que par des stimuli unimodaux ; un protocole évènementiel permet de montrer qu'elles sont plus activées par l'analyse sensorielle que par la réponse ; enfin, elles sont plus activées par des stimuli audiovisuels congruents que par des stimuli incongruents. Notons que le cortex temporal ventral montre une préférence pour les stimuli visuels, contrairement au STS/GTM qui est autant activé par les objets auditifs que visuels. Beauchamp, Lee et coll. (2004) ont également utilisé comme stimuli des vidéos d'actions impliquant des objets, associées aux bruits de ces actions, ce qui ne semble pas modifier l'implication de ces deux aires corticales. Cela suggère que les activations observées sont plutôt de l'ordre d'un accès sémantique aux représentations des objets audiovisuels. Une expérience complémentaire, utilisant ces mêmes vidéos d'actions audiovisuelles (Beauchamp, Argall, Bodurka, Duyn & Martin, 2004), a permis de préciser l'organisation corticale de cette zone du STS/GTM : elle semble être constituée d'un ensemble de sous-aires sensibles soit à la composante auditive, soit à la composante visuelle du stimulus, soit aux deux.

4.5 Conditions limites de l'intégration audiovisuelle

Une autre façon de mettre en évidence des structures cérébrales participant à l'intégration audiovisuelle est de rechercher les structures qui présentent une activité plus importante lorsque les conditions d'une intégration sont réunies que lorsque certaines conditions limites sont dépassées.

Plusieurs effets d'intégration audiovisuelle comportementaux chez l'homme (l'effet McGurk, la ventriloquie) ou électrophysiologiques chez l'animal (la réponse multiplicative des neurones du colliculus supérieur) sont ainsi sérieusement compromis lorsque la coïncidence spatiale ou temporelle des stimuli n'est plus respectée (voir les chapitres 1, 2 et 3). D'où l'idée que certaines interactions multisensorielles n'ont lieu que dans la limite de ces conditions spatiales et temporelles. Deux études ont ainsi comparé une condition dans laquelle les stimuli auditifs et visuels sont synchrones à une condition dans laquelle ils sont décalés temporellement, en utilisant, soit des stimuli simples (bruits et inversion de damiers : Calvert, Hansen, Iversen & Brammer, 2001), soit des stimuli de parole (Calvert, Campbell & Brammer, 2000, voir partie 4.7 page 74). Ces études ont de plus postulé, par référence

directe au comportement multiplicatif des neurones du colliculus supérieur, que les aires aires d'intégration devaient être activées par ces stimuli synchrones au delà de la somme de leurs activations en conditions unimodales seules (super-additivité) et par des stimuli asynchrones en-deçà de cette somme (sous-additivité). Concernant l'étude sur les stimuli simples (Calvert et coll., 2001), un grand nombre de structures respectaient ces deux critères, dont notamment le colliculus supérieur, l'insula et le STS.

Une autre étude d'imagerie fonctionnelle en tomographie par émission de positons (TEP ; Bushara, Grafman & Hallett, 2001) a comparé la réponse hémodynamique dans des blocs de stimuli synchrones et des blocs de stimuli synchrones et asynchrones mélangés, dans lesquels le sujet devait détecter l'asynchronie audiovisuelle. Les aires plus activées dans le bloc asynchrone comprenaient notamment l'insula, dont l'activation était d'autant plus forte que la tâche de détection de l'asynchronie était difficile (avec ici un effet confondu de la tâche et de l'asynchronie puisque la tâche dans les blocs asynchrones était uniquement visuelle et non audiovisuelle). Donc contrairement aux deux études précédentes, l'implication de l'insula était vraisemblablement liée ici à la détection explicite de l'asynchronie et non au succès de l'intégration audiovisuelle.

