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 :

Mr Pascal Barone (Examinateur)
Mme Nicole Bruneau (Rapporteur)
Mr Jean-Luc Schwartz (Rapporteur)
Mme Marie-Hélène Steiner-Giard (Directrice de thèse)
Mr Rémy Versace (Examinateur)

Table des matières

I Revue de la littérature **3**

1 Convergence audiovisuelle en neurophysiologie **5**
1.1 Aires associatives corticales 5
1.1.1 Études électrocorticographique (ECoG) de la convergence multisensorielle 5
1.1.2 Convergence audiovisuelle au niveau du neurone unitaire 8
1.1.3 Aires de convergence dans le cortex frontal 9
1.1.4 Effet de l'anesthésie sur les interactions multisensorielles 9
1.2 Convergence audiovisuelle dans le cortex visuel 10
1.3 Convergence corticale chez l'homme 11
1.4 Convergence sous-corticale 12
1.4.1 Colliculus Supérieur / Tectum optique 13
1.4.2 Autres structures sous-corticales 16
1.5 Études anatomiques de la convergence multisensorielle 17
1.6 Conclusion 19

2 Interactions Audiovisuelles en psychologie **21**
2.1 Effets intersensoriels sur les capacités perceptives 22
2.1.1 Effets dynamogéniques 22
2.1.2 Modèles explicatifs de l'effet dynamogénique 22
2.1.3 Effet dynamogénique et théorie de la détection du signal 24
2.1.4 Modèles de détection d'un stimulus bimodal au seuil 24
2.2 Correspondance des dimensions synesthésiques 25
2.2.1 Établissement des dimensions synesthésiques 26
2.2.2 Réalité des correspondances synesthésiques 27
2.2.3 Correspondance des intensités 29
2.2.4 Résumé 30
2.3 Temps de réaction audiovisuels 31
2.3.1 Premières études 31
2.3.2 Paradigme du stimulus accessoire 33
2.3.3 Paradigme d'attention partagée 36
2.4 Conflit des indices spatiaux auditifs et visuels 42
2.4.1 Ventriloquie 43
2.4.2 Facteurs influençant l'effet de ventriloquie 45
2.4.3 Niveau des interactions dans l'effet de la ventriloquie 46

2.5 Conflit des indices temporels 47
2.6 Conclusion 48

3 Perception audiovisuelle de la parole 49
3.1 Contribution visuelle à l'intelligibilité 49
3.1.1 Complémentarité des informations auditives et visuelles de parole . 50
3.1.2 Redondance des informations auditives et visuelles de parole 51
3.1.3 Facteurs liés à la connaissance de la langue 51
3.2 Effet McGurk 52
3.2.1 L'hypothèse VPAM 53
3.2.2 Intégration audiovisuelle pré-phonologique 54
3.2.3 Influence des facteurs linguistiques et cognitifs 55
3.3 Facteurs spatiaux et temporels 56
3.4 Modèles de perception de la parole audiovisuelle 58
3.4.1 Modèles post-catégoriels 58
3.4.2 Modèles pré-catégoriels 60
3.5 Conclusion 61

4 Intégration AV en neurosciences cognitives 63
4.1 Comportements d'orientation 63
4.1.1 Orientation vers un stimulus audiovisuel chez l'animal 64
4.1.2 Saccades oculaires vers un stimulus audiovisuel, chez l'homme . . . 65
4.1.3 Expériences chez l'animal alerte et actif 66
4.2 Effet du stimulus redondant 67
4.2.1 Premières études 67
4.2.2 Tâches de discrimination 67
4.2.3 Tâche de détection 68
4.3 Perception des émotions 69
4.4 Objets écologiques audiovisuels 70
4.5 Conditions limites de l'intégration AV 71
4.6 Illusions audiovisuelles 72
4.6.1 Intégration audiovisuelle pré-attentive 72
4.6.2 Application du modèle additif 73
4.6.3 Activités corrélées à une illusion audiovisuelle 74
4.7 Perception audiovisuelle de la parole 74
4.8 Conclusion 77

5 Problématique générale 79

II Méthodes 81

6 Approches électrophysiologiques 83
6.1 Bases physiologiques des mesures (s)EEG/MEG 83
6.2 ElectroEncéphaloGraphie (EEG) 84

6.2.1 Enregistrement 84
6.2.2 Analyse des potentiels évoqués (PE) 86
6.3 MagnétoEncéphaloGraphie (MEG) 90
6.3.1 Champs magnétiques cérébraux 90
6.3.2 Procédure d'enregistrement 91
6.4 StéréoElectroEncéphaloGraphie (sEEG) 92
6.4.1 Localisation des électrodes 92
6.4.2 Procédure d'enregistrement 93
6.4.3 Calcul du PE et rejet d'artéfacts 94
6.4.4 Résolution spatiale et représentation spatiotemporelle 94
6.4.5 Étude de groupe et normalisation anatomique 95

7 Approche méthodologique de l'intégration AV 99
7.1 Falsification de l'inégalité de Miller 99
7.1.1 Bases mathématiques et postulats 99
7.1.2 Application de l'inégalité 102
7.1.3 Biais potentiels 104
7.1.4 Analyse statistique de groupe 105
7.2 Modèle additif 106
7.2.1 Falsification du modèle additif en EEG/MEG 107
7.2.2 Interprétation des violations de l'additivité en EEG/MEG 109
7.2.3 Comparaison avec le critère d'additivité en IRM fonctionnelle . . . 109

8 Méthodes statistiques en (s)EEG/MEG 111
8.1 Tests multiples 111
8.2 Tests Statistiques sur les données individuelles 113
8.2.1 Tests sur les essais élémentaires 113
8.2.2 Test du modèle additif par randomisation pour des données non appariées 114
8.2.3 Remarques 115

III Interactions audiovisuelles dans la perception de la parole 117

9 Étude en EEG et comportement 119
9.1 Rappel de la problématique 119
9.2 Méthodes 120
9.2.1 Sujets 120
9.2.2 Stimuli 120
9.2.3 Procédure 121
9.2.4 Expérience comportementale complémentaire 122
9.2.5 Analyse des résultats 122
9.3 Résultats 123
9.3.1 Résultats comportementaux 123
9.3.2 Résultats électrophysiologiques 123

9.4 Discussion 125
9.4.1 Comportement 125
9.4.2 Résultats électrophysiologiques 127

10 Étude en sEEG **131**
10.1 Introduction 131
10.2 Méthodes 134
10.2.1 Patients 134
10.2.2 Stimuli et procédure 134
10.2.3 Calcul des potentiels évoqués 134
10.2.4 Analyses statistiques 135
10.3 Résultats 136
10.3.1 Données comportementales 136
10.3.2 Réponses évoquées auditives 136
10.3.3 Réponses évoquées visuelles 138
10.3.4 Violations du modèle additif 141
10.3.5 Relations entre réponses auditives, visuelles et interactions audiovisuelles 144
10.4 Discussion 145
10.4.1 Activité du cortex auditif en réponse aux indices visuels de parole . 146
10.4.2 Interactions audiovisuelles 149
10.4.3 Comparaison avec l'expérience EEG de surface 151

11 Effet d'indiçage temporel **153**
11.1 Introduction 153
11.2 Expérience comportementale 1 155
11.2.1 Méthodes 156
11.2.2 Résultats 159
11.2.3 Discussion 162
11.3 Expérience comportementale 2 163
11.3.1 Méthodes 164
11.3.2 Résultats 166
11.3.3 Discussion 169
11.4 Discussion générale 170

IV Interactions audiovisuelles en mémoire sensorielle **173**

12 Introduction générale **175**
12.1 MMN Auditive 175
12.2 Rappel de la problématique 176

13 Étude comportementale **179**
13.1 Introduction 179
13.2 Méthodes 180

13.2.1 Sujets 180
13.2.2 Stimuli 180
13.2.3 Procédure 181
13.2.4 Analyses 182
13.3 Résultats 182
13.4 Discussion 183

14 Additivité des MMNs auditives et visuelles 185
14.1 Introduction 185
14.2 Méthodes 187
14.2.1 Sujets 187
14.2.2 Stimuli 187
14.2.3 Procédure 187
14.2.4 Analyses 188
14.3 Résultats 188
14.4 Discussion 191

15 Représentation auditive d'une régularité AV 195
15.1 Introduction 195
15.2 Méthodes 196
15.2.1 Sujets 196
15.2.2 Stimuli 196
15.2.3 Procédure 197
15.2.4 Analyses 197
15.3 Résultats 198
15.4 Discussion 201

16 MMN à la conjonction audiovisuelle 205
16.1 Introduction 205
16.2 Méthodes 207
16.2.1 Sujets 207
16.2.2 Stimuli 207
16.2.3 Procédure 207
16.2.4 Analyses 208
16.3 Résultats 208
16.4 Expérience comportementale complémentaire 210
16.5 Discussion 211

V Discussion générale 215

17 Discussion générale 217
17.1 Interactions audiovisuelles précoces dans la perception de la parole 217
17.2 Représentation d'un évènement audiovisuel en mémoire sensorielle auditive 218
17.3 Interactions audiovisuelles dans le cortex auditif 219

A Données individuelles des patients **223**

B Articles **239**

Bibliographie **287**

Introduction

Nous appréhendons le monde extérieur par différentes modalités sensorielles. Or certains évènements peuvent être perçus par le biais de plusieurs modalités à la fois. Que se passe-t-il lorsque le système cognitif est confronté à un tel évènement, par exemple un stimulus défini par des attributs auditifs et visuels ? À quelles étapes de la chaine des traitements opérés par les différentes structures du système nerveux central, des interactions ont-elles lieu entre les informations provenant des récepteurs visuels et celles provenant des récepteurs auditifs ?

Le phénomène perceptif qui nous intéresse est donc celui de la stimulation simultanée des organes récepteurs des modalités sensorielles auditives et visuelles par un évènement bimodal du monde extérieur (les mêmes questions se posent pour d'autres combinaisons de modalités sensorielles, mais nous nous limiterons ici au cas audiovisuel). Cette façon d'aborder le problème des interactions audiovisuelles est relativement récente dans la littérature scientifique. Même dans la littérature concernée directement par les interactions entre modalités sensorielles auditive et visuelle, beaucoup d'études, surtout les plus anciennes, ont utilisé des stimuli auditifs et visuels qui n'avaient pas forcément de rapport avec un évènement bimodal plausible, et nous verrons que la notion d'évènement audiovisuel, en tant que ce qui donne lieu à des interactions entre le informations auditives et visuelles dans une situation écologique, s'est en fait construite assez progressivement.

Dans l'étude des interactions audiovisuelles, on a coutume de distinguer entre interactions "précoces" et interactions "tardives" (ou convergence) : les premières correspondraient à l'influence que peut avoir une modalité sensorielle sur les traitements propres à une autre modalité sensorielle ; les secondes correspondraient à une convergence des informations auditives et visuelles vers des traitements de plus haut niveau. Une telle distinction suppose implicitement que les traitements auditifs et visuels sont d'abord séparés (pour que des interactions "précoces" puissent avoir lieu), puis convergent à un moment donné vers des traitements communs aux informations des deux modalités (pour pouvoir donner lieu à des interactions "tardives"). De fait, beaucoup d'auteurs ont cherché à caractériser les interactions audiovisuelles en rejetant un modèle de séparation des traitements auditifs et visuels. Nous verrons que cela est vrai aussi bien dans les disciplines biologiques que dans les disciplines psychologiques. Pour beaucoup d'études récentes, le modèle à falsifier est un modèle de convergence tardive dans lequel les traitements auditifs et visuels sont séparés jusqu'à des processus de haut niveau. Or, s'il est évident que les organes récepteurs sont séparés, nous essaierons de montrer, dans une revue de la littérature, que le niveau de traitement, aussi bien en termes temporels, fonctionnels qu'anatomiques, à partir duquel les informations auditives et visuelles convergent n'a jamais réellement fait l'objet d'un

consensus.

Cette revue de la littérature sera organisée à la fois chronologiquement et en fonction des techniques utilisées pour étudier les interactions audiovisuelles.

Nous nous intéresserons d'abord aux données de la neurophysiologie et de la neuroanatomie, qui proviennent essentiellement de l'animal. Dans cette partie nous passerons en revue des études expérimentales, pour la plupart relativement anciennes, qui définissent la convergence audiovisuelle sur des critères neurophysiologiques ou anatomiques.

Ensuite, nous verrons comment des interactions audiovisuelles dans le fonctionnement cognitif humain ont pu être mises en évidence très tôt par des mesures objectives du comportement. Ces études, qui remontent jusqu'au début du siècle dernier, ont mis en évidence des effets intersensoriels de facilitation ou d'inhibition des performances comportementales.

Les résultats concernant la perception de la parole seront regroupés dans une partie indépendante étant donné qu'ils constituent un domaine tout à fait particulier et très riche de la littérature sur les interactions audiovisuelles.

Enfin la dernière partie de l'introduction théorique concernera des études plus récentes qui ont cherché à caractériser les interactions audiovisuelles avec des techniques d'investigation neurophysiologiques en tentant de les relier à des résultats comportementaux chez les mêmes sujets (animaux ou hommes).

Les travaux expérimentaux de cette thèse s'inscriront dans deux axes, soulignés dans cette introduction, pour mettre en évidence des interactions audiovisuelles dans le cortex auditif chez l'homme. Nous étudierons d'une part les processus d'intégration audiovisuelle mis en jeu lors de la perception d'évènements audiovisuels ayant une réalité plausible et nous tenterons d'autre part de relier des mesures neurophysiologiques de ces interactions chez l'homme à des phénomènes de facilitation de traitement mis en évidence de façon comportementale.

Nous nous focaliserons sur deux fonctions cérébrales mettant essentiellement en jeu le cortex auditif. Dans la première partie, nous tenterons de montrer par quels processus et à quelles étapes du traitement, les informations visuelles peuvent influencer le traitement auditif de la parole. Pour cela, nous avons utilisé des mesures comportementales, des mesures électrophysiologiques de surface chez des sujets sains et des mesures électrophysiologiques invasives chez des patients épileptiques. Dans la deuxième partie, nous tenterons de montrer comment des informations visuelles peuvent influencer la représentation des sons en mémoire sensorielle auditive lors de la perception d'un évènement bimodal. Pour cela nous avons utilisé des mesures comportementales, électrophysiologiques et magnétoencéphalographiques chez le sujet sain.

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.

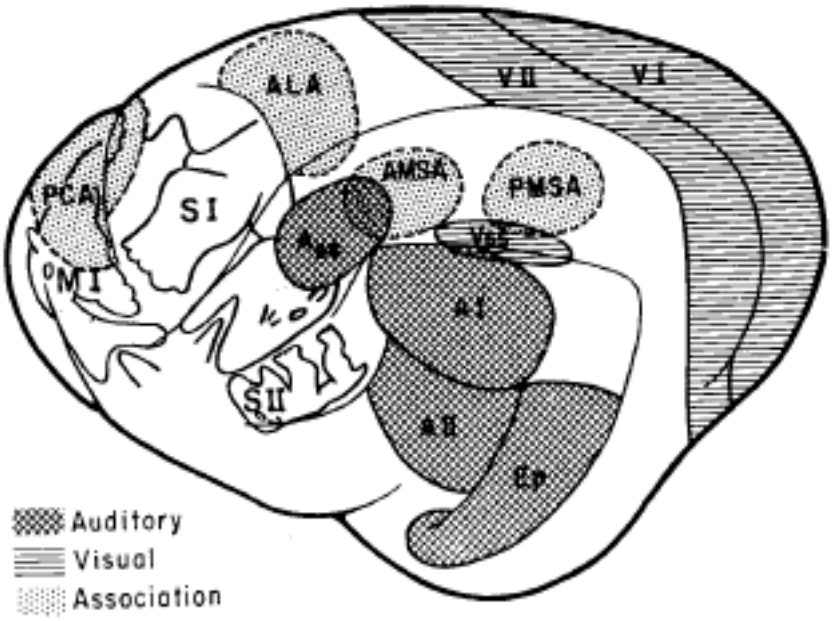

FIG. 1.1 – Localisation des aires unisensorielles et associatives chez le chat. AI : cortex auditif primaire ; AII : cortex auditif secondaire ; ALA : aire latérale antérieure ; AMSA : gyrus suprasylvien antérieur ; MI : cortex moteur primaire ; PCA : aire péricruciée ; PMSA : gyrus suprasylvien postérieur ; 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

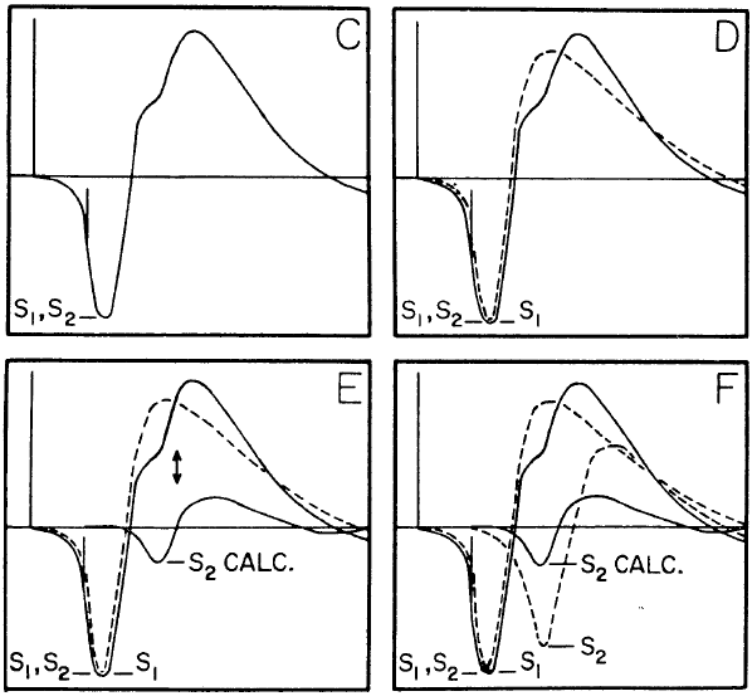

FIG. 1.2 – Illustration de la méthode algébrique utilisée par Thompson, Smith et Bliss (1963). Le but est de savoir si des réponses à deux stimuli S_1 et S_2 enregistrées sur la même électrode sont générées par deux populations neuronales distinctes : si oui, alors la somme des réponses enregistrées 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 5e relai synaptique cortical dans les zones de convergence hétéromodales définies plus haut.

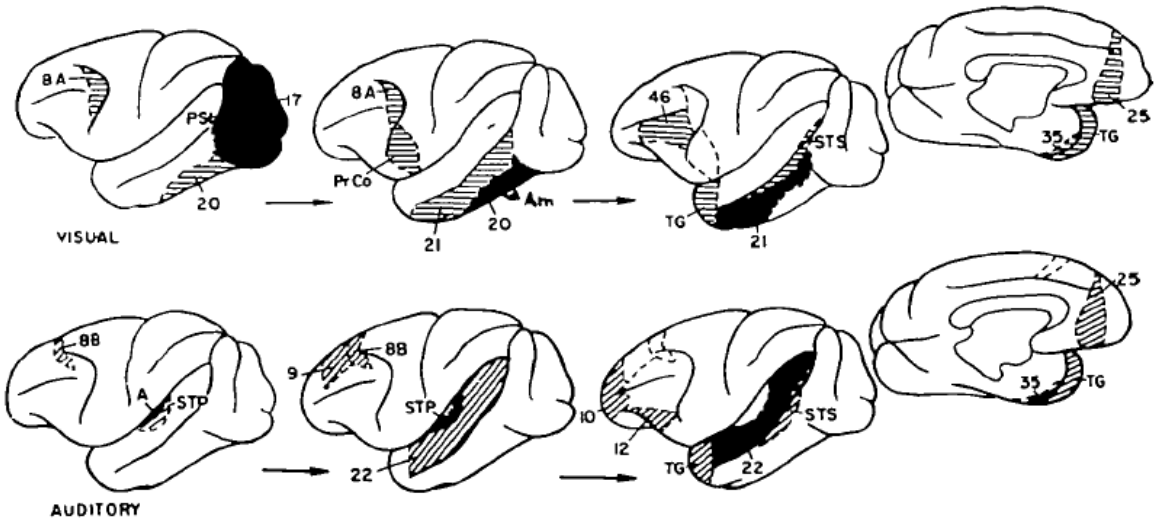

FIG. 1.3 – Schéma récapitulatif des projections cortico-corticales du singe rhésus : sur chaque carte, les zones en noir représentent les zones lésées et les zones hachurées celles où des fibres dégénérées sont trouvées, c'est-à-dire les aires de projection de la zone lésée. Chaque carte représente une 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,

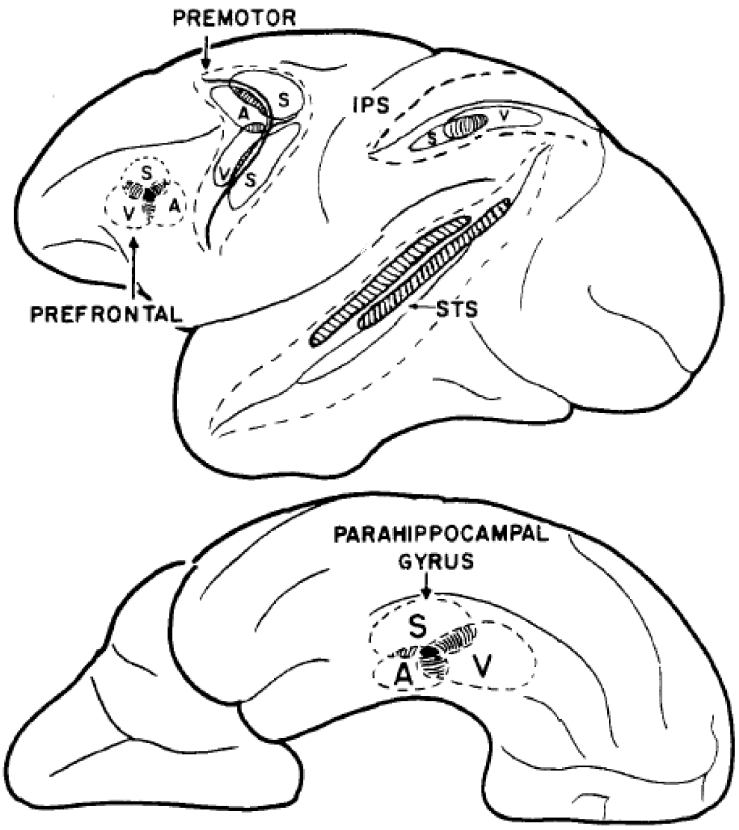

FIG. 1.4 – Aires de convergence définies par la méthode des lésions chez le singe rhésus. A : aires de projection auditive, V : Aires de projection visuelle, S : Aires de projection somesthésique. IPS : sillon intra-pariétal, STS : Sillon temporal supérieur. D'après Pandya et Seltzer (1982)

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).

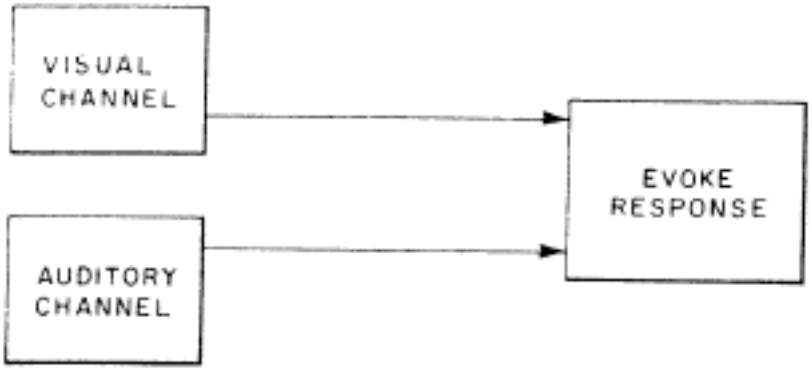

FIG. 2.1 – Illustration du modèle d'indépendance de Hershenson (1962). D'après Nickerson (1973).

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.

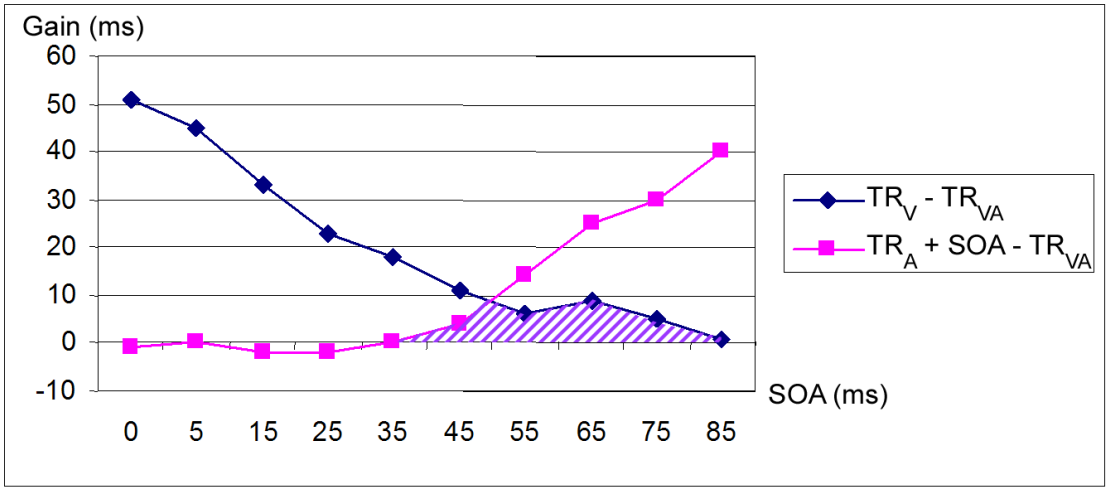

FIG. 2.2 – Facilitation par rapport aux TR unimodaux en fonction du délai séparant le stimulus visuel du stimulus auditif (SOA), sous l'hypothèse que le sujet répond au premier signal traité. La courbe bleue présente le gain de TR en condition bimodale par rapport à la condition visuelle. La courbe rose 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.