À ma connaissance, aucune étude d'imagerie fonctionnelle hémodynamique ou électrophysiologique n'a utilisé la congruence spatiale comme critère pour étudier les interactions audiovisuelles chez l'homme, excepté dans le cas de la parole (Macaluso, George, Dolan, Spence & Driver, 2004, voir partie 4.7 page 74).

4.6 Corrélats neurophysiologiques des illusions audiovisuelles

Dans le même ordre d'idée, certaines études de neuroimagerie chez l'homme ont tiré parti des phénomènes d'illusion audiovisuelle pour étudier les structure impliquées dans l'intégration audiovisuelle. Au moins trois stratégies différentes ont été mises en œuvre.

4.6.1 Intégration audiovisuelle pré-attentive

La première stratégie s'appuie sur la mesure d'une onde des potentiels évoqués appelée négativité de discordance (*Mismatch Negativity, MMN*). La MMN (voir par exemple Näätänen, Tervaniemi, Sussman, Paavilainen & Winkler, 2001, ou la partie 12.1 page 175 pour une revue) est évoquée entre 100 et 300 ms de traitement par tout son déviant présenté dans une suite de sons standards identiques, et ce même si le sujet ne prête pas attention aux sons. La MMN est donc censée refléter des processus auditifs automatiques (on dit souvent pré-attentifs) de détection d'une déviance dans l'environnement sonore.

Dans les illusions McGurk et de ventriloquie, certaines caractéristiques auditives d'un son sont subjectivement modifiées par les informations visuelles qui l'accompagnent. Plusieurs études ont montré que cette modification perceptive suffisait à générer une MMN, dans une situation où la composante auditive du stimulus audiovisuel déviant était identique à celle du stimulus audiovisuel standard, et où seule la composante visuelle changeait entre standards et déviants. Cet effet a été montré pour l'effet McGurk en MEG (Möttönen, Krause, Tiippana & Sams, 2002 ; Sams et coll., 1991) et en EEG (Colin, Radeau,

Soquet & Deltenre, 2004 ; Colin, Radeau, Soquet, Demolin et coll., 2002). Concernant l'effet de ventriloquie, une première étude est parvenu à faire disparaitre la MMN qui aurait normalement dû être générée par un son déviant sur sa position spatiale, en présentant le stimulus visuel concomitant toujours à la même position (Colin, Radeau, Soquet, Dachy & Deltenre, 2002). Une étude plus récente a évoqué une MMN à des sons strictement identiques (de même provenance spatiale), mais dont la localisation auditive apparente était biaisée par un stimulus visuel déviant (Stekelenburg, Vroomen & de Gelder, 2004).

Ces résultats ne signifient cependant pas que l'intégration des informations auditives et visuelles a lieu au niveau de l'étape de traitement correspondant à la MMN, mais plutôt qu'à cette étape pré-attentive de traitement, les informations visuelles ont déjà automatiquement modifié le traitement auditif. Concernant ces deux illusions, la latence de la MMN représente donc un borne temporelle supérieure de l'intégration audiovisuelle. Il faut cependant prendre ces résultats avec prudence dans la mesure où le calcul de la MMN implique ici une soustraction entre deux conditions où les stimuli visuels sont différents. La différence observée pourrait donc refléter un traitement automatique de la déviance visuelle qui a récemment été mis en évidence (pour une revue, voir Pazo-Alvarez, Cadaveira & Amenedo, 2003, ou la partie 14.1 page 185) et non la MMN.

4.6.2 Application du modèle additif

Une seconde stratégie consiste à comparer les activités enregistrées lors d'une stimulation audiovisuelle donnant lieu à une illusion, à la somme des activités enregistrées séparément dans les conditions unimodales de stimulation (modèle additif), l'illusion servant uniquement à montrer qu'une intégration des informations auditives et visuelles a réellement eu lieu (comme dans le cas de la diminution du TR pour un stimulus redondant). C'est la stratégie suivie pour l'illusion "flash/bip". Il s'agit d'une illusion audiovisuelle mise en évidence relativement récemment, dans laquelle le nombre de flashs perçus est influencé par le nombre de stimuli sonores (bips) présentés au même moment (Shams, Kamitani & Shimojo, 2000 ; voir aussi Andersen, Tiippana & Sams, 2004). Dans sa version initiale, l'expérience consiste à présenter un flash unique accompagné de 1, 2 ou 3 bips et à demander au sujet le nombre de flash perçus.

Dans la version EEG, Shams, Kamitani, Thompson et Shimojo (2001), on présente aux sujets soit un flash, soit deux bips, soit les deux en même temps, soit enfin une condition contrôle dans laquelle deux flashs sont réellement présentés. LEs auteurs n'ont sélectionné pour l'analyse que les essais pour lesquels l'illusion s'est produite, c'est-à-dire lorsque le sujet a perçu deux flashs au lieu d'un. L'application du modèle additif montre des interactions vers 180 ms sur les électrodes occipitales — seules celles-ci ont été enregistrées. Ces interactions ressemblent à la différence entre les potentiels évoqués par deux flashs réels et ceux évoqués par un seul flash. Des résultats analogues ont été rapportés par Arden, Wolf et Messiter (2003) et suggèrent également que les interactions audiovisuelles sont d'origine occipitale. Ici encore les effets trouvés pourraient ne pas refléter l'étape d'intégration audiovisuelle mais plutôt les conséquences de cette intégration, c'est-à-dire l'activité visuelle liée à la perception d'un flash illusoire.

4.6.3 Activités corrélées à une illusion audiovisuelle

Une dernière stratégie consiste à comparer des conditions dans lesquelles les mêmes stimuli sont présentés, mais où la perception des sujets diffère selon que l'illusion a eu lieu ou non. Cette stratégie a été mise en œuvre pour étudier les corrélats neurophysiologiques de l'illusion du "croisement/rebond" (*streaming/bouncing*), adaptation au domaine audiovisuel d'un phénomène purement visuel. Dans ce paradigme, le sujet voit deux stimuli visuels identiques en mouvement l'un vers l'autre se croiser puis continuer leur course dans des directions opposées. En l'absence de son, le sujet perçoit dans la plupart des essais deux stimuli qui se croisent. Mais si un son bref est présenté de manière synchrone à la rencontre des deux stimuli, la proportion d'essai dans lequel le sujet perçoit les stimuli rebondir l'un contre l'autre augmente considérablement (Sekuler, Sekuler & Lau, 1997 ; Watanabe & Shimojo, 2001 ; Sanabria, Correa, Lupianez & Spence, 2004). Dans un protocole d'IRMf évènementiel, Bushara et coll. (2003) ont séparé les essais audiovisuels donnant lieu à la perception d'un rebond de ceux donnant lieu à un croisement. La différence entre les deux conditions fait apparaitre un nombre important de structures corticales et sous-corticales qu'il serait trop long de détailler ici. Dans le cas de cette illusion, et contrairement aux effets mis en évidence dans l'illusion "flash/bip", ces activités ne semblent pas uniquement être la conséquence d'une perception différente puisque le même contraste entre rebond et croisement dans une condition visuelle seule ne fait apparaitre aucune activation.