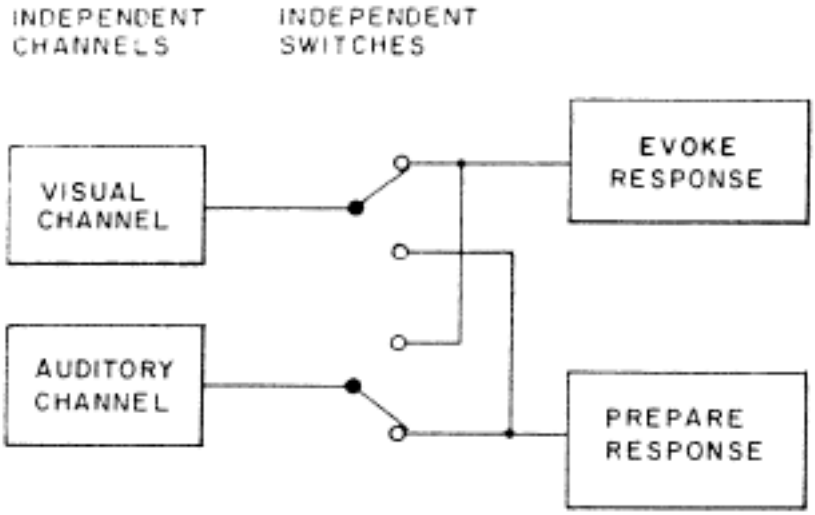

FIG. 2.3 – D'après Nickerson (1973).

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).

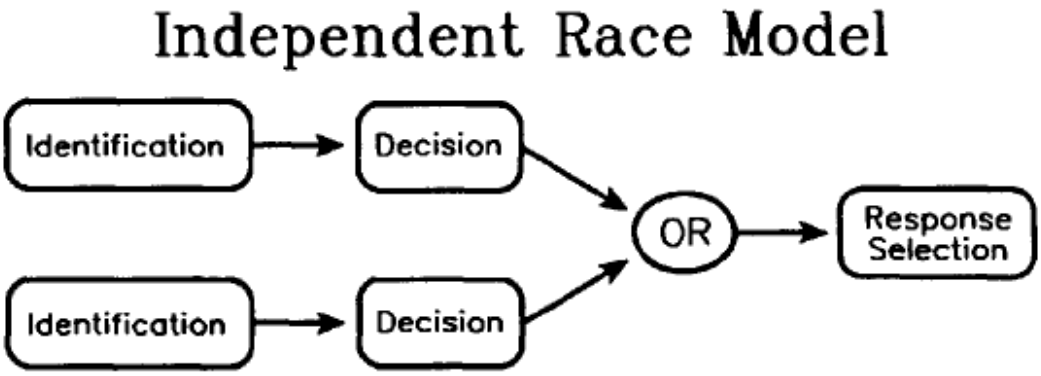

FIG. 2.4 – Modèle d'activations séparées, encore appelé modèle de compétition. Chaque cible auditive ou visuelle est traitée et évaluée indépendamment l'une de l'autre. La première évaluation terminée déclenche la sélection de la réponse et détermine donc le TR bimodal. D'après Mordkoff et Yantis (1991).

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.

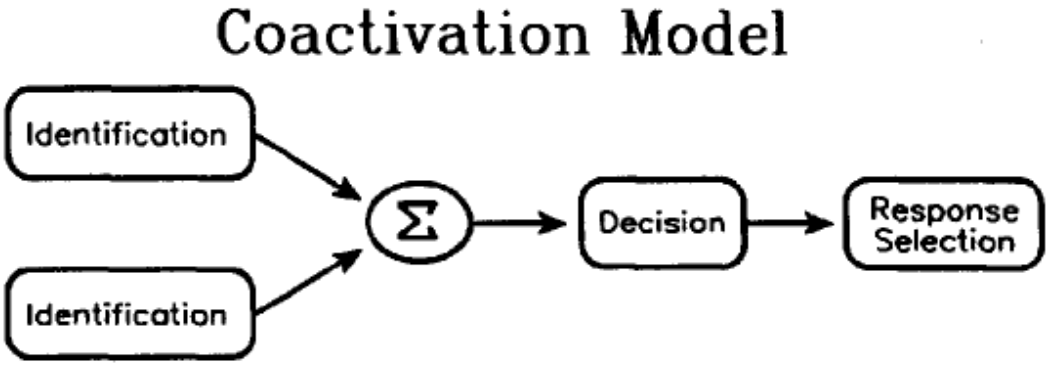

FIG. 2.5 – Modèle de coactivation. D'après Mordkoff et Yantis (1991).

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.

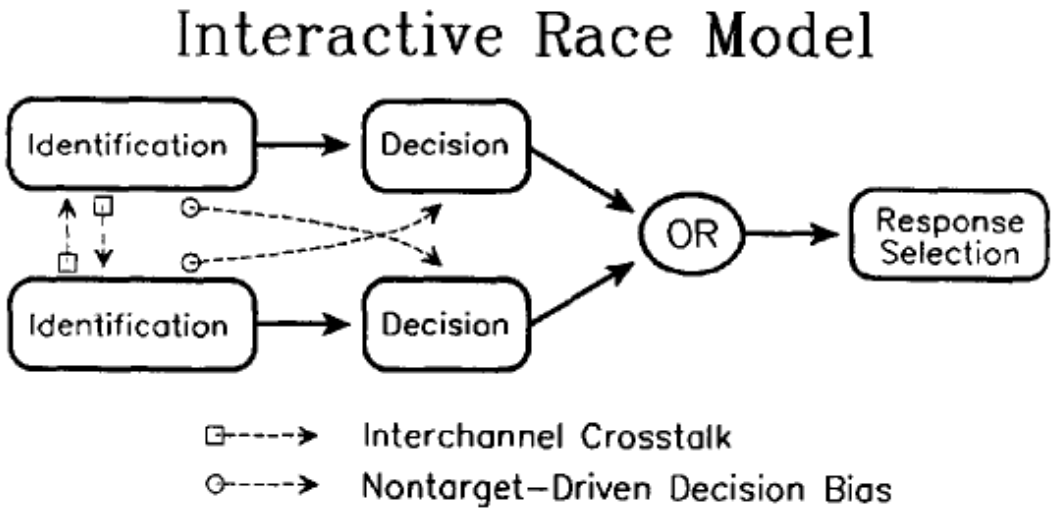

FIG. 2.6 – Modèle de compétition interactif. D'après Mordkoff et Yantis (1991).

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 *prelabeling 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 lorsque qu'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 audiovisuelles 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.

Deuxième partie

Méthodes

Chapitre 6

Approches électrophysiologiques

Les mesures réalisées lors de nos protocoles expérimentaux sont principalement, outre des mesures comportementales de temps de réponse, des mesures de l'activité électrique cérébrale évoquée par des stimulations auditives et/ou visuelles. La technique principalement utilisée est l'ElectroEncéphaloGraphie (EEG). Dans deux expériences, nous avons également utilisé la stéréoElectroEncéphaloGraphie (sEEG) et la MagnétoEncéphaloGraphie (MEG).

6.1 Bases physiologiques des mesures (s)EEG/MEG

Ces trois techniques enregistrent 3 aspects différents d'une activité électrique intracérébrale ayant, a priori, une origine commune. On admet généralement que cette activité électrique reflète les échanges transmembranaires d'ions ayant lieu au niveau cellulaire, lors des potentiels post-synaptiques dans les neurones corticaux de type pyramidal. L'arrivée d'un potentiel d'action sur les terminaisons synaptiques situées sur la membrane d'un neurone provoque l'ouverture de canaux ioniques sur cette membrane, et la formation de puits et de sources de courant vis à vis du milieu extra-cellulaire (enregistrable au niveau cellulaire sous la forme d'un potentiel post-synaptique). Dans les cellules pyramidales, les puits et les sources de courant ont tendance à se répartir de manière ordonnée, les puits au niveau de la dendrite apicale, les sources au niveau du corps cellulaire (figure 6.1.A page suivante), créant l'équivalent d'un dipôle de courant. Ces cellules pyramidales étant disposées parallèlement entre elles et perpendiculairement à la surface du cortex (figure 6.1.B page suivante), une population de tels neurones activés simultanément se comporte, à un niveau macroscopique, comme un dipôle de courant résultant de l'ensemble des dipôles au niveau cellulaire (on parle de dipôle de courant équivalent, figure 6.1.C page suivante). Comme les milieux extracellulaires sont résistifs, cette circulation de courants entraine la formation de champs de potentiel électriques, mesurables soit à l'intérieur de la boite crânienne (sEEG) soit à l'extérieur (EEG et MEG). Les courants électriques créés par les populations de neurones (pyramidaux) diffusent à travers des milieux de conductivité variable (tissu cérébral, liquide céphalo-rachidien, os). Il est important pour l'interprétation de l'EEG et de la MEG de comprendre qu'un seul dipôle équivalent (par exemple l'activation d'une région corticale) induit une distribution particulière de potentiels ou de champs magnétiques sur

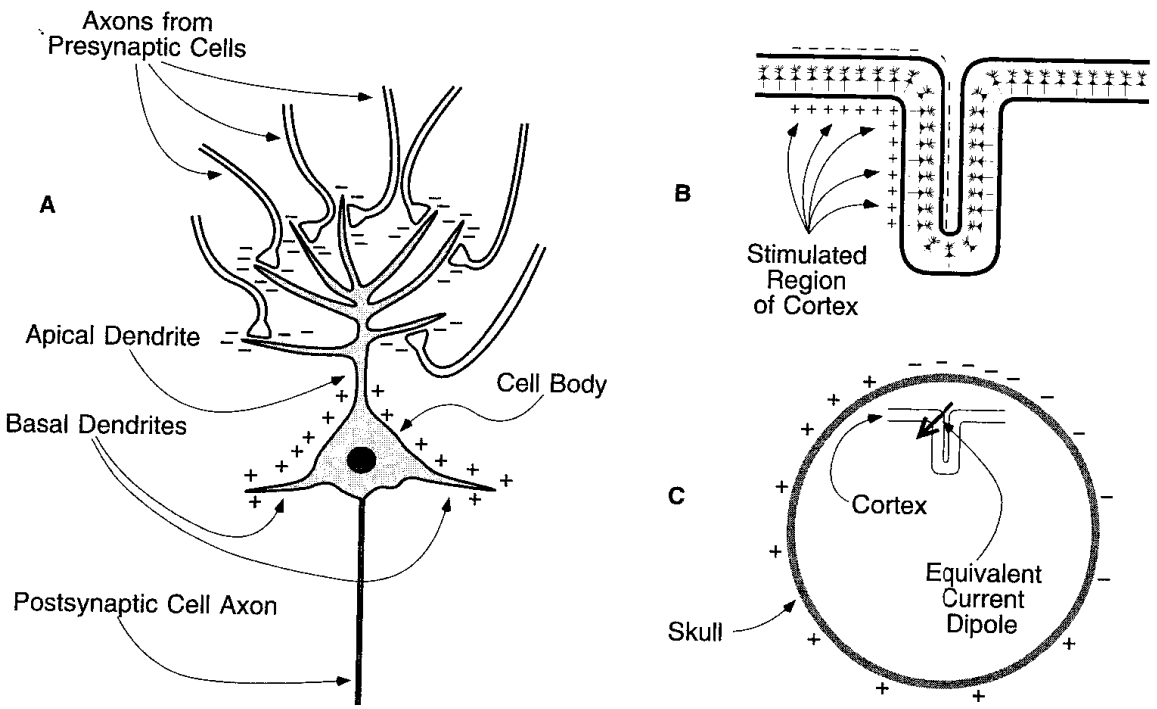

FIG. 6.1 – Bases neuronales du signal électrophysiologique recueilli en EEG de scalp. A. puits (-) et sources (+) de courants dans le milieu extra-cellulaire d'une cellule pyramidale. B. Orientation des cellules pyramidales dans le cortex cérébral. C. Orientation du cortex par rapport à la surface du crâne, dipôle de courant équivalent et potentiels électriques positifs (+) et négatifs (-) recueillis à la surface. D'après Luck (2005, p30).

l'ensemble de la surface du crâne, comme c'est illustré dans la figure 6.1.C.

D'autres types d'activités électrophysiologiques participent sans doute de manière négligeable aux différences de potentiels enregistrées. Il s'agit, entre autres, des échanges ioniques transmembranaires générant les potentiels d'action, et des potentiels post-synaptiques ayant lieu dans des types de cellules nerveuses dans lesquelles les puits et les sources de courant ont une orientation aléatoire (cellules étoilées par exemple), ainsi que dans des structures où les cellules (pyramidales) ne partagent pas la même orientation.

6.2 ElectroEncéphaloGraphie (EEG)

6.2.1 Enregistrement

Toutes les expériences EEG étaient réalisées dans le cadre de la loi relative aux sujets se prêtant à la recherche biomédicale (autorisation RBM-0208). Les sujets participant aux expériences signaient un formulaire de consentement les informant du déroulement de l'expérience.

Pour toutes les expériences en EEG de surface, l'enregistrement des potentiels électriques était réalisé grâce à 35 électrodes Ag/AgCl disposées sur le cuir chevelu des sujets

selon le Système International 10/20 (voir la figure 6.2). Pour des raisons pratiques, nous avons utilisé un bonnet à électrodes (Easy cap) sur lequel l'emplacement des électrodes avait été préalablement déterminé à l'aide d'un système de pose informatisé (Echallier, Perrin & Pernier, 1992). Le contact entre l'électrode et le scalp était réalisé grâce à une pâte conductrice qui facilite la transmission du courant électrique. L'impédance des électrodes était vérifiée lors de la pose des électrodes et devait être inférieure à $5k\Omega$ pour chacune d'entre elles.

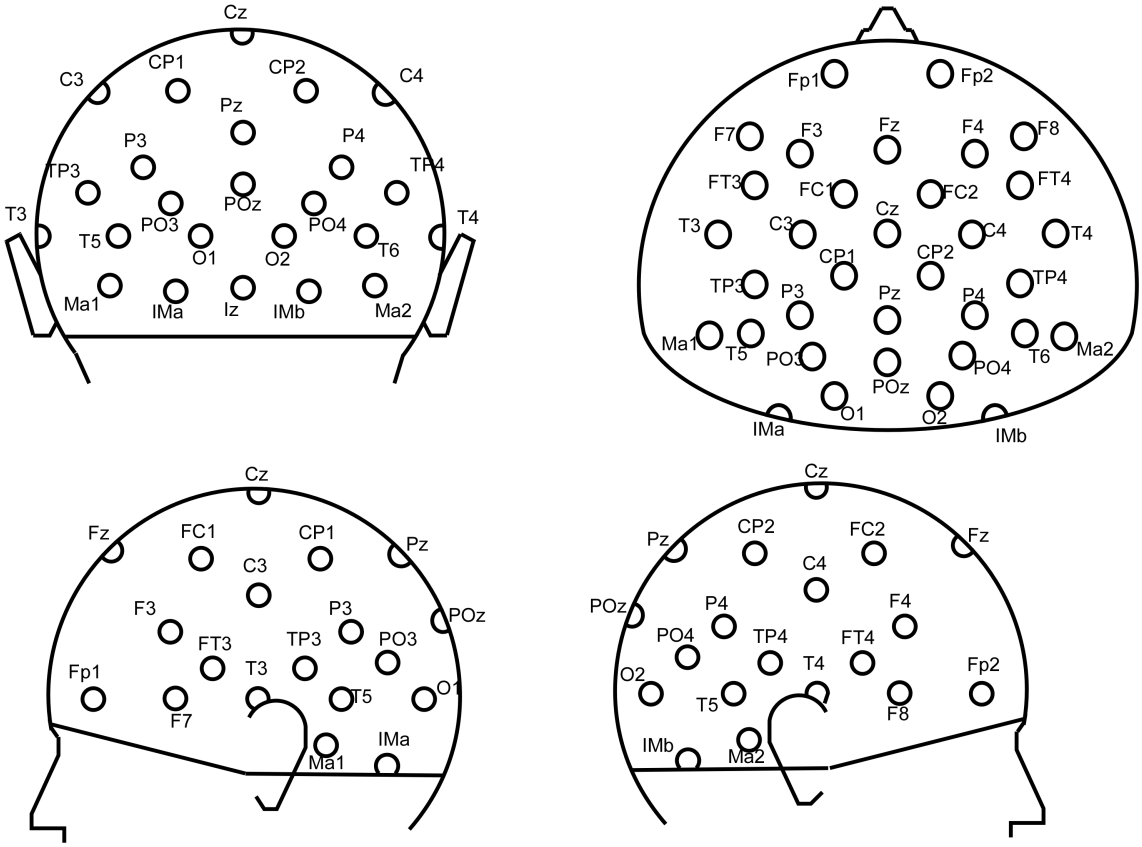

FIG. 6.2 – Électrodes utilisées pour l'enregistrement des potentiels évoqués de scalp. Dans le nom de l'électrode, la lettres indiquent son emplacement sur le scalp : F = Frontal, C = Central, T = temporal, P = Pariétal, O = Occipital, M = Mastoïde, I =Inion ; le chiffre indique l'hémiscalp : chiffre impair = côté gauche, chiffre pair = côté droite, z = ligne médiane.

L'EEG enregistré à chaque électrode est la variation dans le temps de la différence de potentiel entre chacune de ces électrodes (électrodes actives) et une même électrode de référence. La position de l'électrode de référence doit résulter d'un compromis entre un point suffisamment éloigné des sources actives pour être le plus neutre possible du point de vue de l'activité cérébrale, mais suffisamment proche pour éviter l'inclusion de potentiels parasites provenant du reste de l'organisme. Nous avons choisi de placer cette électrode de référence sur le nez. Les signaux étaient amplifiés dans des amplificateurs différentiels de marque Neuroscan Compumedics (64 voies). Afin de réduire le bruit électrique ambiant commun à l'électrode active et à l'électrode de référence, l'amplificateur différentiel amplifie

en réalité, d'une part, la différence de potentiel entre l'électrode active et une électrode de terre placée sur le front du sujet et, d'autre part, la différence de potentiel entre l'électrode de référence et cette terre. Le signal analogique, amplifié avec une bande passante de 0,1 à 200 Hz, était ensuite digitalisé à une fréquence d'échantillonnage de 1000 Hz (un échantillon par milliseconde). Tous les signaux étaient enregistrés en continu pendant les différents blocs expérimentaux. En outre, l'activité électro-oculaire était enregistrée entre une électrode posée près de canthus externe de l'œil droit et l'électrode de référence, afin de contrôler les mouvements oculaires horizontaux. Les mouvements oculaires verticaux étaient estimés dans les signaux des deux électrodes de scalp les plus frontales, Fp1 et Fp2.

Durant l'enregistrement, les sujets étaient confortablement assis dans un fauteuil, dans une pièce peu éclairée et isolée du bruit. Ils avaient pour consigne de se détendre afin de limiter toute activité myographique parasite. Le fauteuil était disposé de façon à ce que le sujet se trouve à 130 cm du moniteur par lequel étaient présentés les stimuli visuels. Les stimuli auditifs étaient (sauf mention contraire) présentés en champ libre au moyen de haut-parleurs situés à environ 1 m de part et d'autre de l'écran. Les consignes et les tâches propres à chaque expérience seront décrites en temps voulu.

Excepté dans les expériences sur l'effet d'indiçage temporel dans la perception de la parole et celle sur la MMN à la conjonction audiovisuelle, l'enchainement des stimuli visuels et sonores était contrôlé grâce au logiciel Vison, développé au laboratoire par Jean-François Echallier, Claude Delpuech et Pierre-Emmanuel Aguera. Ce logiciel fonctionne sous le système d'exploitation non graphique MS/DOS, ce qui permet de contrôler le temps de présentation à la milliseconde près. Les 2 expériences pré-citées ont été réalisées grâce au logiciel Presentation (Neurobehavioral Systems) fonctionnant sous Windows XP. Dans tous les cas, chaque évènement visuel, sonore, ainsi que chaque réponse du sujet, était associé à un code binaire envoyé par le logiciel de présentation des stimuli, de façon synchrone à la stimulation, au système d'acquisition des signaux EEG et permettait de marquer temporellement l'échantillon EEG ayant coïncidé avec cet évènement. Ce marquage permettait le calcul des potentiels évoqués et celui des temps de réaction.

De manière générale, les stimulations étaient présentées par séquences de 2 à 3 minutes, le sujet ayant la possibilité de se reposer entre chaque séquence et décidant lui-même du départ de la séquence suivante. Le temps total d'enregistrement utile ne dépassait pas 45 minutes et une pause était imposée au sujet à la moitié de l'enregistrement.

6.2.2 Analyse des potentiels évoqués (PE)

Toutes les analyses décrites dans cette partie ont été réalisées grâce au logiciel Elan, conçu au laboratoire par Olivier Bertrand et Pierre-Emmanuel Aguera.

Calcul du PE

Une façon d'étudier les processus cérébraux évoqués par une stimulation en EEG est d'estimer les variations de potentiel qui se reproduisent d'une présentation à l'autre du même stimulus dans une situation comparable et qui, a priori, reflètent le traitement de ce stimulus. Ces variations de potentiel évoquées par un stimulus, appelées potentiels évoqués (PE) ont en général une faible amplitude par rapport à l'activité EEG spontanée enregistrée

à tout instant sur le scalp, considérée en l'occurrence comme du bruit physiologique. Une technique simple pour isoler cette activité consiste à calculer la moyenne des variations de potentiel enregistrées suite à la présentation d'un grand nombre de stimuli identiques (entre 100 et 300). Pour chaque échantillon temporel t, on calcule donc la moyenne des potentiels enregistrés à cet échantillon t à travers l'ensemble des présentations du stimulus. On fait l'hypothèse que ce potentiel est la somme de potentiels invariables d'un essai à l'autre, correspondant à l'activité évoquée, et d'un potentiel dont la distribution sur l'ensemble des essais a une espérance égale à zéro, correspondant à l'activité physiologique spontanée ou non calée à la stimulation. La moyenne à un échantillon temporel donné va donc tendre vers la valeur du potentiel évoqué par la stimulation à cet échantillon, et ce d'autant plus que le nombre d'essais sera grand. Ce calcul de moyenne est réalisé à chaque échantillon temporel autour de l'évènement correspondant à l'envoi de la stimulation, sur une période s'étendant de 300 ms avant la stimulation à 600 ms après, dans nos expériences.

À ce stade, les PE peuvent contenir des potentiels non nuls avant la stimulation. Pour isoler les variations qui suivent la stimulation, on recentre les valeurs de potentiel autour de zéro dans une période précédant la stimulation (appelée ligne de base). Les PE obtenus par moyennage puis correction en ligne de base apparaissent comme une série de déflections de polarité positive ou négative. Leur polarité dépend de mécanismes excitateurs et inhibiteurs complexes ayant lieu au niveau synaptique et on ignore leur signification fonctionnelle. Le PE moyen (PEM) pour le groupe de sujets était calculé en faisant la moyenne des PE individuels à chaque électrode et à chaque échantillon temporel.

Notons que certaines activités reproductibles d'un essai à l'autre peuvent avoir lieu sans être exactement calées à la stimulation (en particulier les activités oscillatoires), auquel cas les variations de potentiel associées à ces activités ont tendance à s'annuler dans l'opération de moyennage. L'étude de ces variations de potentiel induites (par opposition aux activités évoquées) nécessite l'emploi de techniques d'analyse différentes et n'a pas été réalisée dans ce travail.

Artéfacts d'enregistrement

Si le moyennage sur quelques centaines d'essais permet d'annuler l'activité EEG spontanée, non calée à la stimulation, ce nombre peut s'avérer insuffisant pour éliminer des variations d'amplitude plus importantes provoquées par des clignements de paupière ou des mouvements des yeux. L'activité électro-oculographique associée à ces mouvements peut s'étendre sur une grande partie de l'EEG enregistrée sur la partie antérieure du scalp. Pour éviter que l'estimation des potentiels évoqués ne soit contaminée par de telles variations, les essais dans lesquels ces mouvements se produisaient ont été éliminés avant moyennage par une procédure de rejet automatique : tous les essais dans lesquels un échantillon avait une valeur de potentiel supérieure à $\pm 100\mu V$ dans la fenêtre d'analyse ont été éliminés du moyennage. De la même façon, une activité musculaire, en particulier au niveau du cou ou des tempes peut augmenter le niveau de bruit et compromettre le moyennage des potentiels évoqués. Lorsque cette activité musculaire était confinée à une ou deux électrodes, les valeurs de potentiel ont été remplacées par une interpolation des valeurs mesurées aux autres électrodes (grâce à des fonctions splines sphériques, voir la partie 6.2.2 page suivante). Lorsque le bruit musculaire s'étendait à un nombre supérieur de capteurs, les données du

sujet ont été exclues de l'analyse. Pour éliminer le bruit résiduel, les PE après moyennage étaient numériquement filtrés entre 1 et 30 Hz.

Cartes de potentiel

La distribution spatiale, ou topographie, des PE sur le scalp dépend bien sûr de la position et de l'orientation des générateurs intracérébraux activés en réponse au stimulus et permet donc, dans une certaine mesure, de localiser ces générateurs. Afin de visualiser cette distribution à un instant donné, la valeur du potentiel en tout point du scalp était interpolée à partir des valeurs réellement enregistrées aux électrodes. Ces valeurs étaient interpolées par des fonctions splines sphériques et les amplitudes étaient représentées sur une échelle de couleur (Perrin, Pernier, Bertrand & Echallier, 1989). L'utilisation de fonctions splines présente un double avantage : les extrémums de la distribution de potentiels ne sont pas nécessairement à l'emplacement d'une électrode et ces fonctions ont des dérivées spatiales continues, ce qui permet d'estimer la distribution d'une autre grandeur électrique appelée "densité radiale de courant sur le scalp". Les données interpolées étaient ensuite projetées radialement (conservation des distances entre les électrodes) sur une surface. Nous avons utilisé des vues gauche (projection centrée sur T3), droite (T4), et arrière (entre O1 et O2) comme indiqué sur la figure 6.3.

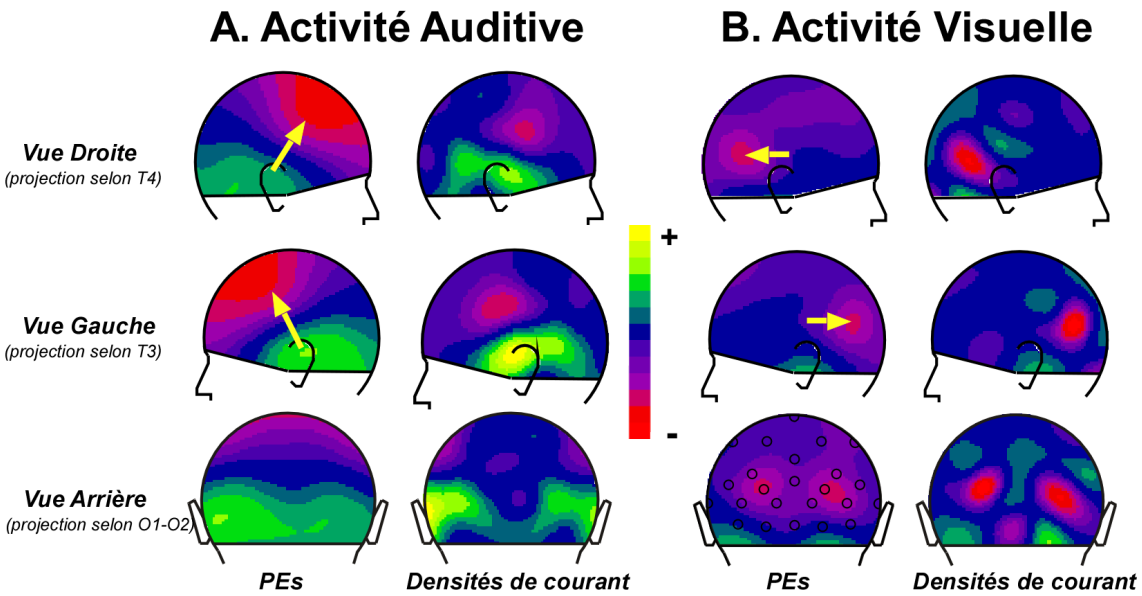

FIG. 6.3 – Topographie des potentiels évoqués et des densités radiales de courant typiquement évoquées par des stimulations auditives et visuelles. Les flèches jaunes indiquent la direction approximative des dipôles équivalents de courant. On peut remarquer sur la carte montrant l'emplacement des électrodes que les extrémums des fonctions splines ne se situent pas nécessairement sur des points de mesure.

La figure 6.3 donne également deux exemples de topographies de PE correspondant, l'une, à une activité auditive, et l'autre, à une activité visuelle. Ces topographies permettent d'illustrer la différence entre des générateurs à orientation plutôt tangentielle (réponse auditive) ou plutôt radiale (réponse visuelle). Comme on l'a dit plus haut, une composante

évoquée enregistrée sur le scalp correspond vraisemblablement à l'activation d'une population de neurones pyramidaux parallèles entre eux et perpendiculaires à la surface du cortex, équivalents à un dipôle de courant perpendiculaire au cortex. Or les circonvolutions du cortex font que ce dipôle peut avoir différentes orientations par rapport à la surface du crâne. Dans le cas de l'activité auditive (figure 6.3.A.), celle-ci est vraisemblablement due à une activité dans le cortex auditif dont l'orientation est perpendiculaire à la surface du crâne puisqu'il se trouve dans la scissure de Sylvius. L'orientation du dipôle de courant équivalent est donc parallèle (ou tangentielle) à la surface du scalp. Cette orientation particulière permet d'observer les potentiels positifs et négatifs correspondants. Dans le cas de l'activité visuelle (figure 6.3.B), celle-ci est sans doute générée par deux dipôles radiaux, c'est-à-dire perpendiculaires à la surface du scalp, dont on ne voit donc qu'un pôle, en l'occurrence le pôle négatif.

Les inférences sur la localisation des générateurs correspondant aux activités sensorielles auditives et visuelles, à partir des distributions de potentiel, sont basées sur les connaissances acquises durant quelques dizaines d'années de recherche sur l'électrophysiologie sensorielle. Lorsqu'on est confronté à une activité pour laquelle on n'a pas d'hypothèses fortes, il est beaucoup plus difficile de faire des inférences précises uniquement à partir des cartes de potentiel, pour plusieurs raisons : d'abord la distribution des potentiels à un instant donné reflète en général l'activité de plusieurs générateurs simultanés. Par ailleurs, la distribution de potentiels créée par chaque générateur est très étalée sur le scalp, en raison des différences de conductivité des tissus traversés. Si bien que la distribution des potentiels enregistrés à un instant donné correspond à la somme algébrique de plusieurs distributions de potentiels, non nuls sur une grande partie du scalp. Comme on n'a, en général, pas d'hypothèse précise sur ces différents générateurs, il est impossible de séparer de façon unique les sources des différentes activités. Pour faciliter la localisation visuelle des générateurs intracérébraux, il est toutefois possible de calculer la distribution d'une autre grandeur électrique sur le scalp : la densité radiale de courant.

Cartes de densité radiale de courant

La densité de courant radial en un point du scalp peut se définir comme la quantité de courant par unité de volume ayant traversé, radialement à la surface, les différents milieux conducteurs jusqu'au scalp. Les cartes de densité de courant représentent les zones du scalp d'où émergent les lignes de courant (sources de courants) et celles où les lignes de courant retournent vers le cerveau (puits de courants). Elles sont estimées à partir de la dérivée spatiale seconde des fonctions splines utilisées dans l'interpolation des champs de potentiel et sont exprimées en mA/m^3 (Perrin, Bertrand & Pernier, 1987 ; Perrin et coll., 1989). Il s'agit d'une grandeur locale, indépendante de tout modèle ou hypothèse sur les générateurs impliqués.

Les densités de courant radial ont une topographie moins diffuse que celle des potentiels, et leurs extrémums sont moins étalés que les pôles positifs et négatifs des cartes de potentiel (cette différence est illustrée dans la figure 6.3 page ci-contre, par la comparaison entre les distributions de potentiel et de densité radiale de courant correspondant à l'activité auditive ou visuelle). Les distributions de densité radiale de courant offrent ainsi l'avantage de pouvoir dissocier des "composantes" (activité d'un ensemble de neurones) qui seraient

superposées dans les cartes de champs de potentiel. Elles sont, d'autre part, indépendantes de la position de l'électrode de référence. Enfin, l'amplitude des champs de courant s'atténue plus rapidement que celle des potentiels quand le (ou les) générateurs sont situés plus en profondeur (Perrin et coll., 1987) : les cartes de densité de courant reflètent donc l'activité de générateurs corticaux relativement proches de la surface et sont aveugles aux sources profondes. L'analyse conjointe des distributions de potentiel et de densité de courant pourra donc apporter des éléments qualitatifs importants sur l'orientation et la profondeur des générateurs intracérébraux.

Notons que lors du moyennage de plusieurs sujets, il existe une certaine invariance de la position des générateurs par rapport aux électrodes. Ceci vient du fait que les différentes électrodes sont placées par rapport à des repères anatomiques propres à chaque sujet, ce qui induit une normalisation spatiale approximative et implicite des distributions de potentiel ou de densité radiale de courant. Comme nous le verrons plus loin, ce n'est pas le cas pour la MEG.

6.3 MagnétoEncéphaloGraphie (MEG)

6.3.1 Champs magnétiques cérébraux

Une autre façon d'améliorer la localisation des générateurs électriques cérébraux est d'en enregistrer un autre aspect, à savoir les champs magnétiques qu'ils engendrent : un dipôle électrique génère en effet un champ magnétique tournant autour de son axe, tel qu'illustré dans la figure 6.4.A page suivante). Lorsqu'une population de neurones corticaux équivalente à un dipôle tangentiel est activée, des champs magnétiques extrêmement faibles entrent et sortent de la tête (figure 6.4.B). Le crâne provoque très peu de perturbation sur ces champs magnétiques ce qui permet une précision spatiale meilleure qu'en EEG. Un inconvénient de ces signaux par rapport à l'EEG est qu'un dipôle radial ne génère pas de champs magnétique enregistrable à l'extérieur du crâne et qu'on n'enregistre donc que des populations de neurones pyramidaux plutôt parallèles aux capteurs.

La variation des champs magnétiques au cours du temps peut être enregistrée avec des capteurs très sensibles appelés SQUID (*Superconducting Quantum Interference Device*), au fonctionnement complexe (voir Pernier & Bertrand, 1997, pour une introduction) et qui nécessitent des températures très basses pour leur fonctionnement. Les capteurs SQUID sont donc baignés dans de l'hélium liquide à 4,2°K et sont de ce fait disposés de façon rigide, en formant un casque dans lequel le sujet place sa tête.

De la même façon que l'on calcule les potentiels évoqués, on peut calculer les champs magnétiques évoqués (CME) par une stimulation, en utilisant les mêmes méthodes de moyennage et de traitement du signal. Grâce au nombre important de capteurs, on peut également représenter la distribution des CME sur des projections bidimensionnelles, à l'exception près qu'une carte représente ici la distribution des champs magnétiques au niveau des capteurs, donc au niveau du casque rigide, et non au niveau du scalp des sujets. Cela a pour conséquence que le moyennage de cartes de plusieurs sujets ajoute une variabilité due au fait que tous les sujets n'ont pas une tête de la même taille et qu'ils peuvent l'orienter

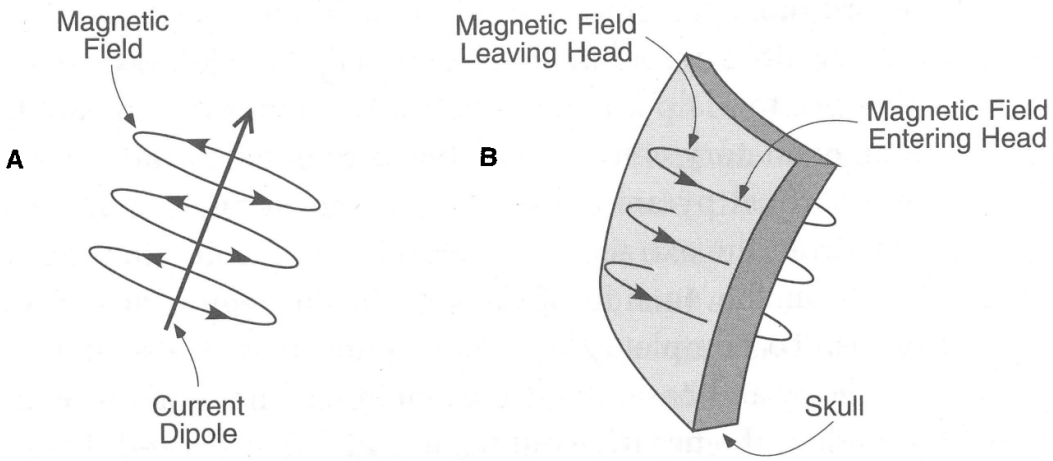

FIG. 6.4 – A. Champ magnétique généré par un dipôle de courant. B. Orientation du champ magnétique créé par un dipôle de courant tangentiel situé derrière le scalp. D'après Luck (2005).

différemment dans le casque MEG. Pour l'interprétation des topographies des CME, il est important de noter que les activités générées en MEG par un dipôle tangentiel montrent une inversion de polarité perpendiculaire à celle de l'activité générée en EEG, comme on peut le constater en comparant les activités auditives MEG de la figure 16.1.A page 209 et les activités auditives EEG de la figure 6.3.A page 88.

6.3.2 Procédure d'enregistrement

L'expérience MEG a été réalisée dans le carde de la loi relative aux sujets se prêtant à la recherche biomédicale (autorisation 2005-091/A). Les sujets participant aux expériences signaient un formulaire de consentement les informant du déroulement de l'expérience.

Les enregistrements MEG ont eu lieu au centre MEG de Lyon, sur un système de marque CTF, 275 capteurs, situé dans une chambre blindée en mu-métal de façon à éviter tout perturbation du champ magnétique enregistré. Le sujet était confortablement assis dans un fauteuil attenant au système MEG, la tête au fond du casque, tout en préservant un champ de vision suffisant pour contenir l'écran de stimulation. Tout objet susceptible de générer un champ magnétique étant banni de la chambre blindée, les stimulations visuelles étaient projetées de l'extérieur, sur un écran translucide placé en face du sujet. La taille des stimuli était calculée de façon à correspondre au même angle visuel que dans les expériences d'EEG. Les stimulations acoustiques étaient générées par un transducteur piézo-électrique créant une vibration transmise par des tubes plastiques souples aux oreilles du sujet. Ce mode de stimulation acoustique était donc différent de la stimulation en champ libre, utilisée en EEG. Les enregistrements ont été effectués de façon continue, avec une fréquence d'échantillonnage de 600 Hz.

6.4 StéréoElectroEncéphaloGraphie (sEEG)

Il existe un moyen d'accéder directement à l'activité électrique cérébrale, mais il n'est utilisé que dans un cadre thérapeutique, celui du traitement de l'épilepsie. Bien qu'il existe des traitements pharmacologiques de l'épilepsie, certains patients sont résistants à ces traitements et l'unique façon d'atténuer ou de supprimer les symptômes est de recourir à la neurochirurgie. La résection de certaines structures corticales ou sous-corticales à l'origine des crises épileptiques nécessite d'identifier le plus précisément possible la provenance de ces crises. La sEEG est l'un des examens destinés à affiner le diagnostic pré-chirurgical. Elle consiste à implanter directement dans le cerveau des patients des électrodes multicontacts, afin d'y enregistrer la variation des champs de potentiel locaux au cours du temps et d'étudier la propagation des potentiels pathologique lors des crises. Les patients sont implantés pour une période de deux semaines environ, de façon à pouvoir enregistrer l'activité EEG intracérébrale pendant au moins une crise. En collaboration avec le docteur Catherine Fischer, responsable du Service d'Exploration Fonctionnelle de l'hôpital neurologique et neurochirurgical Pierre Wertheimer, et Olivier Bertrand (U821) nous avons pu soumettre certains de ces patients à l'un de nos protocoles expérimentaux et enregistrer les potentiels intracérébraux évoqués par des stimulations auditives et/ou visuelles.

6.4.1 Localisation des électrodes

Les électrodes multicontacts présentent 5, 10 ou 15 contacts de 2 mm de longueur, alignés et espacés de 3,5 mm de centre à centre. Les électrodes sont insérées de manière orthogonale au plan sagittal dans le repère stéréotaxique (Talairach & Szikla, 1967), jusqu'à atteindre les structures sous-corticales profondes. Les contacts des électrodes sondent donc aussi bien les aires corticales latérales que médianes, ainsi que les scissures et sillons. Un certain nombre de contacts se trouvent dans la matière blanche et dans des noyaux sous-corticaux.

La localisation précise des électrodes était réalisée a posteriori sur la base de l'IRM anatomique du patient réalisée avant l'implantation des électrodes, et de deux clichés radiographiques montrant la position des électrodes par rapport au crâne, l'un selon une vue sagittale et l'autre selon une vue coronale, dans le repère stéréotaxique utilisé par le chirurgien pour l'insertion des électrodes. Le repère de Talairach du sujet est défini par le plan médian séparant les deux hémisphères cérébraux et un plan orthogonal passant par la ligne reliant les commissures antérieure (AC) et postérieure (PC) (voir la figure 6.5 page 96). Ces points de repères étaient définis visuellement sur l'IRM anatomique et les axes du repère de Talairach étaient ensuite reportés sur le cliché radiographique sagittal par comparaison avec la coupe IRM sagittale médiane, ce qui permettait de relever sur les deux clichés les coordonnées tridimensionnelles des contacts dans le repères de Talairach du sujet[1]. Les coordonnées des contacts étaient ensuite converties dans le système de coordonnées des images IRM anatomiques pour identifier précisément les structures tra-

[1] Cette étape suppose que le plan sagittal stéréotaxique est confondu avec le plan sagittal dans le repère de Talairach du sujet, ce qui n'était pas toujours le cas : un estimation de l'angle de déviation de ces deux plans pouvait être faite grâce à la comparaison d'une coupe coronale de l'IRM et du cliché coronal. Lorsque cet angle était trop grand, il a été pris en compte dans le calcul des coordonnées.

versées par les électrodes. Cette procédure a une précision de l'ordre de 2 mm, comme on a pu le constater pour un patient dont on pouvait voir les traces des électrodes sur des images IRM anatomiques réalisées après la désimplantation. Toutes les manipulations sur les images IRM ont été réalisées grâce au logiciel Activis développé par Marc Thévenet (U280 et Institut des Sciences Cognitives), Claude Delpuech et Pierre-Emmanuel Aguera (Unité 821).

La position des électrodes pouvait être visualisée sur des représentations en trois dimensions de parties isolées de cortex, afin de faciliter l'identification des structures enregistrées, en particulier pour le cortex auditif enfoui dans la scissure de Sylvius, et la comparaison entre patients. La segmentation du cortex était réalisée avec le logiciel Freesurfer et les représentations tridimensionnelles du cortex et des électrodes étaient visualisées grâce à un programme Matlab écrit par Françoise Bauchet (Centre MEG) et Olivier Bertrand (Unité 821).

6.4.2 Procédure d'enregistrement

En raison du faible nombre de patients traités disponibles pour ce genre d'étude, les enregistrements ont été réalisés sur une période de 2 ans. Pour les 5 premiers patients, le matériel utilisé pour l'enregistrement sEEG était le même que celui utilisé au laboratoire pour l'enregistrement de l'EEG chez les sujets sains. Les patients étaient testés dans une pièce isolée de l'hôpital, dans des conditions très similaires aux conditions d'enregistrement des sujets sains au laboratoire, si ce n'est que les stimulations sonores étaient présentées au moyen d'un casque audio à oreillettes. Les 5 patients suivants ont été testés assis dans leur lit d'hôpital, le signal sEEG étant enregistré grâce à des amplificateurs de marque Micromed (128 voies), à une fréquence d'échantillonnage de 512 Hz. Les patients avaient entre 9 et 15 électrodes implantées, pour un nombre maximum de 225 contacts. En raison du nombre limité de canaux d'amplification, nous avons dû choisir 64 (ou 128) de ces contacts, sur la base des informations notées par le chirurgien et du site d'implantation des électrodes, c'est-à-dire des structures cérébrales explorées. La plupart des patients présentant une épilepsie d'origine temporale, les électrodes étaient souvent situées dans le lobe temporal et nous avons surtout ciblé nos enregistrements sur les aires supérieures du cortex temporal (cortex auditif, STS...).

Comme pour l'EEG, on enregistre une différence de potentiel entre un contact actif et un contact de référence. Les variations de potentiel à tous les contacts actifs ont été enregistrées avec une référence intracérébrale unique (montage monopolaire). Nous avons choisi une référence intracérébrale afin d'équilibrer l'impédance entre l'électrode active et l'électrode de référence. Une électrode de scalp posée sur le front du patient était reliée à la terre. L'inconvénient de choisir une référence intracérébrale est que le contact choisi peut a priori présenter une activité évoquée par les stimulations. Dans ce cas, on court le risque d'interpréter une activité au niveau de la référence comme une activité au niveau du contact actif. Pour minimiser ce risque, on choisissait comme référence un contact situé dans la matière blanche, à une position la plus éloignée possible des 64 ou 128 contacts choisis pour l'enregistrement.

6.4.3 Calcul du PE et rejet d'artéfacts

Le signal sEEG était filtré numériquement entre 0,2 et 100 Hz, avec une encoche (*notch*) à 50 Hz de façon à exclure les interférences électromagnétiques provenant du réseau électrique. Le fait de garder les hautes fréquences entre 30 et 100 Hz permet d'observer les réponses précoces dans le cortex auditif primaire qui ont un décours temporel plus rapide. Le calcul des PE suit le même principe qu'en EEG de surface. Les artéfacts d'enregistrements sont cependant de nature différente. Puisque l'on enregistre directement l'activité intracérébrale, celle-ci n'est pas contaminée par les mouvements des yeux ou l'activité musculaire. De plus, le rapport signal sur bruit est bien meilleur qu'en EEG et on peut facilement observer les réponses évoquées les plus amples sur un essai élémentaire. Toutefois l'activité cérébrale enregistrée peut présenter certains aspects pathologiques, même en dehors des crises. L'activité sEEG enregistrée chez les patients épileptiques présente en général des pointes intercritiques qui sont de grandes déflections d'amplitude bien supérieure à l'amplitude des potentiels évoqués intracérébraux.

Pour éviter l'inclusion de ces pointes dans le calcul du potentiel évoqué, nous avons utilisé une procédure de rejet automatique, proposée par Jean-Philippe Lachaux (U821). Pour chaque échantillon temporel dans la fenêtre d'analyse, on a calculé son écart-type sur l'ensemble des essais correspondant à la même stimulation (intervenant dans le calcul du potentiel évoqué). Tout essai dans lequel au moins un échantillon temporel sur au moins un contact déviait du potentiel évoqué de plus de 5 écart-types était exclu du moyennage, ce qui permet une exclusion des essais contaminés par les pointes. Cette procédure a été appliquée pour tous les types de stimulation. Pour éviter que le nombre d'essais ainsi rejetés ne soit trop important en raison de certains contacts présentant un nombre élevé de pointes intercritiques, tout contact participant au rejet de plus de 6% des essais était exclu de l'analyse des potentiels évoqués.

Cette procédure semi-automatique pouvait être adaptée manuellement de façon à conserver certains contacts intéressants participant au rejet de plus de 6% des essais, ce qui se traduisait par un pourcentage d'essais rejetés plus important. Lors de cette procédure, un compromis était donc constamment réalisé entre la conservation du plus grand nombre d'essais possible et celle du plus grand nombre de contacts possible, tout en garantissant l'exclusion des essais contenant des pointes intercritiques. Cette procédure était réalisée grâce à un programme Matlab développé par J.P. Lachaux et adapté par mes soins.

6.4.4 Résolution spatiale et représentation spatiotemporelle

Contrairement à l'EEG ou à la MEG, la sEEG bénéficie d'une excellente résolution spatiale puisque l'activité cérébrale électrique peut être enregistrée directement à sa source. Cependant, en montage monopolaire, la différence de potentiel reflète a priori la somme algébrique de tous les courants générés dans l'encéphale. L'atténuation de ces courants avec la distance fait que le potentiel est dominé par les courants générés à proximité du contact (à condition que le contact de référence ne présente pas de variation notable de son activité calée à la stimulation) et certains auteurs estiment que les signaux enregistrés en montage monopolaire représentent majoritairement des courant générés à une distance maximale de 1 à 2 cm (Lachaux, Rudrauf & Kahane, 2003).

Cette spécificité spatiale est encore améliorée si l'on calcule la différence de potentiel entre deux contacts successifs sur une électrode (montage bipolaire) car les différences de potentiel s'atténuent alors encore plus rapidement avec la distance par rapport à la source. Le montage bipolaire présente en outre l'avantage d'être indépendant de la référence choisie pour l'enregistrement. L'inconvénient des signaux bipolaires est qu'ils sont aveugles aux courants qui affectent de la même façon les potentiels aux deux contacts du bipôle. Les potentiels évoqués en montages bipolaire et monopolaire donnent donc des informations complémentaires sur la localisation des sources enregistrées. La contrepartie de cette bonne résolution spatiale locale est la couverture spatiale du cerveau qui est limitée à une dizaine d'électrodes multicontacts, implantées chez chaque patient en fonction de considérations thérapeutiques uniquement.