4.7 Corrélats neurophysiologiques de la perception de la parole audiovisuelle

Les études les plus anciennes concernant les corrélats neurophysiologiques de l'intégration des indices auditifs et visuels de parole chez l'homme sont issues de la neuropsychologie et ont essentiellement porté sur les différences interhémisphériques. Certaines études de cas de patients cérébrolésés ont tenté de relier la susceptibilité des patients à l'effet McGurk à la latéralité de leur lésion (Campbell, 1992 ; Campbell et coll., 1990 ; Campbell, Landis & Regard, 1986). D'autres ont étudié l'avantage relatif d'un hémisphère cérébral dans le traitement audiovisuel de la parole en évaluant la probabilité d'un effet McGurk lorsque les stimuli visuels sont présentés de façon tachistoscopique dans un des deux hémichamps visuels (Baynes, Funnell & Fowler, 1994 ; Diesch, 1995). Les résultats de ces études sont largement contradictoires, certaines concluant à une dominance de l'hémisphère gauche, d'autres à celle de l'hémisphère droit, d'autres enfin à l'implication obligatoire des deux hémisphères. Une explication de ces contradictions pourrait tenir à la difficulté de séparer dans les variables affectant l'effet McGurk, celles qui sont imputables au traitements unimodaux, de celles qui sont directement liées à l'intégration des informations auditives et visuelles.

Les premières études de neuroimagerie se sont souvent contenté d'exposer plus ou moins passivement les sujets à des conditions de présentation de la parole auditive, visuelle et audiovisuelle et ont recouru à divers critères pour isoler les interactions audiovisuelles.

Dans une étude en MEG, (Sams & Levänen, 1998) comparent les champs magnétiques évoqués par des syllabes auditives, visuelles et audiovisuelles, présentées dans des blocs expérimentaux séparés. Les syllabes audiovisuelles évoquent une onde tardive vers 450 ms après le son qui ne s'explique pas par la somme des réponses unimodales. Cette onde peut être modélisée par un dipôle de courant qui ressemble à celui de l'onde N1 auditive, d'origine principalement supratemporale.

Puis deux expériences en IRMf vont utiliser deux critères différents : Calvert et coll. (1999) exposent leurs sujets à des blocs de mots (chiffres) auditifs, visuels et audiovisuels, que les sujet doivent se répéter intérieurement (les sujets sont capables de lire les dix chiffres sur les lèvres). L'analyse recherche les voxels qui sont à la fois plus activés en condition audiovisuelle qu'en condition visuelle seule et plus activés en condition audiovisuelle qu'en condition auditive seule. Ces aires comprennent la jonction occipito-pariétale (aire V5) et une partie du gyrus temporal supérieur (cortex auditifs primaire et secondaire).

Dans une seconde expérience Calvert et coll. (2000), le critère utilisé est différent puisqu'il consiste à identifier les voxels montrant une activité super-additive (voir la partie 4.5 page 72). Dans cette expérience les stimuli sont des phrases. Les structures identifiées selon ce critère comprennent une partie du gyrus occipital médian s'étendant jusqu'à V5, le STS antérieur, le cortex auditif primaire, le gyrus frontal médian, le lobule pariétal inférieur. Cette expérience comprenait également une condition audiovisuelle dans laquelle les phrases entendues et vues sur le visage du locuteur ne correspondaient pas. Les auteurs ont postulé que les aires d'intégrations devraient montrer une activation sous-additive dans cette condition. La seule aire respectant le critère de sous-additivité, ainsi que celui de super-additivité pour la condition audiovisuelle congruente, est le STS. Cette aire avait déjà été identifiée avec les mêmes critères pour des stimuli autres que la parole (Calvert et coll., 2001).

D'autres études ont tenté d'isoler les aires cérébrales plus activées lorsque le stimulus audiovisuel respectait les règles de coïncidence spatiale et temporelle que lorsqu'il ne les respectaient pas : Olson, Gatenby et Gore (2002) ont comparé une condition de présentation de mots audiovisuels synchrones à une condition de présentation où les informations auditives et visuelles étaient séparées d'une seconde, dans une expérience où l'attention des sujets n'était pas contrôlée. Les structures activées de manière différentielle sont le claustrum (une structure sous-corticale située derrière l'insula) et le pôle temporal. Macaluso et coll. (2004) ont étudié les effets de la séparation spatiale et de la séparation temporelle des mots auditifs et visuels dans une tâche où les sujets devaient réaliser une tâche sémantique. Les aires corticales activées de façon préférentielle lorsque les indices sont spatialement et temporellement congruents sont les cortex occipitaux latéral et dorsal. Étant donné la résistance connue des effets d'intégration de la parole à la séparation spatiale (voir la partie 3.3 page 57), les zones activées préférentiellement par les stimuli synchrone, quelle que soit la séparation spatiale, sont susceptibles d'être des aires d'intégration audiovisuelle de la parole. Dans cette étude, les aires comprennent le gyrus fusiforme et le STS.