Comme les variations de potentiel au cours du temps étaient enregistrées à différents contacts d'une même électrode, on avait également accès à la variation du potentiel dans l'espace, le long de l'axe de l'électrode, ce qui a permis d'observer les profils spatiaux des potentiels au cours du temps. Ce profil spatiotemporel était approximé par interpolation bilinéaire des quatre points les plus proches (dans les dimensions de temps et d'espace). Les profils spatiaux permettent de mieux caractériser la source des potentiels observés. Ainsi, une inversion focale de potentiel électrique monopolaire sur deux contacts voisins signifie que ces contacts se trouvent de part et d'autre du plan orthogonal à la source de courant. Plus cette inversion est focale, plus on peut en déduire qu'ils sont proches de la source. L'observation des profils spatiotemporels des potentiels bipolaires permet également de mieux apprécier la proximité de la source. En particulier une inversion de polarité entre deux contacts montre une variation très locale du gradient de potentiel, ce qui peut indiquer que la source est très proche (ou un changement de conductivité du milieu)

6.4.5 Étude de groupe et normalisation anatomique

La comparaison des résultats de différents patients était uniquement qualitative et n'a pas fait l'objet de tests statistiques (en effet, les implantations des différents patients ne sont pas comparables). Pour réaliser cette comparaison, il est cependant nécessaire de rapprocher les résultats des analyses individuelles dans des structures cérébrales comparables d'une patient à l'autre. Une première solution consiste à réaliser des rapprochements sur la base de l'identification individuelle des structures cérébrales explorées chez chaque patient.

Dans certains cas cependant, il peut être intéressant d'avoir une vue d'ensemble des résultats dans un repère commun à tous les patients. Pour cela, il est nécessaire de normaliser les coordonnées des électrodes de différents patients. Nous avons utilisé la méthode transformation linéaire par cadrans employée par Talairach et Tournoux (1988). Cette méthode consiste à définir une boite entourant le cerveau et tangentielle à celui-ci (voir la figure 6.5 page suivante). Cette boite est subdivisée en 12 cadrans (6 pour chaque hémisphère) dont les limites sont définies par les plans horizontal et sagittal joignant les commissures antérieures et postérieures ainsi que par les deux plans perpendiculaires passant respectivement par AC et PC.

Normaliser deux cerveaux consiste à identifier chaque cadran du premier cerveau avec le cadran analogue de l'autre cerveau. Pour convertir les coordonnées d'un point situé dans le premier cerveau dans le système de coordonnées du second cerveau, on réalise

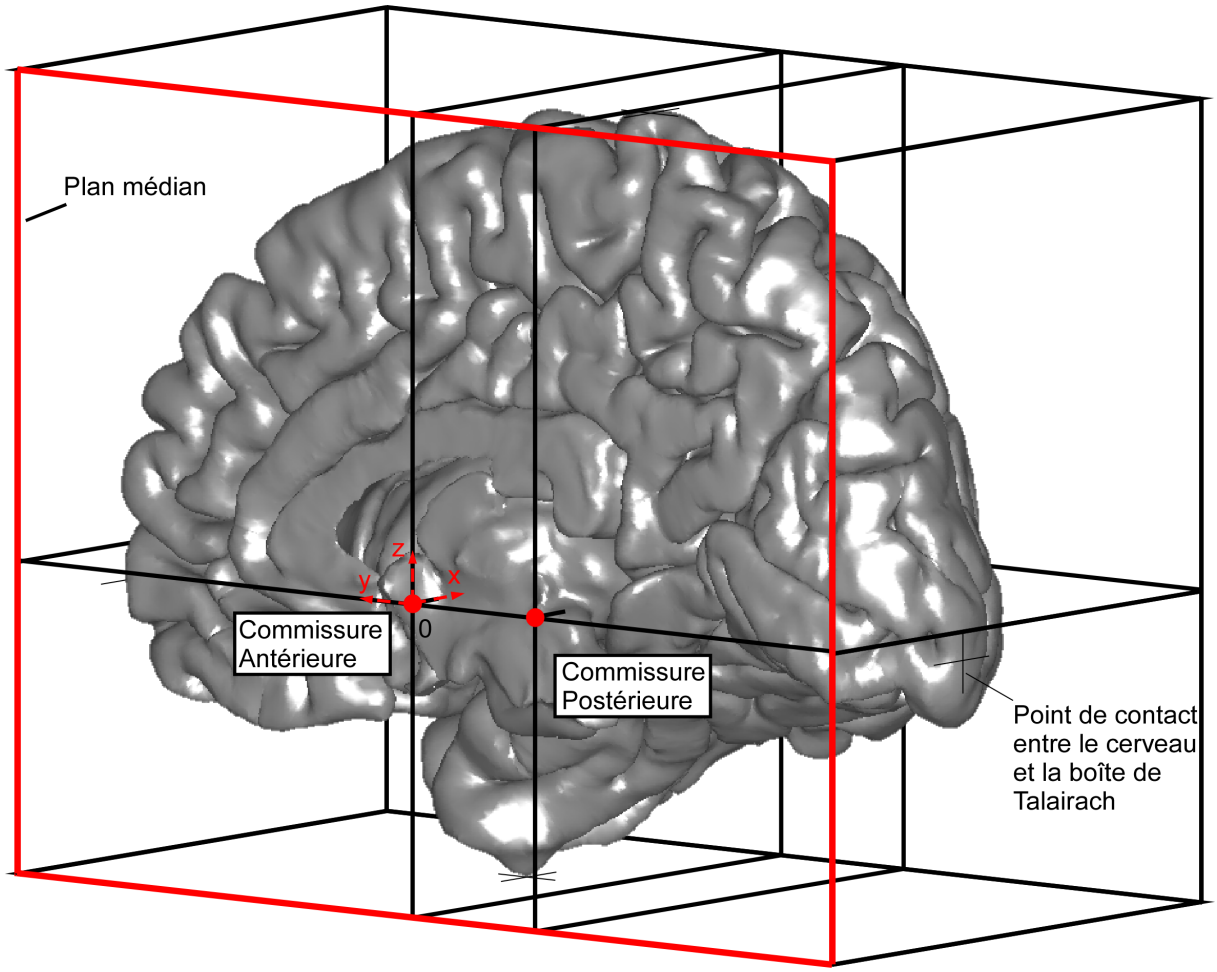

FIG. 6.5 – Boite de Talairach : les commissures antérieure et postérieure et le plan sagittal passant par ces deux points (plan médian) sont définis visuellement à partir de l'IRM anatomique du patient. Ils définissent une boite dont les dimensions correspondent aux points les plus extrêmes du cortex. Cette boite est subdivisée en 12 cadrans (6 par hémisphère). Le repère de Talairach est indiqué par les axes x, y et z.

une transformation linéaire des coordonnées tridimensionnelles du point différente selon le cadran : L'origine de la transformation linéaire est AC pour les 4 cadrans antérieurs et PC pour les 4 cadrans postérieurs. Le coefficient appliqué pour la transformation de chacune des coordonnées est égal au rapport des dimensions des cadrans des deux cerveaux. Pour les 4 cadrans situés entre les deux commissures, l'origine de la transformation linéaire peut être indifféremment la commissure antérieure ou postérieure (le résultat est identique).

On peut ainsi normaliser les coordonnées des électrodes de tous les patients vers un repère arbitraire, qui est traditionnellement celui correspondant au cerveau décrit dans l'atlas de Talairach et Tournoux (1988). Les coordonnées obtenues dans ce cas sont ce que l'on appelle les coordonnées de Talairach.

Pour la représentation visuelle des coordonnées normalisées, nous avons choisi de reporter les coordonnées des électrodes sur l'IRM anatomique du cerveau fourni par l'Institut de Neurologie de Montréal (MNI), qui est souvent utilisé comme cerveau "standard" dans

les études IRMf. Pour ce faire, les axes du repère de Talairach ont été repérés visuellement sur cette IRM anatomique, de la même manière que pour les patients. Les coordonnées des électrodes des 10 patients ont été transformées suivant la méthode décrite ci-dessus vers le repère du cerveau du MNI. Nous avons également segmenté le ruban cortical de ce cerveau afin d'en reconstruire une représentation tridimensionnelle.

Cette méthode de normalisation anatomique globale comporte des inconvénients qu'il convient de garder à l'esprit lors de l'interprétation de telles figures. En effet, en raison de la variabilité importante de l'anatomie sulco-gyrale, la situation d'une structure cérébrale donnée par rapport aux commissures antérieure et postérieure peut varier d'un individu à l'autre de façon relativement importante, même en tenant compte des facteurs d'échelle. La normalisation introduit donc une incertitude qui peut aisément faire passer un point d'un côté à l'autre d'un sillon et fausser l'interprétation des résultats. Pour la localisation anatomique des électrodes proprement dite, on s'en tient donc à l'analyse individuelle de l'IRM anatomique du patient.

Chapitre 7

Méthodes propres à l'étude de l'intégration audiovisuelle

7.1 Falsification de l'inégalité de Miller

À plusieurs reprises, nous avons couplé nos mesures de l'activité cérébrale à des mesures de temps de réaction, qu'il s'agisse de détection ou de discrimination. Un gain comportemental (TR plus court pour traiter le stimulus bimodal que le même stimulus présenté dans chaque modalité séparément) permettait de s'assurer que les traitements unisensoriels avaient bien interagi en condition bimodale. Nous avons choisi cette mesure comportementale car, contrairement à la performance, la rapidité de traitement n'atteint pas de plafond en l'absence de bruit, c'est-à-dire dans des conditions similaires à celles de l'enregistrement des potentiels évoqués.

Nous avons vu (voir la partie 2.3.3 page 36) que la simple présence d'un gain en temps de réaction dans la condition bimodale par rapport à l'une et l'autre des conditions unimodales n'est pas une preuve suffisante de l'existence de processus d'intégration, car ce gain peut s'expliquer, dans un modèle simple de convergence tardive des voies sensorielles, par un phénomène de facilitation statistique. Pour mettre en évidence l'existence de réelles interactions audiovisuelles, nous avons choisi le critère proposé par J. O. Miller (1982), basé sur la comparaison des distributions de TR dans les conditions auditive, visuelle et audiovisuelle et qui permet de rejeter un modèle d'activations séparées. Bien qu'il soit communément accepté que la violation de l'inégalité de Miller révèle de véritables interactions audiovisuelles, il est important de comprendre les détails mathématiques et différents postulats nécessaires à son application, qui, d'une certaine manière, limitent l'interprétation de ce critère.

7.1.1 Bases mathématiques et postulats

Dans le modèle d'activations séparées proposé par J. O. Miller (1982), les deux canaux sensoriels auditif et visuel sont parallèles et convergent vers des processus communs. Ce modèle simple lui permet de faire des prédictions sur la distribution des TR bimodaux à partir des distributions de TR unimodaux en faisant un nombre limité d'hypothèses. Dans

ce modèle, on part du principe que le TR pour un essai bimodal sera déterminé par le premier des traitements unisensoriels déclenchant les processus communs liés à la réponse, comme dans le modèle d'indépendance de Raab (1962). Dans un essai bimodal donné, le temps de traitement (TT) à l'instant de déclenchement des processus communs est donc le plus petit des deux TT auditif ou visuel. L'ensemble des essais bimodaux correspondent à une distribution bivariée des TT auditifs et visuels, c'est-à-dire une distribution de couples (TT_A, TT_V). À chaque essai audiovisuel, c'est le plus petit des deux TT qui définit le TT audiovisuel, donc la distribution des TT bimodaux (résultant de la compétition) est égale à la distribution des minima de cette distribution bivariée.

Pour savoir si les données expérimentales sont explicables par le modèle, il faut donc pouvoir estimer indépendamment la distribution des $min(TT_A, TT_V)$ et la distribution de temps de traitement audiovisuels TT_{AV} à partir de données observables.

Pour ce faire, deux hypothèses doivent être faites :

1. le temps pris par les processus communs est constant dans tous les essais et quelque soit le signal (auditif ou visuel) qui les déclenche. Ceci permet d'estimer la distribution des TT audiovisuels à partir des TR audiovisuels. Cette hypothèse n'oblige cependant pas à se prononcer sur les niveaux de traitement inclus dans ce temps fixe (décision, programmation motrice, exécution motrice).
2. puisqu'on n'a pas accès aux TT unimodaux en condition bimodale, on doit les estimer à partir des conditions unimodales. Il faut alors supposer que la distribution des temps de traitement ne dépend pas du contexte unimodal ou bimodal de présentation. Ce postulat est appelé postulat d'indépendance au contexte. Il n'est pas formulé explicitement par J. O. Miller (1982) mais sera rendu explicite par plusieurs auteurs par la suite (Colonius, 1990 ; Townsend, 1997). En termes statistiques, ce postulat implique que les distributions marginales de la distribution bivariée des TT audiovisuels soient égales aux distributions des TT auditif et visuel en conditions unimodales. Bien entendu, c'est l'hypothèse d'invariance du temps des processus communs qui permet d'estimer les distributions des TT auditifs et visuels à partir des distributions des TR auditifs et visuels.

Ces suppositions faites, on peut donc prédire que, dans le modèle d'activations séparées, la distribution des TR bimodaux, qui est observable, est égale à la distribution des minima de la distribution bivariée dont les distributions marginales sont les distributions des TR unimodaux, qui sont toutes deux également observables.

Une façon d'appliquer ce modèle est de partir des moyennes des TR unimodaux, de postuler la normalité et l'égalité des variances de leurs distributions et d'en déduire la distribution des minima (en postulant au passage l'indépendance des distributions des TR auditifs et visuels, voir plus loin) et donc leur moyenne. Cette moyenne peut être alors être directement comparée au TR moyen obtenu en condition bimodale pour rejeter ou accepter le modèle d'activations séparées. C'est la méthode retenue par Raab (1962). Afin de se passer de l'hypothèse de normalité, J. O. Miller (1982) (ainsi que Gielen et coll., 1983) cherchent au contraire à estimer cette distribution de minima à partir des distributions effectives des TR unimodaux. Pour cela, il est commode d'utiliser les fonctions de répartition des TR ou des TT.

Soient $p(TR_A < t)$, la fonction de répartition des TR auditifs et $p(TR_V < t)$, la fonction de répartition des TR visuels. Pour un t donné, $p(TR_A < t)$ désigne donc la probabilité qu'un TR auditif soit inférieur à une certaine valeur t, et $p(TR_V < t)$ la probabilité qu'un TR visuel soit inférieur à t. De même, $p[min(TT_A, TT_V) < t]$ désigne la fonction de répartition des minima de la distribution bivariée des TT unimodaux en condition bimodale. Dans la condition audiovisuelle, le TT unimodal minimum sera inférieur à une valeur t si le TT auditif est inférieur à t ou si le TT visuel est inférieur à t, ou encore si les deux temps de traitements sont inférieurs à t, ce qui s'écrit :

$$p[min(TT_A, TT_V) < t] = p(TT_A < t \cup TT_V < t), \forall t$$

cette prédiction s'étend aux TR, en vertu des hypothèses posées précédemment, donc :

$$p[min(TR_A, TR_V) < t] = p(TR_A < t \cup TR_V < t), \forall t$$

or les propriétés élémentaires des probabilités indiquent que

$$p(A \cup B) = p(A) + p(B) - p(A \cap B)$$

donc :

$$p[min(TR_A, TR_V) < t] = p(TR_A < t) + p(TR_V < t) - p(TR_A < t \cap TR_V < t), \forall t$$

Les deux premiers termes $p(TR_A < t)$ et $p(TR_V < t)$ peuvent être estimés, mais pas le dernier. Pour le connaitre, il faudrait pouvoir accéder à la distribution bivariée des TT auditifs et visuels en condition audiovisuelle. Or cette dernière n'est pas observable ; autrement dit, on n'a aucun moyen de savoir comment se combinent les TT auditifs et visuels sur l'ensemble des essais audiovisuels.

En fait, on peut définir une infinité de modèles d'activations séparées selon le degré de corrélation des distributions des temps de traitement unimodaux pour une essai bimodal. Ainsi, il se peut qu'un TT rapide pour un stimulus auditif soit plus souvent associé à un TT rapide du stimulus visuel (corrélation positive) ou plus souvent associé à un traitement lent du stimulus visuel (corrélation négative), ou que toutes les associations soient également probables (indépendance). Certains auteurs ont postulé, à la suite de Raab (1962), une indépendance des distributions des temps de traitement unimodaux[1] (Gielen et coll., 1983). Dans ce cas, on a :

$$p(TR_A < t \cap TR_V < t) = p(TR_A < t) \times p(TR_V < t), \forall t$$

et donc

$$p[min(TR_A, TR_V) < t] = p(TR_A < t) + p(TR_V < t) - p(TR_A < t)p(TR_V < t), \forall t$$

[1]Cette hypothèse d'indépendance ne doit pas être confondue avec le postulat d'indépendance au contexte qui renvoie au fait que les distributions marginales de la distribution bivariée des temps de traitement unimodaux en condition bimodale sont considérées comme identiques au distributions des temps de traitement unimodaux dans les conditions unimodales.

Tous les termes étant observables, on peut calculer $p[min(TR_A, TR_V) < t]$ et comparer la fonction de répartition obtenue à la fonction de répartition $p(TR_{AV} < t)$ obtenue à partir de la distribution effective des TR bimodaux. Selon ces auteurs le modèle doit être rejeté s'il existe au moins une valeur t pour laquelle

$$p(TR_{AV} < t) > p[min(TR_A, TR_V) < t]$$

En effet, dans ce cas, les TR bimodaux sont inférieurs à ceux prédits par le modèle (graphiquement, cela correspond au cas où la fonction de répartition des TR audiovisuels passe au dessus de la fonction de répartition des minima de la distribution bivariée). En fait, strictement parlant, le modèle devrait être rejeté si l'égalité n'est pas respectée, que ce soit dans un sens ou un autre. Plusieurs auteurs ont utilisé ce critère (par exemple : Laurienti, Kraft, Maldjian, Burdette & Wallace, 2004 ; Molholm et coll., 2002 ; Senkowski, Molholm, Gomez-Ramirez & Foxe, 2006), qui est parfois, à tort, confondu avec l'inégalité de Miller.

Cependant, divers arguments peuvent être avancés contre l'indépendance des distributions des TT unimodaux. Une corrélation positive peut être postulée si on estime que des facteurs fluctuant au cours de l'expérience affectent de la même façon le traitement dans les deux canaux sensoriels (attention, fatigue...). À l'inverse, une corrélation négative est envisageable si chacun des canaux sensoriels est en compétition pour certaines ressources (par exemple attentionnelles) : si les ressources attentionnelles sont portées sur le canal auditif, elles sont moins disponibles pour le canal visuel et il s'ensuit que les TR sont corrélés négativement. Une corrélation négative, en particulier, va diminuer les TR prédits par le modèle d'activations séparées car pour chaque couple de la distribution bivariée (TT_A, TT_V), l'un sera plutôt rapide et l'autre plutôt lent, ce qui aura pour effet que la distribution des minima comptera plus de TT courts que si les temps de traitement n'étaient pas corrélés. Donc, certains modèles d'activations séparées prédisent des TR plus rapides que le modèle d'activations séparées sous l'hypothèse d'indépendance. Cela se manifeste par le fait que pour des distributions corrélées négativement, le terme $p(TR_A < t) \times p(TR_V < t)$ tend vers 0 (voir aussi la figure 7.1 page suivante).

Pour éviter de faire des hypothèses sur l'indépendance des TR unimodaux, J. O. Miller (1982) remarque que le terme $p(TR_A < t \cap TR_V < t)$ est toujours positif ; il en déduit que

$$p[min(TR_A, TR_V) < t] \leq p(TR_A < t) + p(TR_V < t), \forall t$$

Cette inégalité (la véritable inégalité de Miller) est satisfaite par tous les modèles d'activations séparées, quelle que soit la dépendance des distributions des temps de traitement auditifs et visuels. Donc si la distribution des TR audiovisuels observés est telle qu'il existe une valeur de t telle que

$$p(TR_{AV} < t) > p(TR_A < t) + p(TR_V < t)$$

alors la distribution des TR audiovisuels ne peut s'expliquer par aucun modèle d'activations séparées, quelle que soit la corrélation existant entre les distribution des TT unimodaux.

7.1.2 Application de l'inégalité

Il est très facile de vérifier graphiquement si l'inégalité de Miller est respectée ou non. Il suffit de tracer la fonction de répartition des TR audiovisuels et la somme des fonctions de

répartition des TR auditifs et visuels. Graphiquement, l'inégalité est falsifiée et le modèle d'activations séparées rejeté si, à n'importe quel TR t, la fonction de répartition des TR bimodaux se trouve au-dessus de la somme des fonctions de répartition unimodales.

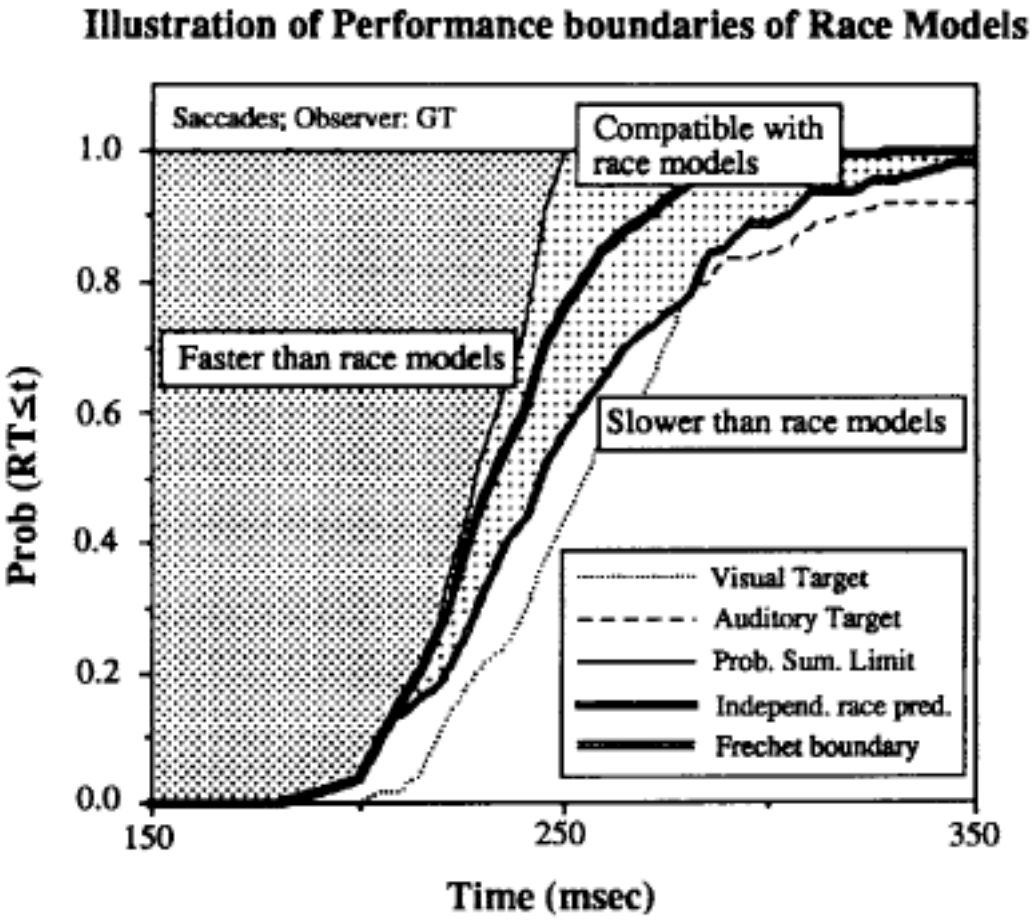

FIG. 7.1 – Illustration graphique de la falsification des modèles de compétition. De gauche à droite : la première courbe correspond à la somme des distributions de TR auditifs et visuels ; la deuxième, la plus épaisse, est la courbe prédite par un modèle d'activations séparées, sous l'hypothèse d'indépendance des distributions unimodales ; la troisième correspond au maximum des distributions unimodales et représente la distribution des TR prédits sous l'hypothèse d'une dépendance négative parfaite. Le cas qui nous intéresse est celui où la fonction de répartition des TR bimodaux se trouve dans la zone grisée à gauche des courbes : si c'est le cas, les TR bimodaux sont trop rapides pour être explicable par tout modèle d'activations séparées, quelle que soit la corrélation entre les distributions unimodales de TR. D'après Hughes et coll. (1994).

Plusieurs remarques sont cependant nécessaires pour appliquer et interpréter correctement l'inégalité. D'abord, il faut souligner que le terme de droite de l'inégalité de Miller, autrement dit la somme des fonctions de répartition des TR unimodaux, ne spécifie pas la répartition des TR prédite par un modèle spécifique d'activations séparées, mais la répartition des TR minimaux prédite par n'importe quel modèle d'activation séparées tels qu'ils sont définis par J. O. Miller (1982). Une conséquence en est que, lorsque t devient assez grand, cette somme devient supérieure à 1. Cela ne signifie pas que l'inégalité est

fausse mais simplement qu'elle ne spécifie pas les contraintes d'un unique modèle mais de plusieurs modèles à la fois. Pour des valeurs de t suffisamment grandes, l'inégalité est donc forcément respectée. Mais les valeurs de t intéressantes sont plutôt les valeurs faibles, puisque l'on s'attend, en cas d'interactions audiovisuelles à une accélération du temps de réaction et dans l'idéal à ce que les plus petits TR bimodaux soient inférieurs aux plus petits TR unimodaux.

Comme cela a déjà été souligné dans la partie 2.3.3 page 36, falsifier un modèle de compétition n'implique pas un modèle alternatif unique (tel que le modèle de coactivation), ni même que strictement tous les modèles d'activation séparées doivent être rejetés (voir par exemple le modèle d'activation séparées interactif de Mordkoff & Yantis, 1991). Certains modèles alternatifs sont cependant plus plausibles que d'autres et parmi les modèles plausibles, tous incluent des interactions entre les processus auditifs et visuels. Par ailleurs l'existence d'effets de compatibilité suggèrent que cette interdépendance est au moins en partie informationnelle. Mais la violation du modèle de l'inégalité ne permet pas en elle même de telles conclusions.

De nombreux articles ont utilisé l'ampleur de la violation, définie implicitement ou explicitement comme la surface comprise entre la fonction de répartition audiovisuelle et la somme des fonctions de répartition unimodales, comme une mesure directe de l'effet de facilitation intersensorielle. Une telle interprétation est tentante dans la mesure où elle met en relation les TR audiovisuels avec à la fois les TR auditifs et visuels. Or, on peut s'interroger sur la signification de cette valeur dans la mesure ou elle représente une déviation par rapport à une classe de modèle que l'on finit par rejeter. Pour autant, c'est une approximation qui parait raisonnable dans la mesure où cette valeur sera d'autant plus grande que la facilitation est grande[2].

7.1.3 Biais potentiels

Plusieurs biais potentiels dans l'application de l'inégalité de Miller à des données expérimentales ont été soulevés : Le premier biais est que l'inégalité n'est valable que pour les essais dans lesquels les stimuli sont analysés par le sujet et où il répond en fonction de cette analyse. Or on sait que, dans une certaine proportion des essais d'une expérience, le sujet est susceptible de répondre au hasard. Si de plus la vitesse de ces réponses faites au hasard est plus importante que celle des réponses où le stimulus est analysé, ce qui est tout à fait plausible, l'amplitude de la violation de l'inégalité peut être sous-estimée du fait de la présence de deux fois plus de ces essais au hasard dans la partie droite de l'inégalité (Eriksen, 1988). Une première façon de réduire ce biais est de ne prendre en compte que les essais pour lesquels le sujet a produit une réponse juste. Mais même dans ce cas, la

[2] (Colonius & Diederich, 2006) ont montré que cette mesure équivaut à la différence entre le TR bimodal moyen et le TR moyen prédit par un modèle d'activation séparée à dépendance maximalement négative, pour peu qu'on utilise comme mesure la surface entre la fonction de répartition bimodale et la fonction $min[1, p(TR_A < t) + p(TR_V < t)]$, c'est à dire la somme des fonctions de répartition unimodale bornée par 1, et non la somme $p(TR_A < t) + p(TR_V < t)]$ et que l'on prenne en compte la surface négative en la retranchant à la surface positive de la violation. Cette équivalence avec une différence de moyennes permet également de justifier l'utilisation de tests statistiques classiques pour comparer l'amplitude de la violation entre différentes conditions expérimentales.

présence d'essais pour lesquels le sujet répond juste par chance (*fast guesses*) peut biaiser les résultats. Pour réduire ce biais, Eriksen (1988) introduit un technique, qu'il appelle « tuer le jumeau » (*"kill-the-twin"*), consistant à faire l'hypothèse que la distribution des réponses justes par chance est la même que celle des réponses fausses. On peut alors retrancher cette fonction de répartition des réponses fausses à la fonction de répartition des TR des essais justes pour chaque condition auditive, visuelle et audiovisuelle. J. O. Miller et Lopes (1991) montrent par des simulations que cette technique réduit considérablement le biais. Toutefois, nous ne l'avons pas mise en œuvre car : d'une part, le biais va dans le sens d'une sous-estimation de la violation, donc si une violation est mise en évidence, son existence ne peut pas être remise en cause par cet argument ; d'autre part, notre but est de mettre en évidence une violation et pas forcément de la mesurer de manière exacte.

Le second biais a été évoqué dès les premières études (J. O. Miller, 1982), il s'agit du cout du changement de modalité. Il est en effet connu que le temps de réponse à un stimulus dans un modalité est plus court s'il suit un essai dans cette même modalité que s'il suit un stimulus d'une autre modalité (par exemple Turatto, Benso, Galfano & Umilta, 2002). Dans une expérience où les essais unimodaux et bimodaux sont présentés aléatoirement, il y a une probabilité plus forte qu'un essai unimodal montre ce cout par rapport à un essai bimodal, ce qui peut résulter en une surestimation de la violation de l'inégalité (et de la facilitation en général). Ce biais a été exclu comme principale cause de la violation par J. O. Miller (1986), puis Gondan, Lange, Rösler et Röder (2004), bien qu'il participe de façon négligeable à l'effet de facilitation.

7.1.4 Analyse statistique de groupe

La falsification de l'inégalité de Miller s'applique en principe pour chaque sujet et n'est pas un test d'hypothèse, donc elle ne garantit pas que la facilitation audiovisuelle n'est pas due à un biais d'échantillonnage des TR au niveau d'un sujet. Afin d'évaluer statistiquement la facilitation audiovisuelle, nous avons choisi de tester si la falsification de l'inégalité de Miller au niveau du groupe de sujets était ou non attribuable à un biais d'échantillonnage des sujets. Nous faisons l'hypothèse que si un effet est significatif au niveau du groupe, c'est qu'il reflète un effet réel au niveau des sujets.

Puisque l'application de l'inégalité de Miller utilise des distributions et non des moyennes de TR, il nous fallait rassembler les distributions des différents sujets sans perdre l'information d'appariement des distributions auditives, visuelles et audiovisuelles de chaque sujet. On ne pouvait donc se contenter de comparer à chaque valeur t les moyennes des effectifs cumulés des sujets car cela aurait pu gommer les violations de l'inégalité dans le cas où les sujets présentaient des différences importantes de TR moyens entre eux.

Pour obtenir les distributions de groupe, nous avons utilisé une technique de regroupement des distributions connue sous le nom de vincentisation, proposée à l'origine par Vincent (1912) et appliquée par J. O. Miller (1982) puis Giray et Ulrich (1993) au test statistique de l'inégalité de Miller dans un groupe de sujets. Elle consiste à calculer un fractile donné de la distribution de groupe comme la moyenne de ce fractile à travers les sujets. Cette façon de moyenner les distributions constitue une sorte de normalisation puisqu'elle permet d'éviter l'injection de variabilité due à des différences de TR absolus entre

sujets et de faire ressortir les différences de distributions présentes chez tous les sujets (Ratcliff, 1979). Nous avons donc, pour chaque sujet, calculé la somme de ses distributions de TR unimodales. Puis nous avons, pour chaque sujet, calculé les 19 fractiles d'ordre 20 pour cette distribution et pour la distribution des TR audiovisuels. Pour chacun des 19 fractiles, nous avons pu tester statistiquement la différence de moyenne grâce à un test de Student afin de voir si le fractile audiovisuel était plus faible que le fractile de la somme des distributions de TR unimodaux, c'est-à-dire si l'inégalité était violée.

En principe, on devrait corriger le risque de première espèce pour le rejet de l'hypothèse nulle, dans la mesure où l'on réalise un test pour chacun des 19 fractiles (voir la partie 8.1 page 111). Cependant, J. O. Miller (1982, note 3) a montré, grâce à des simulations que lorsque l'on garde un seuil $p < 0,05$ pour les 19 tests, le rejet erroné, sous l'hypothèse nulle, d'un modèle d'activations séparées ne dépasse 5% qu'au-delà du 7ème fractile, et seulement si la correlation négative est inférieure à -0,7. Comme les violations qui nous intéressent sont celles ayant lieu pour les TR les plus faibles, nous avons considéré cette garantie suffisante pour ne pas corriger le seuil.

7.2 Critère neurophysiologique d'intégration audiovisuelle

De nombreux critères ont été proposés pour identifier les structures cérébrales dans lesquelles ont lieu la convergence et l'intégration des informations de différentes modalités lors de la perception d'un évènement multisensoriel. Ces critères dépendent des méthodes d'investigation neurophysiologiques utilisées, mais beaucoup reposent sur la comparaison d'une condition de stimulation bimodale à des conditions de stimulation dans chacune des modalités séparément.

Historiquement, les première aires de convergence multisensorielle ont été identifiées en observant quelles structures montraient une activation similaire pour des stimulations dans différentes modalités sensorielles (voir la partie 1.1.1 page 5). C'est de cette manière qu'a été découverte la convergence dans la formation réticulée et les aires corticales associatives. L'activation d'une même aire corticale par différents stimuli en ECoG ne garantit cependant pas que ces stimuli activent les même cellules. Il est en effet possible qu'au sein d'une même population de neurones des sous-populations différentes soient activées par des stimuli de différentes modalités. Avec le développement des techniques d'enregistrement des réponse unitaires, il a toutefois pu être montré que certains neurones répondaient individuellement à des stimuli de différentes modalités ; de tels neurones ont été trouvés non seulement dans les aires associatives (voir la partie 1.1.2 page 8) mais également dans les cortex dits unisensoriels (le cortex visuel en particulier, voir la partie 1.2 page 10). Cependant, les auteurs qui se sont intéressés à ces questions n'ont tout d'abord pas envisagé l'existence possible de réponses associées spécifiquement à la présentation concomitante de stimuli dans plusieurs modalités, comme l'attestent les protocoles utilisés, dans lesquels les composantes auditives et visuelles étaient toujours séparées par un intervalle temporel. Les premières études s'étant intéressées à la stimulation bimodale simultanée sont en fait celles sur le colliculus supérieur au début des années 80. Ces études ont montré l'existence de réponses neuronales multiplicatives propres à la présentation simultanée de stimuli de différentes modalités sensorielles (voir partie 1.4.1 page 14).

Avec l'application des techniques d'imagerie non invasives de l'activité cérébrale humaine pour étudier l'intégration multisensorielle a commencé à se poser la question de l'identification de ces réponses spécifiques chez l'homme. Les techniques non invasives enregistrent typiquement l'activation de grandes populations de neurones, et on ignore en grande partie les relations existant entre l'activation au niveau de la population et l'activité au niveau cellulaire. Les critères proposés pour l'EEG ou l'IRMf ont donc été, de fait, relativement indépendants des principes d'intégration découverts au niveau cellulaire, même s'ils y ressemblent ou en sont parfois inspirés (voir la partie 4.5 page 72).

7.2.1 Falsification du modèle additif en EEG/MEG

(Cette discussion du modèle additif en électrophysiologie a fait l'objet d'une publication dans la revue Cognitive Processing, jointe en annexe page 240)

En ce qui concerne les potentiels évoqués, un critère d'identification proposé par Giard et Peronnet (1999), et que nous utiliserons pour les expériences sur la parole et dans une expérience sur la mémoire sensorielle, est que la réponse évoquée par un stimulus audiovisuel soit différente de la somme des réponses évoquées séparément par un stimulus auditif et un stimulus visuel. Ce critère est basé sur le principe de la sommation linéaire des potentiels électriques : de la même façon que l'activité en tout point du scalp est la somme linéaire de tous les courant générés à un instant donné dans le cerveau (voir la partie 6.2.2 page 89), si les traitements des composantes auditive et visuelle d'un stimulus sont totalement indépendants alors l'activité électrique générée par le traitement d'un stimulus audiovisuel devrait être égale à la somme des activités électriques générées par ses deux composantes présentées séparément (d'où le nom de modèle additif). Le critère proposé est donc celui du rejet de ce modèle additif, autrement dit du rejet de l'hypothèse de non convergence des informations auditives et visuelles : si cette égalité n'est pas respectée, c'est que les informations visuelles et auditives ont convergé ou interagi à un instant donné.

Ce critère avait en fait été appliqué plusieurs fois sous des formes légèrement différentes : Berman (1961) l'avait utilisé pour étudier les interactions audio-tactiles à la surface du cortex de rats, dans le cas de stimulations auditives et somesthésiques successives dans le temps. Il a également été utilisé par L. K. Morrell (1968b) en potentiels évoqués chez l'homme pour étudier les corrélats de l'effet d'un stimulus auditif accessoire sur le TR visuel. Plus récemment, il a été utilisé pour identifier les aires d'interactions audiovisuelles chez les rat en ECoG (Barth et coll., 1995). Dans ce contexte, ce critère permettait de voir à quelles latences et au dessus de quelles zones corticales, le traitement du stimulus bimodal différait du traitement de ses constituant unimodaux et ainsi d'identifier des zones de convergence ou d'intégration.

Dans le cas de l'EEG de scalp, cependant, si la violation du modèle additif à une latence donnée suffit à affirmer qu'à cette latence les traitements auditifs et visuels interagissent ou ont interagi, il est plus difficile d'identifier les structures dans lesquelles peuvent avoir lieu ces interactions. Ainsi, si une violation du modèle additif est observée à une électrode de scalp donnée, rien n'indique que la source de cette interaction se trouve sous cette électrode en raison de la diffusion des potentiels sur le scalp. Il est donc nécessaire, comme c'est le cas dans toute recherche de localisation des sources en EEG/MEG, de prendre en compte la distribution de la violation du modèle additif sur tout le scalp. Cette distribution peut

permettre de localiser l'origine des effets observés (dans les limites de résolution spatiale de l'EEG/MEG, voir la partie 6.2.2 page 88).

Biais possibles dans l'application du modèle additif

Les interprétations possibles des interactions audiovisuelles, estimées par la violation du modèle additif, sont multiples et il faut avoir à l'esprit les diverses limites que cette méthode présente. Les limites les plus évidentes tiennent au fait qu'en faisant cette estimation, on s'éloigne dangereusement des standards de la démarche expérimentale puisque l'on compare une condition expérimentale à 2 autres conditions expérimentales. Toute variable qui affecterait de manière identique la variable dépendante (les potentiels évoqués) dans les trois conditions de stimulation apparaitrait nécessairement, à tort, comme une interaction audiovisuelle. Il faut donc tenter d'identifier les conditions dans lesquelles de tels effets pervers peuvent avoir lieu et éviter l'application du critère dans ces situations, ou proposer des modifications du paradigme expérimentale ou de l'analyse des signaux permettant d'éviter l'expression de ces variables.

Tout d'abord, certaines composantes des potentiels évoqués reflètent des processus de sélection de la réponse et des processus moteurs. Si le sujet a pour tâche de répondre aux stimuli auditifs, visuels et audiovisuels, on s'attend nécessairement à observer des violations de l'additivité puisque les activités liées à la réponse seront ajoutées une fois et retranchées deux fois. D'une certaine manière, ces interactions reflètent une sorte de convergence audiovisuelle à un niveau tardif de traitement : des stimuli différents accèdent à un même processus et l'on sait que la réponse comportementale à un stimulus bimodal n'est pas égal à la somme des réponses comportementales à ses composantes unimodales (quelle que soit la façon dont on mesure ces réponses). Entre les stimuli et les réponses, il existe forcément une étape où l'additivité n'est plus respectée. Excepté dans le cas où l'on n'observe aucune interaction avant ces processus de réponse, une non-additivité à ce stade n'est pas très intéressante. De manière générale, on ne s'attend pas à observer des potentiels évoqués liées à la réponse avant environ 200 ms de traitement chez l'homme (voir Hillyard, Teder-Sälejärvi & Munte, 1998, pour une revue, ainsi que la partie 1.3 page 11). Il est donc prudent de limiter l'application du modèle additif aux traitements ayant lieu avant 200 ms.

D'autres composantes communes aux trois conditions de stimulation peuvent apparaitre dans les potentiels évoqués, en particulier dans les paradigmes expérimentaux où le sujet doit réaliser une réponse chronométrée : il s'agit de réponses anticipatoires lentes, visibles dès la période pré-stimulus, quelle que soit la condition de présentation et qui peuvent donc faire apparaitre des effets dans le calcul des interactions audiovisuelles à des latences précoces (Teder-Sälejärvi et coll., 2002). Ces composantes anticipatoires sont d'autant plus fortes que la survenue de la stimulation est prédictible. Une façon de les atténuer est de présenter les stimulations avec un intervalle inter-stimulus aléatoire qui réduit la prédictibilité. Comme ces composantes sont lentes, on peut aussi les éliminer assez efficacement dans l'analyse en filtrant les potentiels évoqués avec un filtre passe-haut à 1,5 ou 2 Hz.

Les écueils dus aux composantes communes disparaissent totalement dans un cas tout à fait particulier, mais qui se présentera dans notre première expérience électrophysiologique sur la mémoire sensorielle : ici, le modèle additif sera appliqué à des différences de

PE calculées séparément pour des déviances unimodales et bimodales et non plus directement aux potentiels évoqués par des stimulations sensorielles. Les composantes communes disparaissent dans les différences, et le modèle additif peut alors être appliqué sans risque.

Un autre biais peut apparaitre si les stimulations dans les différentes conditions auditive, visuelle et audiovisuelle sont présentées dans des blocs distincts. En effet, plusieurs études en IRM fonctionnelle ont montré que la stimulation continue dans une modalité sensorielle pouvait diminuer le débit sanguin cérébral dans les aires corticales spécifiques des autres modalités (Haxby et coll., 1994 ; Kawashima, O'Sullivan & Roland, 1995 ; Laurienti et coll., 2002). Si ces effets de déactivation se manifestent dans les potentiels évoqués, ils apparaitront dans les interactions audiovisuelles puisqu'ils n'ont, a priori, pas leur équivalent en condition audiovisuelle, où les deux modalités sensorielles sont sollicitées. Une façon d'éviter ces effets est de présenter les stimuli des différentes conditions de manière complètement aléatoire et équiprobable.

7.2.2 Interprétation des violations de l'additivité en EEG/MEG

Si ces quelques précautions sont respectées, cela devrait permettre de limiter les effets indésirables et de mettre en évidence des traitements spécifiques à l'intégration des informations auditives et visuelles. Ces traitements spécifiques peuvent prendre différentes formes : ils peuvent soit correspondre à l'activation de structures qui ne sont activées par aucune des deux stimulations unisensorielles présentées séparément. Dans ce cas la topographie des effets d'interaction sur le scalp devrait être différente de celles observées dans l'une et l'autre des conditions unisensorielles. Mais les traitements spécifiques à l'intégration peuvent aussi correspondre à l'influence des informations d'une modalité sensorielle sur les traitements dans l'autre modalité sensorielle. Dans ce second cas, les interactions devraient refléter une modulation de l'activité unisensorielle et avoir une topographie identique à celle évoquée par la stimulation dans la modalité sensorielle modulée. Notons ici que la polarité positive ou négative des interactions calculées dans la cadre de la violation du modèle additif n'indique pas directement si une telle modulation correspond à une diminution ou à une augmentation de l'activité unimodale, puisque cette activité unimodale peut se manifester elle-même par des polarités positives ou négatives — les deux pouvant même refléter une composante unique dans le cas d'un dipôle équivalent tangentiel (voir la partie 6.2.2 page 88). C'est la comparaison entre la polarité des interactions et la polarité de l'activité unimodale qui permettra de se prononcer sur le fait que les interactions reflètent une diminution ou une augmentation de l'activité unimodale.

7.2.3 Comparaison avec le critère d'additivité en IRM fonctionnelle

Le modèle additif ressemble à d'autres critères d'additivité utilisés dans l'étude des interactions multisensorielles avec d'autres techniques de neuroimagerie. Ainsi, Calvert (2001) propose un critère d'identification des zones d'intégration audiovisuelle en IRMf qui consiste également à comparer les activations (les augmentations du débit sanguin cérébral) en condition de stimulation bimodale à la somme des activations dans les conditions

de stimulation unimodale (voir la partie 4.5 page 72). Cette ressemblance est cependant trompeuse car les implications d'une non-additivité dépendent de la nature des variables enregistrées. Je me contenterai ici de souligner une différence fondamentale entre l'application de l'additivité en EEG/MEG et en IRMf. D'autres implications méthodologiques propres à l'IRM fonctionnelle ont été discutées plus spécifiquement par Calvert et Thesen (2004), puis Laurienti, Perrault, Stanford, Wallace et Stein (2005).

Si on fait l'hypothèse que l'activation cérébrale pour un stimulus bimodal est égale à la somme des activations pour ses composante unimodales, alors la loi de superposition des potentiels électriques implique que les potentiels générés par le stimulus bimodal sont égaux à la somme des potentiels générés par les deux stimuli unimodaux. Si l'on doit rejeter l'additivité au niveau électrique, cela implique logiquement qu'on doive la rejeter au niveau physiologique (modus tollens). Or on ne dispose pas d'une telle loi biophysique dans le cas de la réponse hémodynamique : on ne sait pas comment se comporterait la variation de la réponse hémodynamique en un voxel sous l'hypothèse d'une additivité physiologique (par exemple si l'additivité est due au fait que deux populations neuronales indépendantes sont activées, cette additivité n'apparaitra pas nécessairement au niveau du flux sanguin, car ce dernier augmente peut-être plus vite ou moins vite que le nombre de neurones à irriguer). En d'autres termes, on n'a pas de raison de supposer une relation linéaire entre l'activité neuronale et la variation du débit sanguin cérébral (alors que cette relation linéaire se justifie pour l'enregistrement de l'activité électrique ou magnétique). Par conséquent, si l'additivité des variations du débit sanguin cérébral n'est pas respectée, cela n'implique pas qu'elle ne l'est pas au niveau neuronal. L'interprétation d'une non-additivité des variations du débit sanguin cérébral est donc très hasardeuse.

Calvert (2001) a justifié l'utilisation du critère d'additivité en IRMf par le fait qu'au niveau neuronal, certaines cellules montrent des réponses bimodales super-additives (voir partie 1.4.1 page 14). Le raisonnement est le suivant : si certaines structures contiennent de tels neurones multisensoriels, alors la réponse de ces structures à un stimulus bimodal devrait être supérieure à somme des réponses à ses composantes unimodales. Il parait cependant difficile d'extrapoler ainsi directement des critères, basés sur les propriétés multiplicatives du taux de décharge des neurones multisensoriels, à l'analyse de mesures macroscopiques, car celles-ci dépendent de variables physiologiques différentes. En effet, l'activité observable au niveau macroscopique résulte probablement plus de l'activité post-synaptique que des potentiels d'actions, aussi bien en EEG/MEG (voir la partie 6.1 page 83) qu'en IRMf (Logothetis, Pauls, Augath, Trinath & Oeltermann, 2001 ; Logothetis, 2003). Sans modèle précis et quantitatif des relations entre le taux de décharge neuronal et ces variables physiologiques macroscopiques, il est donc impossible d'extrapoler le critère de super-additivité aux mesures non invasives chez l'homme. Comme le soulignent Laurienti et coll. (2005), cet argument est valable aussi bien pour le critère d'additivité en IRMf qu'en EEG/MEG : l'observation d'une violation de l'additivité en EEG/MEG ou en IRMf n'implique pas la présence, dans les structures à la source de cette violation, de neurones bimodaux présentant un comportement intégratif multiplicatif.

Chapitre 8

Méthodes statistiques appliquées à l'électrophysiologie chez l'homme

Afin d'atteindre un certain degré de généralisabilité des résultats, les expériences d'EEG et de MEG ont été menées sur de petits échantillons de sujets (entre 10 et 20), censés être représentatifs de la population (jeune et étudiante) générale. Les potentiels évoqués montrent une variabilité intersujet certaine et il est nécessaire de s'assurer que les effets observés dans les potentiels évoqués moyens du groupe reflètent une tendance générale et non la contribution des potentiels évoqués de quelques individus, c'est-à-dire qu'ils ne sont pas la conséquence du hasard de l'échantillonnage (au sens statistique). Dans nos premières expériences, les tests statistiques des analyses de groupe étaient des tests de Student classiques, à mesures répétées puisque l'on disposait pour chacun des sujets de ses PE dans chacune des conditions de stimulation. Pour tester le modèle additif, on calculait donc pour chaque sujet la violation du modèle additif à tous les échantillons temporels et à toutes les électrodes et on comparait la valeur obtenue à zéro grâce à un test de Student pour chaque échantillon de chaque électrode.

8.1 Tests multiples

Un problème classique de ce type d'approche est que le risque d'obtenir un test significatif alors que la différence est due au hasard (risque de première espèce, noté α) augmente considérablement avec le nombre de tests réalisés. Si l'on accepte un risque $\alpha = 0,05$ à chaque test, la probabilité de ne pas se tromper à chaque test est de $1 - \alpha = 0,95$. Donc la probabilité de ne jamais se tromper sur m tests est de $(0,95)^m$. Le risque global α_{global} de se tromper au moins une fois en effectuant m tests est donc $\alpha_{global} = 1 - (0,95)^m$. Dans notre première expérience sur la parole, nous avons réalisé 7000 tests (35 électrodes sur une période 200 ms échantillonnée toutes les millisecondes), donc le risque α_{global} était égal à $1 - (0,95)^{7000}$ c'est-à-dire presque 1. Il fallait donc trouver une méthode pour limiter α_{global} à 0,05.