Certaines études enfin ont utilisé les phénomènes comportementaux connus de l'influence visuelle sur la perception de parole pour identifier les aires impliquées dans ces effets com-

portementaux, en particulier l'amélioration de l'intelligibilité dans le bruit et l'effet McGurk. Pour la perception de la parole dans le bruit, deux études ont cherché à identifier les aires cérébrales montrant une influence plus forte des indices visuels dans le bruit que sans le bruit (ce qui correspond à une interaction entre la présence d'indices visuels et la présence de bruit). Dans une étude en EEG (Callan, Callan, Kroos & Vatikiotis-Bateson, 2001), dans laquelle le sujet devait identifier un mot auditif accompagné ou non des indices visuels correspondants, dans le bruit ou dans le silence, ce critère a permis d'isoler deux composantes des activités oscillatoires dans la bande de fréquence 45-70 Hz (à l'issue d'une analyse en composante indépendante) : l'une entre 150 et 300 ms de traitement, compatible avec l'activation de la partie supérieur du cortex temporal, l'autre soutenue dans le temps compatible avec l'activation d'un réseau fronto-pariéto-temporo-occipital. Cette étude a porté un sujet unique. Dans une étude de groupe en IRMf, utilisant à peu près le même protocole expérimental et une anlyse analogue (Callan et coll., 2003), les structures remplissant le critère étaient la partie supérieure du cortex temporal, dont le cortex auditif primaire, le GTM, le gyrus temporal supérieur (GTS) et le STS, ainsi que le pôle temporal, V5, l'aire de Broca, l'insula, le claustrum et les ganglions de la base.

En ce qui concerne l'effet McGurk, j'ai déjà mentionné les études qui ont montré l'existence d'un MMN à la déviance auditive illusoire d'une syllabe McGurk dans la partie 4.6.1 page 72. Ces études montrent qu'au stade de traitement correspondant à la MMN, l'intégration audiovisuelle a déjà eu lieu. D'autres études vont tenter d'identifier les structures cérébrales qui sont plus activées lorsque des syllabes incongruentes donnent lieu à la perception d'une syllabe illusoire (fusion) que lorsque l'illusion n'a pas lieu. La première étude (Sekiyama, Kanno, Miura & Sugita, 2003), réalisée en IRMf et en TEP, tire parti du fait que les locuteurs japonais sont plus sensibles à l'effet McGurk dans le bruit et compare une condition audiovisuelle incongruente dans le bruit donnant une proportion importante d'illusions à une condition audiovisuelle incongruente sans bruit donnant moins d'illusion. Le problème avec cette analyse, c'est qu'elle confond l'effet du bruit acoustique et l'effet lié à l'existence d'une illusion. Une seconde étude en IRMf (J. A. Jones & Callan, 2003) manipule la proportion d'illusions McGurk en faisant varier la synchronie entre la syllabe auditive et visuelle. Ici encore, le fait de comparer les conditions synchrones et asynchrones ne permettait pas de différencier les effets de l'asynchronie de ceux liés à l'illusion. Néanmoins l'analyse choisie consistait à rechercher les activations dans les conditions audiovisuelles incongruentes (estimée à partir d'un condition contrôle dans laquelle les sujets voient un visage immobile) qui corrèlent significativement avec la proportion d'illusions McGurk effectivement mesurée chez les sujets, quelle que soit la synchronie. Cette analyse montre que l'activation de la jonction temporo-occipital, proche de V5 est corrélée négativement à la proportion d'illusions. Notons que dans cette même étude, une condition audiovisuelle congruente permettait d'identifier des aires différemment activées par des syllabes audiovisuelles congruentes et incongruentes, à savoir le gyrus supra-marginale et le lobule pariétal inférieur.

Le STS ayant été impliqué à plusieurs reprises dans les études précédentes, certaines études d'IRMf se sont spécifiquement intéressées à cette structure. Dans un protocole d'IRMf évènementiel, Wright, Pelphrey, Allison, McKeown et McCarthy (2003) ont com-

paré la réponse hémodynamique à des stimuli auditifs, visuels et audiovisuels. Contrairement au STG qui montre une activité audiovisuelle supérieure ou égale à la somme des activités auditives et visuelles sur toute sa longueur (avec une réponse hémodynamique visuelle nulle ou négative), les aires bordant le STS peuvent montrer soit une super-additivité, soit une sous-additivité (dans la partie postérieure du STS). Beauchamp, Argall et coll. (2004) ont, de leur côté, montré que des stimuli audiovisuels de parole activaient le STS postérieur de la même manière que des évènements audiovisuels non langagiers, avec la même répartition de sous-aires auditives, visuelles et audiovisuelles (voir la partie 4.4 page 70).

Comme on peut le constater, la plupart des premières études des corrélats neurophysiologiques de l'intégration audiovisuelle dans la perception de la parole ont été réalisées en imagerie fonctionnelle hémodynamique. Les études électrophysiologiques, en EEG ou en MEG, n'ont pas tardé à suivre, à partir de 2003, en même temps que nous finissions de réaliser notre première étude d'EEG. Afin de respecter la chronologie des évènements, les résultats de ces études seront exposés dans les discussion de nos différentes études sur la parole.

4.8 Conclusion

L'impression qui se dégage des résultats de la neuroimagerie chez l'homme, c'est la multiplicité des sites cérébraux activés spécifiquement par la présentation d'un stimulus audiovisuel, selon les types de stimuli, les critères et les paradigmes expérimentaux utilisés. Comme beaucoup de résultats sont issus de l'IRMf, il est souvent difficile de savoir à quels stades de traitements correspondent les différentes activations observées, en dépit des critères d'intégration utilisés. Les études en EEG montrent cependant que ces activations peuvent avoir lieu à de multiples stades de traitement et impliquer les cortex unisensoriels dès les première étapes de l'analyse. Ces données électrophysiologiques obtenues chez l'homme ne s'accordent guère avec un modèle de convergence tardive tel qu'il a été exposé dans la partie 1.5 page 17 et qui est communément accepté dans le domaine des neurosciences cognitives.

Chapitre 5

Problématique générale

Il semble que l'on puisse conclure à l'issue de cette revue que l'intégration multisensorielle lors de la perception d'un évènement audiovisuel n'est décidément pas un phénomène unitaire. Au niveau neurophysiologique et anatomique, les mécanismes neuronaux pouvant en rendre compte sont multiples et différents modes de convergence semblent coexister dans le système nerveux central (chapitre 1 page 5). Au niveau comportemental, les effets d'interaction entre modalités sensorielles sont nombreux et une partie d'entre eux au moins implique l'existence de stades d'interactions précoces et d'échanges d'informations entre systèmes sensoriels (chapitre 2 page 21). L'utilisation de la neuroimagerie (chapitre 4 page 63) a confirmé la multiplicité et la spécificité des réseaux réseaux impliqués dans différentes tâches.