L'approche la plus directe est la correction de Bonferroni qui consiste à diviser le risque global accepté (par exemple $\alpha = 0,05$) par le nombre de tests effectués et d'appliquer ce

risque corrigé à chacun des tests, ce qui permet de garder le risque global à $\alpha_{global} = 0,05$[1]. Dans notre exemple, l'application de cette correction aurait nécessité de choisir un risque local d'environ 10^{-5}. Or c'est un seuil beaucoup trop stricte pour l'EEG de scalp, en particulier pour détecter des effets aussi subtils que des violations du modèle additif. Si la correction de Bonferroni est trop stricte, c'est parce qu'elle ne tient pas compte du fait que les potentiels sont corrélés dans le temps et dans l'espace. Lorsque des tests statistiques sont effectués sur des mesures corrélées, le risque global n'augmente pas aussi rapidement avec le nombre de tests que pour des mesures indépendantes (Manly, McAlevey & Stevens, 1986).

Pour résoudre ce problème, dans le cas particulier des potentiels évoqués sur un grand nombre d'échantillons temporels, Guthrie et Buchwald (1991) proposent, non pas de corriger le risque localement pour chaque test, mais d'imposer un nombre minimum d'échantillons significatifs successifs qui garantit un risque $\alpha_{global} = 0,05$. Ce risque global est calculé grâce à une statistique globale n_{max} : le nombre maximal de tests de Student significatifs à p<0,05 successifs obtenus sur une fenêtre temporelle d'une taille donnée. Sous l'hypothèse nulle, cette statistique a une certaine distribution et il existe une valeur de n_{max} qui n'a pas plus de 5% de chance de se produire. Si le nombre de test significatifs successifs obtenu sur les données est supérieur à cette valeur, alors la probabilité qu'ils aient été obtenus par chance est inférieure à 5%, donc le risque global de premier espèce reste limité à 5% sur l'ensemble des echantillons temporels. Cette valeur critique de n_{max} dépend non seulement du nombre de sujet, de la taille de la fenêtre et des risques locaux et globaux, mais également de l'auto-corrélation temporelle des potentiels évoqués. En faisant un certain nombre d'hypothèses sur la structure temporelle des potentiels évoqués, Guthrie et Buchwald (1991) ont tabulé, grâce à des simulations, ces valeurs critiques de n_{max} en fonction de ces différents paramètres.

Pour les tests statistiques du modèle additif dans l'expérience sur la parole en EEG de scalp, nous avons utilisé cette table afin de tenir compte des tests multiples, au moins à chaque électrode. Cette méthode ne me paraissait cependant pas très satisfaisante pour plusieurs raisons : d'une part, la table proposée est trop limitée en ce qui concerne la fenêtre d'analyse, et d'autre part, des hypothèses sont faites sur la structure temporelle du signal.

Une autre méthode proposée par Blair et Karniski (1993) permet d'éviter ces hypothèses en calculant la distribution de la statistique n_{max} dans le cas particulier des données sur lesquelles ont réalise le test. À partir de cette distribution, on peut calculer les valeurs critiques de n_{max} permettant de limiter le risque α_{global} à $0,05$. Pour estimer la distribution de la statistique n_{max}, on utilise une méthode de permutation : sous l'hypothèse nulle de l'absence de différences entre deux conditions, les données correspondant aux deux conditions sont interchangeables pour un individu donné, puisque les deux échantillons sont réputés être tirés de la même population. On réalise un nombre maximum de 2^N permutations aléatoires des données et on calcule à chaque permutation la probabilité

[1] C'est une approximation de la valeur exacte, ou correction de Sidak, due au fait que $1-(1-a)^b$ vaut environ $a \times b$ lorsque b est assez grand et que a est assez petit. Donc lorsque le risque non corrigé α est assez petit, donc que m est assez grand, le risque global vaut $\alpha_{global} = 1-(1-\alpha)^m \approx \alpha \times m$.

associée au test de Student et le nombre maximum d'échantillons significatifs au risque local $\alpha = 0,05$ dans la fenêtre d'analyse (lors des permutations, les échantillons temporels conservent leur structure et ne sont jamais permutés entre eux). On obtient donc une distribution de n_{max} sous l'hypothèse nulle, spécifique de la fenêtre d'analyse choisie, du nombre de sujets et aussi de la structure temporelle particulière des données testées (qui peut être différente à chaque électrode). Grâce à cette distribution, on trouve la valeur critique de n_{max} correspondant à un risque de $\alpha_{global} = 0,05$ (c'est-à-dire le nombre maximal de tests significatifs successifs obtenu dans moins de 5% des permutations sous l'hypothèse nulle). Lors des tests de Student sur les données non permutées, on ne considère alors que les successions significatives supérieures à cette valeur critique, ce qui permet de limiter le risque global à 5%.

Cette méthode est appliquée par Blair et Karniski (1993) aux tests de Student, mais peut être étendue à tout type de test d'hypothèse, paramétrique ou non paramétrique. En collaboration avec Pierre-Emmanuel Aguera, nous avons développé un programme permettant d'appliquer cette correction à des tests de permutation pour des mesures répétées. Dans ce cas, la significativité de la différence à chaque échantillon temporel est estimée dans un permier temps par la méthode des permutations en calculant la distribution des différences de moyenne pour $m = 2^N$ permutations et dans un deuxième temps, la méthode des permutations est à nouveau appliquée pour calculer les valeurs critiques de n_{max}. Ces valeurs critiques sont ensuite utilisée comme critères pour conserver ou non les différences significatives trouvées lors de la première application des permutations. Nous avons appliqué cette méthode statistique dans la deuxième expérience EEG sur la mémoire sensorielle et une méthode analogue (appliquée aux échantillons indépendants) dans le traitement statistique des données sEEG (voir la partie 8.2.2 page suivante).

Dans la première expérience de PE sur la mémoire sensorielle, nous avions suivi une stratégie plus classique consistant à ne prendre en compte la significativité des tests statistiques que dans la période de temps et sur les électrodes sur lesquels nous nous attendions à observer des effets, ce qui permet de réduire le nombre de tests et donc le problème des tests multiples.

Notons que les méthodes décrites dans cette partie ne prennent en compte les tests multiples que dans la dimension temporelle et non dans la dimension spatiale.

8.2 Tests statistiques appliqués aux données individuelles en sEEG et en MEG

8.2.1 Tests sur les essais élémentaires

Les données sEEG enregistrées chez les patients doivent être traitées différemment des données EEG de surface car, étant donné que chaque patient a une implantation d'électrodes particulière, on ne dispose pas de valeurs de potentiels comparables d'un patient à l'autre et qui pourraient donner lieu à une analyse de groupe. Il n'est donc pas possible d'évaluer statistiquement la généralisabilité des résultats à une population de patients et

encore moins à la population générale[2]. Néanmoins, pour s'assurer de la validité des résultats au niveau d'un patient, il était nécessaire d'évaluer statistiquement les effets au niveau individuel. Les potentiels évoqués étant la moyenne d'observations réalisées sur un groupe d'essais élémentaires, on peut comparer deux (ou trois) groupes d'essais, correspondant aux deux (ou trois) conditions de stimulation et tester l'hypothèse nulle que les échantillons ont été tirés par hasard d'une même population d'essai. Si on peut rejeter cette hypothèse avec un certain risque inférieur à 5%, alors on considère que les deux potentiels évoqués ont été moyennés à partir d'essais individuels reflétant des traitements différents, chez un patient particulier.

Dans les analyses de groupe réalisées en EEG ou en MEG, on considère généralement que la distribution des potentiels évoqués à chaque échantillon temporel de chaque électrode suit une loi de distribution normale et on utilise souvent des tests paramétriques basés sur cette hypothèse. Pour les analyses sur les essais élémentaires, nous avons préféré ne pas faire cette hypothèse et nous avons utilisé des tests d'hypothèse non paramétriques.

Dans le cas où l'on voulait tester l'émergence d'une activité par rapport à la ligne de base, nous avons utilisé un test de Wilcoxon pour comparer l'amplitude du potentiel à la valeur moyenne de la ligne de base (ce qui revient à tester la différence entre l'amplitude corrigée en ligne de base à zéro). Donc, dans ce cas, les échantillons étaient appariés.

Lorsqu'il s'agissait de comparer deux conditions, nous avons dû utiliser des tests non paramétriques pour groupes indépendants puisque les deux groupes d'essais dans les deux conditions ne pouvaient être appariés. C'était le cas pour les données individuelles en MEG lorsque nous avons voulu tester la différence entre les CME évoqués par les stimuli standards et déviants. Pour ces tests, nous avons utilisé une méthode de randomisation pour comparer les lois de distributions de deux groupes indépendants (voir par exemple Edgington, 1995).

Ce cas s'est présenté également pour le test du modèle additif en sEEG, où les essais étaient répartis en 3 groupes de mesures indépendantes : un groupe d'essais auditifs, un groupe d'essais visuels et un groupe d'essais audiovisuels. Or, ce cas est tout à fait particulier puisque l'application d'un test statistique au modèle additif, telle que nous l'avons exposée plus haut, nécessite en principe que les données soient appariées afin de pouvoir calculer une distribution des sommes de potentiels auditifs et visuels. Comme les essais élémentaires ne sont pas appariés, j'ai conçu un test d'hypothèse spécifique au test du modèle additif pour des groupes indépendants, que nous avons implémenté pour le traitement des données sEEG, avec l'aide de P-E. Aguera.

8.2.2 Test du modèle additif par randomisation pour des données non appariées

Le principe du test est basé sur une méthode de randomisation pour comparer les lois de distribution de deux échantillons indépendants. Soient N_A, N_V et N_{AV} les effectifs des

[2]c'est également en partie vrai pour la MEG, où la position de la tête varie beaucoup par rapport aux capteurs entre les différents sujets, ce qui a pour effet d'introduire une variabilité supplémentaire dans les données de groupe.

groupes d'essais auditifs, visuels et audiovisuels. Le modèle additif équivaut à l'hypothèse nulle que la somme d'un essai visuel et d'un essai auditif est tiré de la même population qu'un essai audiovisuel. Sous cette hypothèse nulle, on peut donc "mélanger" les essais audiovisuels et les sommes d'essais unimodaux et les attribuer aléatoirement à l'une des deux conditions. De même que dans le test de permutation, on peut estimer la distribution de la statistique qui nous intéresse (ici la violation du modèle additif), sous l'hypothèse nulle, en réalisant un grand nombre de tirages aléatoires. Afin que les effectifs des modalités A et V participant au calcul de la somme soient identiques, nous avons arbitrairement conservé les $min(N_A, N_V)$ premiers essais élémentaires A et V.

Chaque randomisation se déroulait de la façon suivante : pour constituer les sommes, on appariait au hasard les essais auditifs et les essais visuels, sans remise, de façon à obtenir $min(N_A, N_V)$ sommes. Chaque somme était obligatoirement la somme d'un essai auditif et d'un essai visuel. Ces $min(N_A, N_V)$ sommes étaient ensuite mélangées aux N_{AV} essais audiovisuels et l'on répartissait au hasard ces essais dans des groupes d'effectifs $min(N_A, N_V)$ et N_{AV}, sans remise. Ces deux nouveaux groupes d'essais servaient à calculer la valeur de la violation du modèle additif sous l'hypothèse nulle pour cette randomisation. Lors de toutes les attributions aléatoires, les différents échantillons temporels à toutes les électrodes d'un essai donné étaient bien sûr solidaires et restaient toujours associés à cet essai.

On effectuait 10 000 randomisations de ce type, ce qui permettait d'établir la distribution de la violation du modèle additif sous l'hypothèse nulle, pour chaque échantillon temporel de chaque électrode. On calculait ensuite la violation du modèle additif sur les données réelles non randomisées, c'est-à-dire la différence entre le potentiel évoqué moyen audiovisuel et la somme des potentiels évoqués unimodaux. Pour chaque échantillon temporel de chaque électrode, on pouvait situer cette valeur de différence moyenne par rapport à la distribution, calculée par permutations, correspondant à cet échantillon, et donc estimer la probabilité que cette valeur soit due au hasard sous l'hypothèse nulle.

Pour corriger les tests multiples, nous avons utilisé une méthode analogue à celle décrite plus haut pour les analyses de groupe, basée sur l'estimation de la distribution du nombre maximum d'échantillons significatifs successifs (voir la partie 8.1 page 112). Dix mille randomisations étaient réalisées une seconde fois afin de déterminer, pour chaque électrode, le nombre n_{max} de tests successifs, significatifs avec un risque local $\alpha = 0,05$, nécessaires pour garder le risque global à $0,05$.

8.2.3 Remarques

En ce qui concerne la généralisabilité des résultats, on ne disposait pas de méthode quantitative dans le cas de l'étude sEEG puisque les implantations de chaque patients n'étaient pas strictement comparables. L'interprétation des résultats et leur généralisation au cas de la population normale était donc uniquement qualitative, essentiellement en faisant l'hypothèse que la pathologie des patients n'avait aucune incidence sur les processus étudiés et qu'un effet significatif observé chez plusieurs patients dans une zone anatomiquement équivalente avait un certain degré de généralisabilité.

Il important de noter que le fait d'utiliser des tests non paramétriques ne signifie pas que l'on ne fait aucune hypothèse sur la distribution des variables. S'il est vrai que dans ces tests,

on n'a pas besoin de supposer la normalité des variables, il peut être nécessaire de poser l'égalité des variances lorsque l'on cherche à tester une différence entre deux conditions. Ainsi, les hypothèses nulles utilisées dans les tests de permutations et de randomisations que nous avons utilisés sont des hypothèses d'égalité des distributions et non d'égalité des moyennes : donc, si le test conclut à une différence de distribution des deux conditions, il est nécessaire de poser l'hypothèse supplémentaire que les variances sont égales pour pouvoir conclure à une différence de moyenne, qui est en général la conclusion que l'on cherche à atteindre.

Troisième partie

Interactions audiovisuelles dans la perception de la parole

Chapitre 9

Étude en EEG et comportement

Cette première étude a été réalisée lorsque j'étais en DEA à l'unité 280, sous la direction de Marie-Hélène Giard, et l'analyse des données s'est poursuivie au début de ma thèse. Cette étude ayant fait l'objet d'une publication (Besle, Fort, Delpuech & Giard, 2004), elle ne sera que brièvement présentée ici. Les détails en sont décrits dans la publication, intégrée au manuscrit en annexe (page 245).

9.1 Rappel de la problématique

Nombre de données comportementales ont montré que des indices visuels de parole (les mouvements des lèvres en particulier) pouvaient influencer la perception auditive de la parole. Une partie de ces interactions a vraisemblablement lieu, entre autres, à une étape précoce du traitement, avant la catégorisation phonologique des sons de parole (voir la partie 3.2.2 page 54). La plupart des études de neuroimagerie qui ont traité de la question de l'intégration des indices auditifs et visuels dans la perception de la parole ont cependant utilisé l'IRM fonctionnelle. Ces études ont montré l'implication de plusieurs aires corticales dans cette intégration mais ne pouvaient leur assigner de place dans la chaine des traitements, étant donné la faible résolution temporelle de la technique utilisée.

L'EEG est une technique d'enregistrement particulièrement adéquate pour tenter de mettre en évidence différentes étapes de traitement, où peut opérer l'intégration audiovisuelle. Au moment ou nous avons conduit cette expérience, les seuls résultats en EEG/MEG sur la perception de la parole avaient cependant soit montré des interactions audiovisuelles à des latences très tardives autour de 450 ms (en utilisant donc le modèle additif en dehors de son domaine d'application, Sams & Levänen, 1998), soit uniquement établi une borne temporelle supérieure pour les premières interactions audiovisuelles : c'est le cas des études ayant montré l'existence d'une MMN auditive, vers 200 ms, pour des syllabes déviant sur leur dimension visuelle (Colin, Radeau, Soquet, Demolin et coll., 2002 ; Möttönen et coll., 2002 ; Sams et coll., 1991, voir aussi la partie 4.6.1 page 72). L'utilisation du modèle additif dans les 200 premières millisecondes de traitement devrait donc permettre de mettre en évidence le décours temporel des interactions audiovisuelles ayant lieu en amont.

L'utilisation de techniques d'imagerie et du modèle additif permettent de plus d'étudier la perception de la parole naturelle, sans recourir à la présentation d'informations conflic-

tuelles ou bruitées comme cela a souvent été le cas dans les études comportementales de la perception de la parole bimodale. Nous avons donc enregistré les potentiels évoqués par des syllabes présentées soit uniquement dans la modalité auditive, soit uniquement dans la modalité visuelle, soit dans les deux modalités simultanément et avons comparé le potentiel évoqué audiovisuel à la somme des potentiels évoqués auditifs et visuels, de façon à déterminer les aires cérébrales et les étapes de traitement où ces interactions ont lieu.

9.2 Méthodes

9.2.1 Sujets

Seize sujets droitiers (dont 8 de sexe féminin), âgés en moyenne de 23 ans ont passé cette expérience. Aucun sujet ne souffrait de troubles neurologiques. Ils avaient tous une audition normale et une vision normale ou corrigée.

Treize autres sujets (dont 9 de sexe féminin) âgés en moyenne de 24,3 ans ont participé à l'étude comportementale.

9.2.2 Stimuli

Les stimuli étaient des syllabes /pa/, /pi/ , /po/ et /py/, prononcées par une locutrice de langue maternelle française et enregistrées à une fréquence d'échantillonnage de 25 images/s pour l'image et de 41 kHz pour le son. Trois exemplaires différents de chacune des syllabes ont été sélectionnés, sur un corpus d'une centaine de syllabes enregistrées, de manière à conserver une certaine variabilité naturelle de la parole, nécessaire pour que les sujets traitent les stimuli sur un plan linguistique et ne se contentent pas de discriminer les stimuli sur des traits de surface non pertinents, tels qu'une légère différence d'éclairage ou de position des lèvres au départ de la syllabe. Ces 12 syllabes ont été sélectionnées de sorte qu'elles aient approximativement toutes la même structure temporelle audiovisuelle et ont ensuite été légèrement modifiées de façon à présenter des caractéristiques temporelles véritablement identiques (temps séparant le début du mouvement des lèvres, l'ouverture de la bouche, l'explosion de la consonne et début du voisement) et ainsi minimiser la variabilité des réponses évoquées auditives et visuelles. La structure audiovisuelle des syllabes finales est décrite dans la figure 9.1 page ci-contre. Seule la partie inférieure du visage de la locutrice était présentée aux sujets, la bouche ayant une taille de 2,2° d'angle visuel. Le niveau sonore était confortable.

Dans ces stimuli, les informations visuelles commençaient 240 ms avant le début des informations auditives. Les mouvements des lèvres dans les 6 premières images avant l'ouverture de la bouche étaient toutefois de faible amplitude. Nous avons vérifié dans une pré-expérience sur un groupe de 7 sujets qu'ils ne pouvaient donner d'indice sur l'identité de la syllabe. Nous avons pour cela demandé au sujet de tenter d'identifier la syllabe visuelle, qui pouvait être tronquée à la 6ème, la 8ème ou la 14ème image. Les résultats (table 9.1 page suivante) montrent que les sujets répondent au hasard (26% de bonnes réponses en moyenne) lorsque la syllabe s'arrêtait à la 6ème trame. Dès la 8ème image cependant, les informations visuelles étaient suffisantes pour atteindre la performance ob-

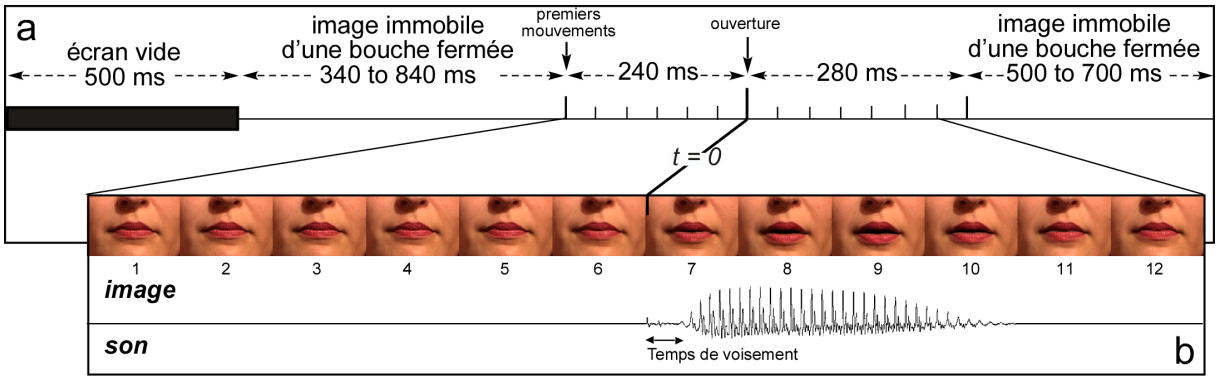

FIG. 9.1 – Structure temporelle d'un essai audiovisuel (a) et d'une syllabe audiovisuelle (b). chaque trame vidéo durait 40 ms. Le temps 0 pour le calcul des PE était pris au début du son.

	/pa/	/pi/	/po/	/pu/	moyenne
6 trames	21%	31%	10%	43%	26%
8 trames	45%	74%	23%	51%	50%
14 trames	81%	81%	37%	27%	53%

TAB. 9.1 – Résultats de la pré-expérience comportementale. Les valeurs indiquent le pourcentage de reconnaissance en fonction du type de syllabe et du nombre de trames présentées

servée lorsque la syllabe était présentée dans sa totalité. L'analyse des erreurs de cette expérience a, par ailleurs, montré que les syllabes /po/ et /py étaient souvent confondues.

9.2.3 Procédure

Les 3 exemplaires des 4 syllabes étaient présentées de façon auditive, visuelle ou audiovisuelle. Tous les essais étaient présentés aléatoirement dans un même bloc de stimuli. Au total, 1116 stimulations étaient présentées, réparties en 16 blocs d'une durée approximative de 2 min 30. Au début de chaque bloc, l'une des 4 syllabes était désignée comme cible (chaque syllabe pouvait donc être cible ou non-cible selon le bloc). Le sujet devait répondre en appuyant sur un bouton lorsqu'il entendait la syllabe cible (seulement pour les essais audiovisuels et auditifs).

Nous avons longuement hésité à demander aux sujets de détecter la cible quelle que soit sa modalité de présentation, y compris en condition visuelle seule, c'est-à-dire en lisant sur les lèvres. Dans ce cas, nous aurions pu lier plus directement les résultats des potentiels évoqués aux résultats comportementaux calculés à partir de l'inégalité de Miller, qui prend en compte les TR auditifs, visuels et audiovisuels et exclut un simple effet de facilitation statistique des TR. Cependant le fait de demander aux sujets une réponse dans les trois conditions aurait nécessité un effort attentionnel plus important en condition de lecture labiale que dans les deux autres modalités. Les effets de cette attention visuelle sur la réponse évoquée auraient pu se manifester de manière plus importante dans la condition visuelle seule, que dans la condition audiovisuelle et auraient donc pu apparaitre de manière

erronée comme des violations de l'additivité (voir aussi Besle, Fort & Giard, 2004, et la partie 7.2.1 page 108 pour une discussion plus détaillée). Nous avons donc demandé au sujet de ne répondre que sur la base des indices auditifs. Les sujet devaient cependant fixer la bouche durant toute l'expérience, et ceci était vérifié grâce à une caméra vidéo.

9.2.4 Expérience comportementale complémentaire

Pour appliquer l'inégalité de Miller et vérifier l'existence d'un gain comportemental audiovisuel, nous avons donc mené une expérience comportementale complémentaire avec un autre groupe de sujets. Les stimuli et les conditions de stimulation étaient identiques, excepté que les sujets devaient répondre dans les 3 conditions de stimulation. Dans cette expérience, seules les syllabes /pa/ et /pi/, plus faciles à discriminer visuellement, pouvaient être cible. Cette expérience complémentaire permettra a minima de conclure que les stimuli utilisés pour le calcul des interactions audiovisuelles, au moyen du modèle additif, sont susceptibles de donner lieu à un effet de facilitation audiovisuelle qui n'est pas dû à une facilitation statistique.

9.2.5 Analyse des résultats

Les TR auditifs et audiovisuels dans l'expérience d'EEG ont été comparés par un test de Student. Les TR auditifs, visuels et audiovisuels dans l'expérience comportementale complémentaire ont été analysés par application de l'inégalité de Miller et comparaison des fractiles de distribution des TR audiovisuels et de la somme des distributions des TR unimodaux (voir la partie 7.1.4 page 105).

Les PE n'ont été calculés que sur les essais non-cibles, pour exclure toute activité motrice dans les signaux analysés. Les essais où le sujet avait répondu par erreur ont également été exclus de l'analyse des PE. Le nombre moyen d'essais était de 160 par condition et par sujet. Le temps zéro utilisé pour le moyennage des PE et à partir duquel étaient mesurées les latences correspondait au début de la syllabe auditive. La ligne de base était prise entre -300 et -150 ms pré-stimulus. Cette fenêtre de latence est un compromis entre la nécessité de rapprocher le plus possible la ligne de base de la fenêtre d'analyse, et celle d'éviter l'inclusion de potentiels évoqués visuels dus au mouvement des lèvres qui commençait 240 ms avant la présentation du son (bien que ces mouvements soient de très faible amplitude).

La différence entre le PE audiovisuel et la somme des PE unimodaux statistiquement a été testée à chaque échantillon temporel et à chaque électrode par un test de Student apparié, dans les 200 premières millisecondes post-stimulus. Les tests multiples ont été pris en compte en exigeant, pour chaque électrode, un nombre minimal de 24 échantillons significatifs successifs, d'après la table proposée par Guthrie et Buchwald (1991, voir la partie 8.1 page 111).

9.3 Résultats

9.3.1 Résultats comportementaux

En ce qui concerne l'expérience en EEG, les sujets ont été plus rapides pour répondre aux cibles audiovisuelles (400 ms) qu'aux cibles auditives (423 ms). Bien que cette différence soit assez faible, elle était très significative ($t(15) = 4,33$; $p < 0,001$). Le pourcentage d'erreurs (oublis ou fausses alertes) était inférieur à 1% dans les deux conditions.

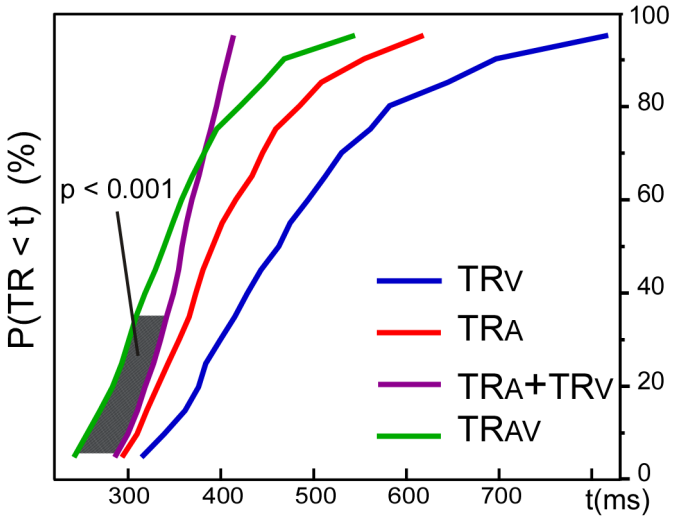

FIG. 9.2 – Application de l'inégalité de Miller. TR_V : fonction de répartition des temps de réaction visuels ; TR_A : fonction de répartition des temps de réaction auditifs ; $TR_A + TR_V$: somme des 2 fonctions de répartition unimodales ; TR_{AV} : fonction de répartition des temps de réaction audiovisuels. La partie hachurée désigne les zones où les fractiles correspondants des deux fonctions de répartition sont significativement différents.

Concernant l'expérience comportementale complémentaire, les TR dans les conditions visuelle, auditive et audiovisuelle étaient respectivement de 496, 418 et 356 ms. La figure 9.2 montre les fonctions de répartition (pour l'ensemble des sujets) des TR visuels, auditifs et audiovisuels, ainsi que la somme des fonctions de répartition unimodales. Pour les 9 premiers fractiles, les TR bimodaux sont significativement inférieurs à ceux prédits par les modèles d'activations séparées et représentés par la somme des fonctions de répartition unimodales.

9.3.2 Résultats électrophysiologiques

la figure 9.3.A (page suivante) montre les PEs obtenus dans chaque modalité dans les 300 premières millisecondes. La réponse visuelle unimodale (courbe bleue) montre principalement un pic négatif vers 40 ms, dont le maximum se situe sur les électrodes occipitales et forme une topographie occipitale bilatérale (non illustrée). La topographie et la latence de cette onde suggèrent qu'il pourraient s'agir de l'onde N1 visuelle dont le pic est habituellement observé vers 180 ms. Dans notre cas, ce pic suivait le début du mouvement des

lèvres de 280 ms, ce qui pourrait s'expliquer par le début très progressif des mouvements, et donc à la fois un temps de traitement plus lent à s'établir et une plus grande variabilité des réponses élémentaires, qui auraient pour effet d'étaler cette composante dans le temps. Il se peut aussi qu'il s'agisse d'une composante spécifique au traitement d'un mouvement. Cette réponse était suivie d'une deuxième composante visuelle négative dont le maximum se situait de façon bilatérale sur les électrodes pariéto-centrale, vers 160 ms post-stimulus.

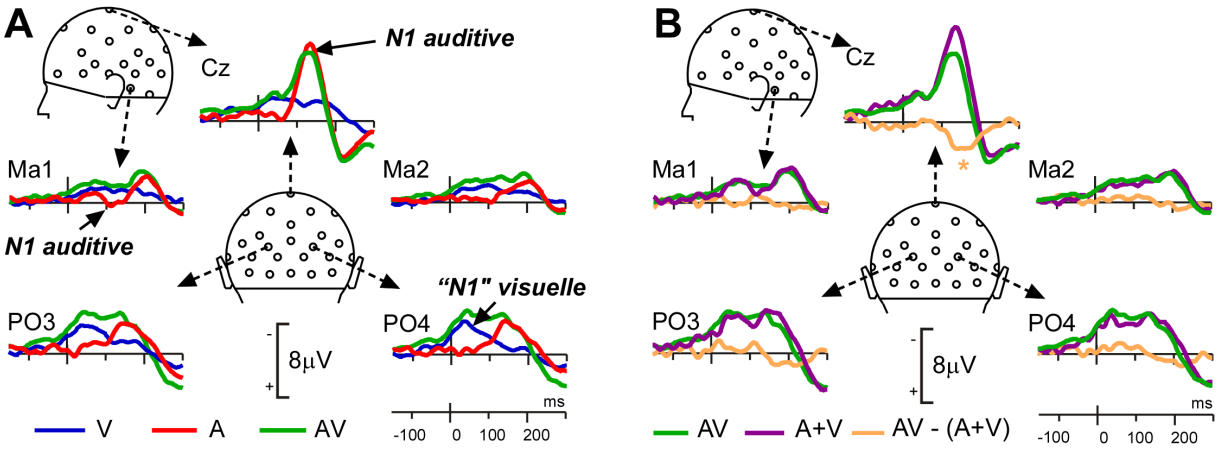

FIG. 9.3 – A. Réponses évoquées par les syllabes auditives (A), visuelles (V) et audiovisuelles (AV) entre -150 et 300 ms à un sous-ensemble d'électrodes. B. Application du modèle additif. A+V : somme des réponses auditives et visuelles. AV-A+V : violation du modèle additif. L'étoile indique les violations significatives au seuil corrigé.

La réponse auditive unimodale (courbe rouge) se caractérisait par une onde négative dont le pic maximum vers 135 ms était associé à une inversion de polarité sur les électrodes mastoïdes. La topographie de cette onde ainsi que celles des densités radiales de courant associées sont visibles sur la figure 9.5 (1re colonne, page 126). C'est une topographie typique d'activités prenant place dans le cortex auditif. Cette onde correspond sans nul doute à l'onde N1 auditive. L'onde N1 était suivie d'un onde de polarité inverse (l'onde P2) dont l'amplitude était maximale à 205 ms post-stimulus.

La figure 9.3.B compare les PE audiovisuels (courbe verte) à la somme des PE unimodaux (courbe mauve). Ces deux courbes sont globalement identiques excepté sur les électrodes fronto-centrales entre 100 et 200 ms, c'est-à-dire dans une fenêtre de temps correspondant à l'onde N1 auditive et à la deuxième composante visuelle. Les résultats détaillés du test statistique du modèle additif sont donnés dans la figure 9.4 page ci-contre. On peut constater que la différence entre la réponse bimodale et la somme des réponses unimodales est significative sur une grande partie des électrodes fronto-centrales d'environ 120 ms à 200 ms post-stimulus et que le nombre d'échantillons significatifs successif dépasse largement 24 ms sur ces électrodes, ce qui permet d'exclure un effet dû au nombre important de tests réalisés. La topographie des interactions audiovisuelles était à peu près stable entre 120 et 190 ms post-stimulus.

Pour tenter de comprendre la nature de ces interactions audiovisuelles, nous avons comparé (figure 9.5 page 126) leur topographie à la topographie des réponses unimodales, à la latence où la violation du modèle additif était la plus significative, ce qui correspond

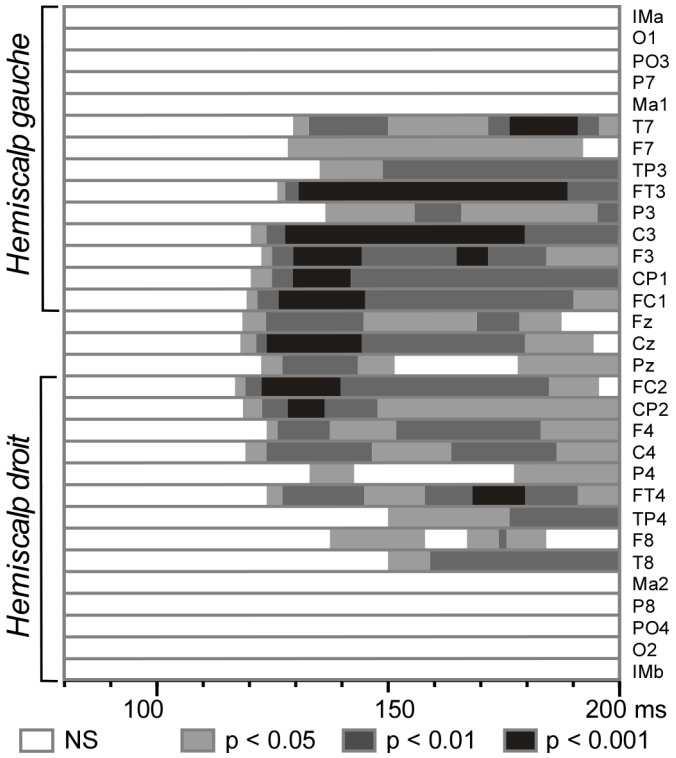

FIG. 9.4 – Résultat des tests statistiques de la violation du modèle additif sur l'ensemble des électrodes entre 80 et 200 ms. Le niveau de gris indique la significativité

au pic de l'onde N1 auditive. La topographie des interactions ressemble clairement plus à celle de l'activité auditive unimodale qu'à celle de l'activité visuelle. En particulier, la configuration des puits et des sources de courant reproduit assez fidèlement celle de l'onde N1 auditive, avec des polarités inversées. Cela suggère que les interactions audiovisuelles observées autour de 135 ms reflètent une diminution d'activité des générateurs de l'onde N1 auditive dans la condition audiovisuelle par rapport à la condition auditive seule.

9.4 Discussion

9.4.1 Comportement

Les résultats comportementaux de l'expérience d'EEG, ainsi que ceux de l'expérience comportementale, montrent que le traitement de la parole peut être accéléré par des indices visuels, même lorsque la performance des sujets a atteint un plafond en termes de pourcentage de réponses correctes. À notre connaissance, c'est la première fois qu'un tel résultat est montré. Peu d'études se sont en fait intéressé aux temps de réactions à des stimuli de parole audiovisuelle. Deux études ont mesuré les TR auditifs et audiovisuels dans un tâche de catégorisation de syllabes commençant par des consonnes différentes, mais elles ont soit rapporté des différences faibles et non reproductibles (Massaro & Cohen, 1983), soit des TR audiovisuels supérieurs à l'un des TR unimodaux (K. P. Green & Gerdeman, 1995).

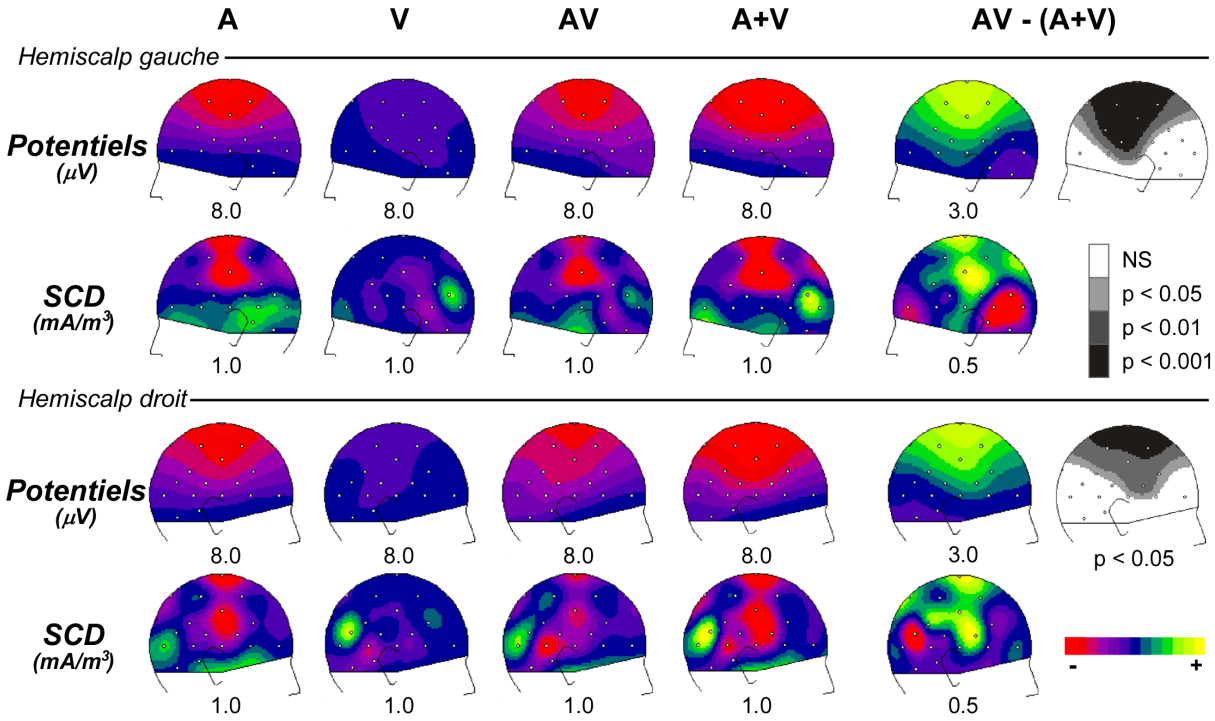

FIG. 9.5 – Topographies des réponse auditives (A), visuelles (V), audiovisuelles (AV), ainsi que de la somme des réponses unimodales (A+V) et de la violation du modèle additif (AV - (A+V)) sur les hémiscalps droit et gauche. La valeur maximale de l'échelle de couleur est indiquée sous chaque carte. La couleur jaune correspond aux potentiels ou aux courants positifs, tandis que la couleur rouge correspond aux potentiels ou aux courants négatifs. Les deux cartes en niveaux de gris à droite donnent la significativité des PE d'interaction (AV-(A+V)). SCD : Densité radiale de courant.

Une étude précédente, rapportée par Massaro (1987), avait cependant testé un modèle d'activations séparées chez deux sujets (sous l'hypothèse d'indépendance des distributions des TR unimodaux, voir la partie 7.1.1 page 101). Les auteurs trouvaient des TR audiovisuels inférieurs en moyenne aux TR unimodaux mais qui restaient prédictibles par le modèle. Dans notre expérience complémentaire, nous montrons au contraire que le gain en temps de réaction observé avec les même stimuli que ceux utilisés dans notre expérience d'EEG ne peut s'expliquer par un tel modèle. Ce résultat implique (dans les limites exposées dans la partie 7.1 page 99) que les canaux auditifs et visuels ont échangé des informations.

On notera cependant que le gain de temps de réaction dans la condition audiovisuelle par rapport à la condition auditive seule est beaucoup plus important dans l'expérience comportementale que dans l'expérience électrophysiologique. Le fait d'avoir utilisé des groupes de sujets différents limite les conclusions que l'on peut tirer de ce résultat. Il est cependant probable que le fait d'attirer l'attention des sujets vers les indices visuels en leur demandant d'effectuer une tâche de lecture labiale ait augmenté la contribution des indices visuels au traitement de l'identité de la syllabe.

Il faut aussi souligner que la tâche de l'expérience d'EEG s'apparente plus à paradigme du stimulus accessoire dans lequel le sujet n'a pas à analyser les informations visuelles pour discriminer les syllabes auditives. On ne peut donc exclure le fait que le gain de TR en condition audiovisuelle représente un effet d'alerte dû à la présence d'un stimulus visuel, d'autant plus que le mouvement des lèvres précédait la syllabe auditive.

9.4.2 Résultats électrophysiologiques

Ce gain comportemental pour la discrimination de syllabes était associé à ce que l'on a interprété comme une diminution d'activité des générateurs de l'onde N1 auditive. Avant de tenter d'interpréter cette diminution et son rôle dans l'intégration des indices auditifs et visuels de parole, soulignons que des résultats analogues ont été rapportés dans la littérature soit en même temps, soit à la suite de notre étude.

Ainsi, Klucharev, Möttönen et Sams (2003) ont testé le modèle additif pour des syllabes auditives, visuelles et audiovisuelles et ont trouvé des interactions audiovisuelles vers 125 ms de traitement, dont la topographie radiale suggère la diminution de certaines composantes seulement, de l'onde N1 auditive (notons que les sujets réalisaient des tâches différentes dans des blocs de stimulations auditifs, audiovisuels et visuels séparés, ce qui pose quelques problèmes pour l'application du modèle additif, voir la partie 7.2.1 page 108).

van Wassenhove, Grant et Poeppel (2005) ont mis en évidence une diminution d'amplitude importante de l'onde N1 auditive en condition audiovisuelle par rapport à un condition auditive seule, dans une tâche de discrimination phonologique, dans des blocs séparés pour les différentes conditions auditives, visuelles et audiovisuelles. Cette diminution d'amplitude était doublée d'une diminution de latence, difficile à interpréter, cependant, en l'absence d'utilisation du modèle additif.

En MEG, Möttönen, Schurmann et Sams (2004) ont montré la même diminution. Si l'utilisation de la MEG limite en pratique le besoin de recourir au modèle additif en raison de la moindre diffusion des champs magnétiques sur le scalp, ces auteurs ne présentent toutefois pas de réponses visuelles seules permettant de s'assurer que la diminution observée était bien due à une modulation de l'activité auditive.

Une étude de Miki, Watanabe et Kakigi (2004) en MEG n'a en revanche pas rapporté une telle diminution. Plusieurs raisons peuvent expliquer cette absence, comme par exemple le fait que les sujets étaient totalement passifs ou que les mouvements de lèvres consistaient simplement en la présentation d'une image de bouche ouverte et non en de véritables mouvements filmés. Malgré les nombreux problèmes méthodologiques que présentent ces études, elle convergent presque toutes vers le même résultat, ce qui suggère que l'effet trouvé est assez robuste.

Une telle diminution de l'onde N1 auditive ne semble pas exister dans des expériences de discrimination ou de détection de stimuli non langagiers dans lesquels une diminution du TR audiovisuel était observé (Fort et coll., 2002a, 2002b ; Giard & Peronnet, 1999 ; Molholm et coll., 2002 ; Teder-Sälejärvi et coll., 2002). Cette effet pourrait donc bien être spécifique de l'intégration audiovisuelle des indices de parole, ou plus généralement

d'évènements bimodaux dans lesquels le stimulus visuel précède le stimulus auditif (notons cependant que Möttönen et coll., 2004 trouvent une diminution de l'onde N1 auditive avec des stimuli de paroles auditifs et visuels dont les débuts sont synchrones).

En revanche, une diminution de l'onde N1 visuelle (vers 180 ms de latence) a été trouvée pour la discrimination de stimuli audiovisuels par rapport à des stimuli visuels seuls (Giard & Peronnet, 1999). Cette onde, générée dans le cortex visuel extrastrié (Mangun, 1995) serait liée à des processus de discrimination visuelle (Vogel & Luck, 2000). Cette réduction avait été interprétée comme le reflet d'une demande énergétique moindre pour discriminer les stimuli visuels, rendu plus saillants par la présence et l'utilisation d'informations auditives. De la même manière, l'onde N1 auditive serait liée à l'analyse séparée des traits acoustiques du stimulus dans le cortex auditif (Näätänen & Picton, 1987 ; Näätänen & Winkler, 1999). La diminution observée pourrait donc refléter la facilitation de traitement des syllabes auditives due à la présence d'informations phonétiques visuelles, à une latence où les différents traits acoustiques n'ont pas encore abouti à une représentation intégrée du stimulus sonore (Näätänen & Winkler, 1999).

Bien que traditionnellement, on situe les générateurs de l'onde N1 auditive dans le cortex auditif, c'est-à-dire sur la partie supérieure du cortex temporal, il est possible que l'onde N1 en réponse à des sons de parole, beaucoup moins étudiée, inclue d'autres générateurs. Plusieurs études ont montré l'implication du STS dans le traitement de sons complexes, avec une préférence pour les sons de parole, qu'ils soient intelligible ou non (revue dans Hickok & Poeppel, 2004). Étant donné que le STS a été impliqué dans plusieurs études de neuroimagerie sur l'intégration audiovisuelles des indices de parole (Beauchamp, Argall et coll., 2004 ; Calvert et coll., 2000 ; Wright et coll., 2003) et que son orientation est parallèle au plan supratemporal, c'est un candidat possible pour la localisation de l'effet observé. Toutefois, le fait que tous les générateurs, visibles sur la carte des densités radiales de courant de l'onde N1 auditive, apparaissent également sur la topographie des interactions suggère qu'il s'agit d'une diminution globale de l'activité auditive seule à cette latence et non d'un seul générateur spécifique au traitement de la parole.

Plusieurs interprétations alternatives de l'effet observé peuvent être proposées. Tout d'abord, cette diminution pourrait refléter une facilitation du traitement due à une meilleure préparation du sujet pour traiter les indices auditifs lorsque ceux-ci ont été précédés de mouvements lui indiquant qu'un son va peut-être lui être présenté dans les 240 ms. En effet, bien que la violation de l'inégalité de Miller suggère l'existence d'échanges d'informations auditives et visuelles, elle a été appliquée à des TR enregistrés dans des conditions différentes qui font qu'on ne peut exclure un pur effet d'alerte dans l'expérience d' EEG. Si tel était le cas, cependant, on s'attendrait à observer plutôt une augmentation de l'onde N1 auditive, analogue aux effets d'attention auditive qui se manifestent sur plusieurs ondes sensorielles auditives, dont l'onde N1 (revue dans Näätänen, 1992 ; Giard, Fort, Mouchetant-Rostaing & Pernier, 2000). De façon intéressante, si des effets d'un indice visuel spatial sur les potentiels évoqués auditifs ont été mis en évidence (McDonald, Teder-Sälejärvi, Heraldez & Hillyard, 2001), ils prennent la forme d'une négativité accrue à la latence de nos effets. De tels effets auraient donc résulté en une augmentation de l'onde N1 auditive. Il n'est toutefois pas dit que les effets d'alerte se manifestent de la même

manière que les effets d'attention spatiale sur les PE auditifs, même si certaines études indiquent des effets analogues pour les deux phénomènes sur la réponse visuelle dans le cortex extrastrié (Thiel, Zilles & Fink, 2004).

Ensuite, plusieurs études ont montré que la lecture labiale pouvait, en elle-même, activer le cortex auditif (par exemple Calvert et coll., 1997). Même si les sujets n'avaient pas pour tâche de lire les syllabes sur les lèvres, certains sujets ont rapporté avoir tenté de le faire. Et, quoiqu'il en soit, la vision des mouvements articulatoires, même sans tentative d'en comprendre le contenu pourrait également activer le cortex auditif en condition visuelle seule. Cette activation du cortex auditif par les mouvements labiaux, si elle avait lieu à la latence de l'effet observé, pourrait apparaitre comme une violation du modèle additif et expliquer la topographie auditive des interactions prenant place entre 120 et 190 ms. Cette explication est cependant très peu probable dans la mesure où l'on n'observe pas de réponse ayant une topographie auditive dans cette fenêtre de latence dans la condition visuelle seule.

Une autre explication enfin serait que les informations visuelles sur l'identité de la syllabe sont disponibles avant les informations auditives, par un phénomène de coarticulation. Ces informations pourraient alors pré-activer des unités phonologiques dans le cortex auditif (ou le STS). Plusieurs expériences ont montré que l'amorçage sémantique, aussi bien unimodal qu'intermodal, pouvait se manifester au niveau neuronal par des diminutions d'activité (Badgaiyan, Schacter & Alpert, 1999 ; Holcomb & Anderson, 1993 ; Holcomb & Neville, 1990). De façon analogue, la diminution d'activité dans le cortex auditif pourrait refléter un effet d'amorçage des informations phonétiques visuelles sur le traitement phonétique ou phonologique auditif (voir aussi Jaaskelainen et coll., 2004). Bien que nous ayons montré dans une pré-expérience que les informations visuelles au moment de l'arrivée du son étaient insuffisantes pour identifier les syllabes (voir la partie 9.2.2 page 120), le traitement intégral de l'amorce n'est pas nécessaire pour observer des effets d'amorçage. Il est toutefois probable que les informations visuelles présentes avant l'ouverture complète de la bouche soient trop subtiles pour participer à l'amélioration audiovisuelle. Munhall, Kroos, Jozan et Vatikiotis-Bateson (2004) ont en effet montré que les fréquences spatiales des informations visuelles participant à l'amélioration de l'intelligibilité de la parole dans le bruit sont assez grossières (inférieures à 7 cycles/visage).