Tous ces éléments indiquent que le traitement d'un évènement audiovisuel peut mettre en jeu différents niveaux de convergence et modes d'intégration des informations auditives et visuelles. Les travaux présentés dans cette thèse visent à caractériser ces interactions chez l'homme à la fois dans leurs dimensions temporelle et spatiale. Pour cela, nous avons utilisé des enregistrements de potentiels évoqués et de champs magnétiques évoqués cartographiques, c'est-à-dire sur l'ensemble du scalp du sujet, ce qui permet à la fois de connaitre avec une grande précision la chronologie des activations cérébrales et, dans une certaine mesure, de localiser les structures cérébrales impliquées. Nous avons également utilisé des enregistrements de potentiels évoqués intracérébraux chez le patient épileptiques, qui permettent à la fois une grande précision temporelle et spatiale.

Les travaux de ma thèse concernent deux aspects de la perception d'un évènement audiovisuel. Le premier volet concerne l'étude des interactions audiovisuelles dans la perception d'un évènement audiovisuel typique : la parole. Le but était d'établir le décours temporel des interactions entre informations auditives et visuelles lors de la perception de la parole naturelle. En effet, les études psycholinguistiques concernant les effets des informations visuelles sur la perception auditive de la parole ont montré que les interactions dans le traitement des deux modalités pouvaient avoir lieu à différents niveaux. Comme on vient de le voir, de nombreuses études d'imagerie fonctionnelle utilisant différents critères pour l'identification des structures impliquées dans cette intégration ont montré des activations dans diverses aires corticales et sous-corticales, principalement en utilisant l'IRMf. Cependant peu d'études s'étaient intéressées à la façon dont ces différents effets peuvent

s'articuler dans le temps. La technique des potentiels évoqués électriques permet d'étudier la dynamique de ces interactions. Des travaux menées précédemment à l'unité 280 par Alexandra Fort, Marie-Hélène Giard et Frank Peronnet avaient introduit l'utilisation du modèle additif pour l'étude de la dynamique des interactions audiovisuelles chez l'homme lors de la perception d'objets bimodaux (voir Fort & Giard, 2004, et la partie 4.2.2 page 67 pour une revue). Il était donc tout naturel d'appliquer ce modèle additif à la perception de la parole.

Le deuxième volet de cette thèse porte sur la représentation d'un évènement audiovisuel en mémoire sensorielle, et ce par le biais d'un marqueur électrophysiologique de cette représentation. Contrairement à une idée fort répandue, et comme cela a été amplement démontré dans l'introduction pour l'audition et la vision, la convergence des informations de différentes modalités ne se fait pas uniquement dans des aires corticales associatives à une étape tardive du traitement. Les données sur les effets d'interaction audiovisuelle dans les structures sous-corticales et dans les cortex modalité-spécifique chez l'animal et chez l'homme, l'existence d'illusions audiovisuelles irrépressibles ou d'effets audiovisuels précoces dans la détection de stimuli audiovisuels ou la perception de la parole, même si elles n'excluent ni une spécificité relative des cortex unisensoriels, ni l'existence d'aires associatives, montre qu'il n'y a pas de ségrégation stricte des différentes modalités sensorielles dans le système nerveux central. On peut donc légitimement se demander si certains processus décrits comme modalité-spécifiques ne sont pas moins spécifiques qu'on ne le pensait auparavant. Le processus qui nous intéresse est celui de la détection auditive du changement. La détection d'un changement dans un environnement acoustique régulier est un processus largement automatique, qui génère dans les potentiels évoqués un onde spécifique : la MMN vers 150 ms après la stimulation. Ce processus automatique implique l'existence d'une trace mnésique des régularités acoustiques à laquelle un son déviant doit être comparé. Étant donné l'existence d'interactions audiovisuelles dès les premiers niveaux de traitement, cette représentation est susceptible d'être modifiée par des informations visuelles, notamment la détection d'un changement visuel.

Les deux processus cognitifs (perception de la parole et mémoire sensorielle auditive) auxquels nous nous sommes intéressés présentent le point commun d'être avant tout des processus auditifs. Nous nous attendions donc surtout, mais pas exclusivement, à mettre en évidence des influences des informations visuelles sur les traitements dans le cortex auditif.