Si des informations phonétiques visuelles ont permis de moduler l'activité auditive de traitement des syllabes, ce sont sans doute celles portées par la forme de l'ouverture de la bouche, qui sont disponibles au même moment que les informations auditives. Dans ce cas, les informations visuelles mettent environ 100 ms à venir moduler l'activité dans les structures traitant la parole auditive.

Chapitre 10

Étude en sEEG

10.1 Introduction

Notre expérience en EEG de scalp a montré l'existence d'importants effets d'interactions audiovisuelles dans la perception de la parole bimodale entre 120 et 190 ms de traitement de la syllabe, reflétant vraisemblablement une diminution d'activité auditive. Contrairement aux études précédentes utilisant le modèle additif dans l'effet du stimulus redondant avec des stimuli non-langagiers et qui avaient mis en évidence des interactions complexes, de topographies différentes à différentes latences, nous n'avons trouvé que cet effet de modulation de l'activité auditive. Il était cependant possible que l'amplitude importante de l'effet de diminution de l'onde N1 auditive ait caché d'autres effets d'interaction dans d'autres structures. Par ailleurs la résolution spatiale limitée de l'EEG de scalp ne permettait pas de s'assurer de la localisation exacte de la diminution d'activité. Celle-ci aurait pu avoir lieu aussi bien dans le planum temporale que sur l'une des aires bordant le STS.

Afin d'étudier plus en détail les interactions audiovisuelles ayant lieu lors de la perception de syllabes bimodales, nous avons fait passer cette expérience à des patients épileptiques portant des électrodes intracérébrales, en collaboration avec O. Bertrand (U821) et le Docteur C. Fischer (Hôpital Neurologique de Lyon). La plupart de ces patients étaient suivis pour des épilepsies d'origine temporale et avaient donc un certain nombre d'électrodes traversant le planum temporale, le gyrus de Heschl, le gyrus temporal supérieur (GTS), le STS et le gyrus temporal moyen (GTM) (ces structures sont indiquées sur la figure 10.4.B, page 138). À l'occasion, d'autres structures ont pu être explorées (insula, gyrus supra-marginal, opercules pré-central et post-central, gyrus temporal moyen postérieur, etc...). Les emplacements de toutes les électrodes de tous les patients ont été reportées sur un cerveau commun dans les figures 10.1 page suivante et 10.2 page 133.

Bien que nous n'ayons pas observé d'activation de type auditif en réponse aux mouvements labiaux présentés isolément dans l'expérience d'EEG, les enregistrements sEEG constituaient aussi une occasion de vérifier l'existence de traitements des indices visuels de parole dans le cortex auditif. La plupart des études d'IRMf ayant étudié la lecture labiale ont montré l'implication, entre autres structures corticales, d'une partie importante du cortex auditif Il existe cependant un débat concernant l'implication du cortex auditif

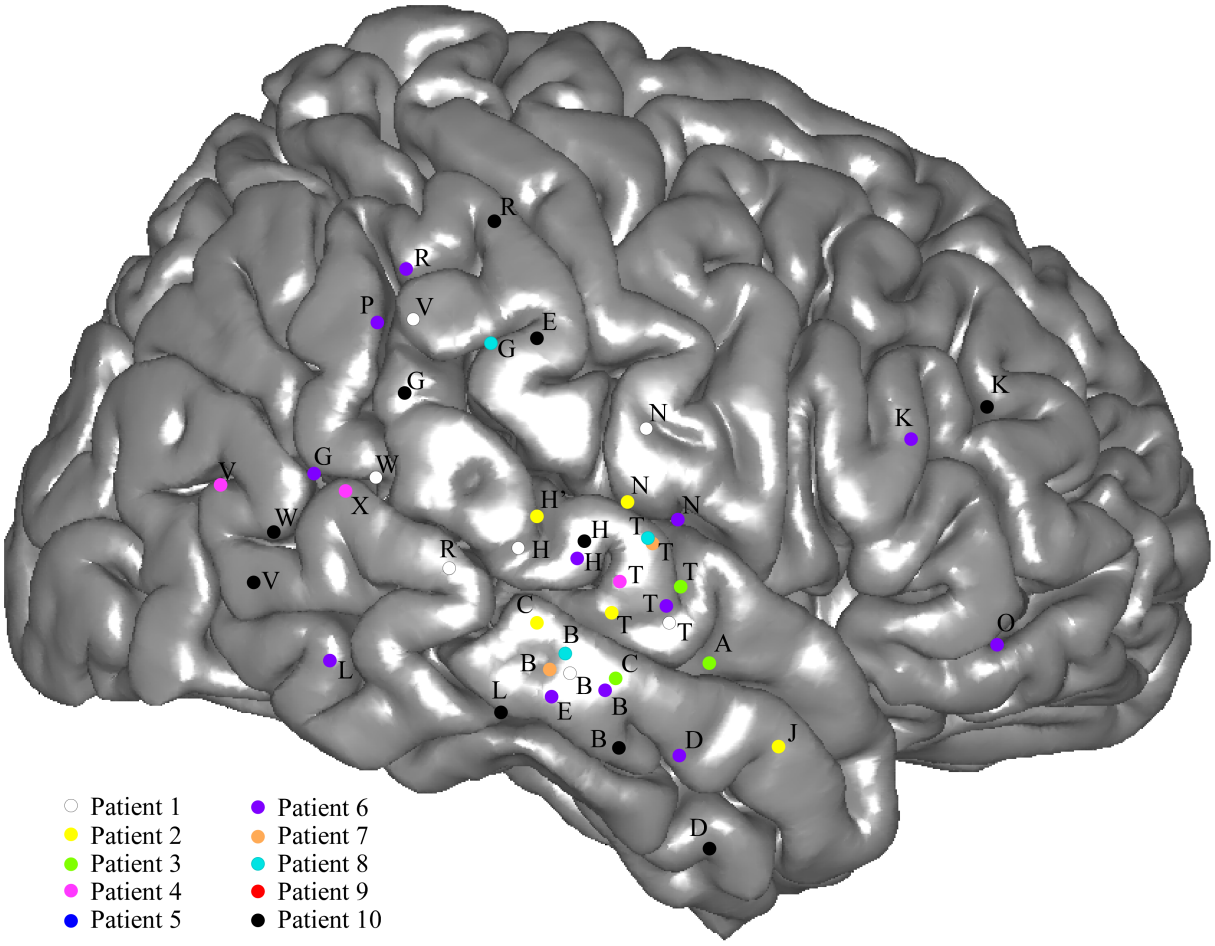

FIG. 10.1 – Emplacements des électrodes de l'hémisphère droit reportés à la surface d'un cerveau standard (single-subject du MNI). Le recalage des électrodes des différents patients a été réalisé par la méthode de Talairach (transformation linéaire par cadrans). Chaque électrode comprend entre 5 et 15 contacts explorant les structures situées à la perpendiculaire du plan de la figure

primaire (aire 41 de Brodmann) dans cette activation. Certaines études ont montré une activation de la partie médiale du gyrus transverse (ou gyrus de Heschl), où se situe le cortex auditif primaire (Calvert et coll., 1997 ; Ludman et coll., 2000 ; MacSweeney et coll., 2001). D'autres ont trouvé une activation de sa partie latérale (Calvert & Campbell, 2003), qui ne correspond déjà plus au cortex primaire ou seulement des cortex secondaires (L. E. Bernstein et coll., 2002 ; Campbell et coll., 2001 ; MacSweeney et coll., 2000, 2002 ; Olson et coll., 2002 ; Paulesu et coll., 2003), dont le planum temporale (aire 42) et le GTS latéral (aire 22). Le cortex auditif primaire étant une structure de petite taille, la variabilité anatomique inter-individuelle est cependant susceptible de cacher des activations dans une étude de groupe et c'est seulement récemment qu'une étude a défini, chez chaque sujet, les zones activées par la lecture labiale d'une part et la position anatomique du gyrus transverse d'autre part : chez 7 sujets sur 10, une activation du cortex auditif primaire a été trouvée (Pekkola et coll., 2005).

Un autre débat concerne la signification fonctionnelle de cette activation. Ainsi l'ac-

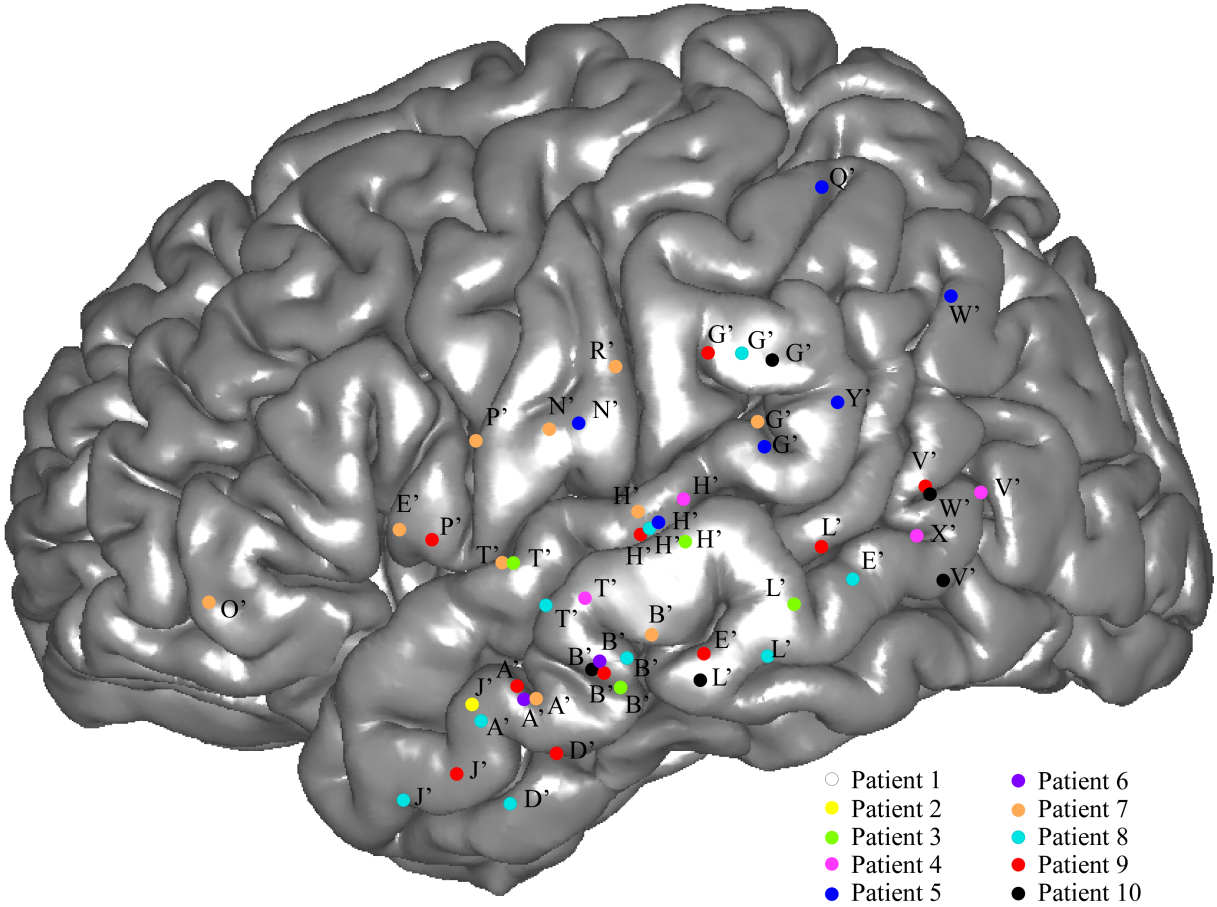

FIG. 10.2 – Emplacements des électrodes de l'hémisphère gauche reportés à la surface d'un cerveau standard (single-subject du MNI). Le recalage des électrodes des différents patients a été réalisé par la méthode de Talairach (transformation linéaire par cadrans). Chaque électrode comprend entre 5 et 15 contacts explorant les structures situées à la perpendiculaire du plan de la figure

tivation du cortex auditif (primaire ou secondaire) pourrait correspondre à de l'imagerie auditive et avoir lieu à une latence tardive : la vision des articulateurs pourrait activer des représentations phonologiques et l'accès à ces représentations permettraient aux sujets d'imaginer les sons de parole correspondant, ce qui pourrait activer le cortex auditif. Certains éléments suggèrent cependant que ce scénario est peu plausible : en effet deux études ont montré une activation du cortex auditif par des mouvements labiaux ressemblant à de la parole mais ne correspondant à aucun mot ou son connu (phonèmes étrangers : Calvert et coll., 1997 ou vidéos passées à l'envers : Paulesu et coll., 2003). Il se pourrait donc que le cortex auditif (primaire ou secondaire) participe au décodage phonologique de la parole visuelle comme il participe à celui de la parole auditive.

La résolution temporelle de la sEEG, ainsi que sa résolution spatiale devraient permettre d'apporter des éléments concernant la signification fonctionnelle des activations du cortex auditif, en donnant la latence d'activation de ses différentes parties, ainsi qu'une preuve directe de l'implication ou non du cortex auditif primaire dans la lecture labiale.

Patient	1	2	3	4	5	6	7	8	9	10	moyenne
V	91	130	125	120	135	81	103	79	92	82	104
A	87	128	130	130	144	84	103	90	81	81	106
AV	92	129	130	135	142	81	106	87	94	77	107

TAB. 10.1 – Nombre d'essais pris en compte pour le calcul des potentiels évoqués et des tests statistiques. V : condition visuelle. A : condition auditive. AV : condition audiovisuelle.

10.2 Méthodes

10.2.1 Patients

10 patients ont participé à cette étude. Aucun de ces patients ne souffrait de troubles auditifs (excepté le patient 1 qui était capable de lire sur les lèvres) ou visuels.

10.2.2 Stimuli et procédure

Les stimuli, la procédure et la tâche des patients étaient identiques à ceux employés dans l'étude d'EEG de scalp, excepté que seuls 8 blocs de 66 stimuli (d'une durée de 2 minutes 15 chacun) étaient présentés. Le nombre total de stimuli non-cibles présentés était de 150 dans chacune des conditions de présentation.

Pour 6 des patients (patients 5 à 10), nous avons ajouté des essais audiovisuels incongruents. Les résultats pour cette condition expérimentale ne seront pas rapportés ici. Afin de ne pas rallonger la durée de l'expérience, le nombre total de stimuli était identique avec et sans syllabe incongruentes, si bien que le nombre d'essais moyen par condition pour les 6 derniers patients était diminué d'un quart (108 essais par condition expérimentale).

10.2.3 Calcul des potentiels évoqués

Les méthodes de calcul des PE intracérébraux ayant été exposées dans la partie 6.4 page 92, nous nous contenterons de rappeler que les essais comprenant des valeurs d'amplitude, supérieures en valeur absolue à 5 écart-types de la distribution des amplitudes sur l'ensemble des essais dans une condition donnée, étaient rejetés avant le moyennage, afin d'éviter la contamination des données par les pointes inter-critiques. Le nombre d'essais retenus après rejet des artéfacts pour l'analyse par conditions et par patients est donné dans la table 10.1.

Les contacts qui participaient à plus de 6% de rejet étaient considérés comme mauvais et exclus de l'analyse. Le nombre de contacts retenus par patients après rejet des artéfacts est donné dans la table 10.2 page suivante. Pour les tests d'émergences des activitis unisensorielles, nous n'avons pas appliqué cette contrainte (voir la partie 10.2.4 page ci-contre).

Rappelons également que, comme pour l'étude en EEG de surfaces, le temps 0 pour le calcul des PE correspondait au début de la syllabe auditive, et que la ligne de base était prise entre -300 et -150 ms.

Patient	1	2	3	4	5	6	7	8	9	10	moyenne
Modèle additif	63	45	63	63	63	65	40	62	51	42	56
A ou V vs 0	63	63	63	63	63	127	124	127	112	127	93

TAB. 10.2 – Nombre de contacts considérés pour les tests statistiques. A ou V vs 0 : test d'émergence

10.2.4 Analyses statistiques

Pour les données comportementales, nous avons comparé le TR pour les syllabes auditives et les syllabes audiovisuelles cibles, pour chaque patient et au niveau du groupe. Les TR moyens de chaque patient étaient comparés par un test de Student pour groupes indépendants et les TR moyens du groupe étaient comparés par un test de Student pour mesures appariées.

Pour tous les tests statistiques portant sur les données électrophysiologiques, le signal a été sous-échantillonné à 50 Hz, l'amplitude à un échantillon temporel donné étant égal à la moyenne du signal dans une fenêtre de 40 ms autour de cet échantillon.

Pour le calcul des interactions audiovisuelles, nous avons testé la violation du modèle additif à chaque échantillon temporel de chaque contact retenu pour l'analyse entre 0 et 200 ms (20 échantillons temporels ; les tests ont en fait été réalisés entre -300 et 600 ms après le stimulus auditif, mais nous ne considèrerons que les violations du modèle additif qui commençaient avant 200 ms post stimulus, voir la partie 7.2.1 page 108). Le nombre moyen de contacts retenus par patient était de 56 (voir la table 10.2), ce qui donne un total de 1120 tests en moyenne par patient.

Les tests multiples étaient pris en compte indépendamment pour chaque patient, dans les dimensions spatiales et temporelles. Dans la dimension temporelle, nous avons utilisé la méthode du minimum d'échantillons consécutifs significatifs (voir la partie 8.1 page 112). Pour tenir compte des tests multiples dans la dimension des capteurs, nous avons appliqué la correction de Bonferroni (voir la partie 8.1 page 111) et exigé que les violations du modèle additif soient significatives à $p<0{,}001$, ce qui correspond à un seuil classique de 0,05 divisé par 50 (le nombre approximatif de contacts par patient). En réalité ce seuil est sans doute trop conservateur car les signaux enregistrés sur des contacts voisins sont souvent corrélés (mais pas toujours, en particulier dans le cas de gradients locaux importants).

Pour l'analyse des réponses visuelles seules et des réponse auditives seules, nous n'avons considéré que les réponses qui différaient significativement de la ligne de base. La significativité était testée par un test non paramétrique apparié (test de Wilcoxon) à chacun des échantillons entre -150 et 600 ms (38 échantillons temporels). À la différence du test du modèle additif, l'émergence des réponses sensorielles a été testée sur l'ensemble des capteurs enregistrés et pas seulement sur ceux conservés lors du rejet des artéfacts, de façon à augmenter l'échantillonnage spatial et parce que l'on s'attend à observer des effets moins sensibles au bruit dans ce cas. Le nombre de tests réalisés par patient était donc en moyenne de 38 échantillons $\times$ 93 capteurs, c'est-à-dire environ 3500 tests. Pour ces tests, je n'ai pas eu le temps d'implémenter la méthode du minimum d'échantillons significatifs consécutifs, nous avons donc corrigé le seuil de significativité par la méthode

de Bonferroni dans les dimensions temporelles et spatiales, c'est-à-dire utilisé une seuil égal à $0,05/3500 = 1,4 \times 10^{-5}$. La conservativité de cette approche est cependant moins problématique ici que dans le cas du modèle additif car les effets sont de manière générale plus robustes.

Afin de localiser le cortex auditif primaire, nous avons en particulier recherché les premières réponses sensorielles auditives corticales qui apparaissent à partir de 10-15 ms. Ces réponses étant des réponses transitoires rapides, les tests de significativité (test de Wilcoxon par rapport à la ligne de base) ont été menés sur les données échantillonnées à 1000 Hz (512 Hz pour la seconde moitié des sujets) entre 10 et 40 ms (respectivement 30 et 15 échantillons), sur les électrodes traversant les gyrus temporal supérieur. Pour ce test, un seuil de $p < 10^{-5}$ était suffisant.

Tous les tests ont été menés à la fois sur les données monopolaires et les données bipolaires. Mais seules les données bipolaires ont été prises en considération pour l'application des critères statistiques, de manière à pouvoir attribuer l'effet à la région traversée par le contact concerné (en particulier pour l'analyse de groupe). Les données monopolaires n'étaient donc utilisées que pour la description et l'interprétation des résultats (excepté dans un cas, qui sera signalé).

Dans tous les cas, lorsqu'un effet (violation du modèle additif ou émergence de la réponse unisensorielle) remplissait les critères statistiques requis, c'est l'ensemble de l'effet présentant une unité spatiale et temporelle qui était pris en compte dans l'interprétation, même s'il ne remplissait pas les critères à tous les contacts et à tous les échantillons temporels concernés. En d'autres termes, lorsqu'un effet était significatif sur un certain nombre d'échantillons consécutifs et sur un certain nombre de contacts voisins, il suffisait qu'au moins un échantillon remplisse les critères, pour que cet effet soit retenu et/ou décrit dans son intégralité.

10.3 Résultats

10.3.1 Données comportementales

La figure 10.3 page suivante montre les temps de réactions des 10 patients pour les syllabes auditives et audiovisuelles. En moyenne, les TR étaient plus rapides en condition audiovisuelle, mais cette différence n'était pas significative (p=0,13). Au niveau individuel, seul le patient 4 montrait une facilitation significative pour détecter les syllabes en condition audiovisuelle. Aucune des autres différences n'était significative.

10.3.2 Réponses évoquées auditives

Les réponses auditives évoquées par les syllabes se manifestaient comme une succession d'ondes transitoires enregistrées principalement dans le gyrus temporal supérieur et d'amplitude beaucoup plus importante que celles enregistrées dans les mêmes régions pour les activités visuelles. Ces activités n'étant pas l'objet principal de cette étude, on se contentera de les décrire de façon globale en négligeant les aspects propres à chaque patient.

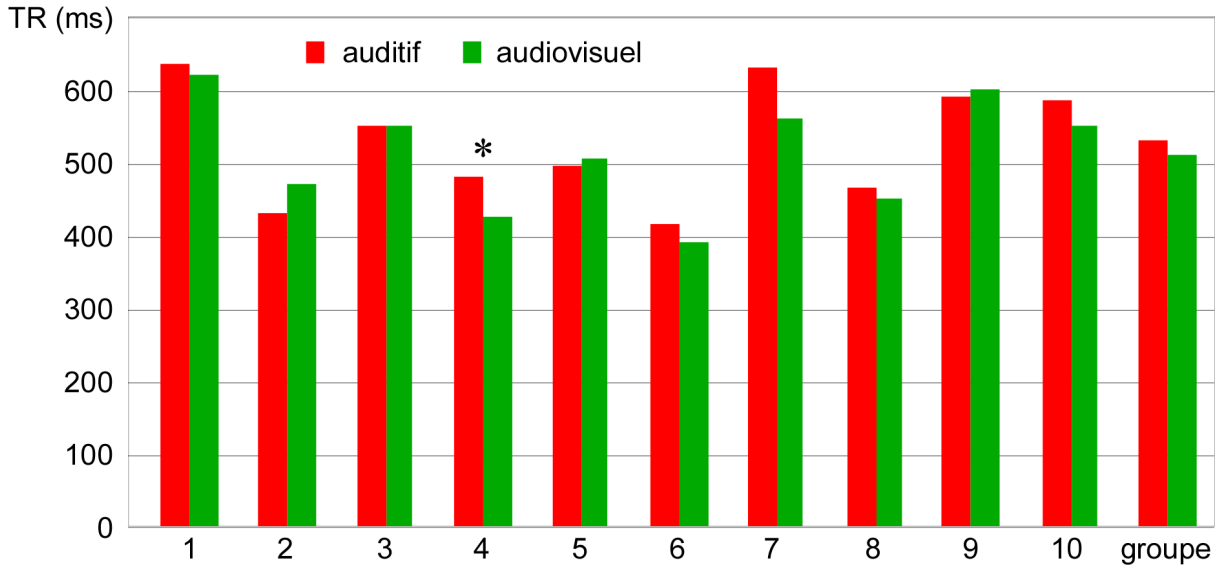

FIG. 10.3 – TR moyens auditifs et audiovisuels par patient et pour le groupe de patients. L'étoile indique une différences significative au seuil p<0,05.

Soulignons simplement que la variabilité des réponses peut être attribuée tout autant à des différences d'implantation, qu'à une variabilité anatomique et fonctionnelle. Malgré cette variabilité, on peut aisément distinguer plusieurs composantes communes à la plupart des patients (figure 10.4 page suivante).

Les premières réponses étaient enregistrées dans la partie médiane du gyrus transverse antérieur (ou gyrus de Heschl) à partir de 15 ms. Le détail des réponses enregistrées dans les 30 premières millisecondes est donné dans la table 10.3 page 139.

Les réponses s'étendaient ensuite dans les parties plus latérales du gyrus transverse ainsi que vers l'arrière sur le planum temporale, à partir de 40 ms post-stimulus. Toutes ces réponses étaient de polarité aussi bien positive que négative (en montage monopolaire). À partir de 70 ms commençait une réponse enregistrée majoritairement comme positive et dont l'amplitude culminait vers 100-130 ms. Cette réponse était enregistrée au niveau du gyrus transverse, du planum temporale, ainsi que sur la partie latérale du gyrus temporal supérieur (GTS) jusqu'à des zones assez postérieures jouxtant le gyrus supramarginal (et correspondant à l'aire Wernicke). Cette composante était suivie par une autre composante d'origine similaire, de polarité majoritairement négative dont le maximum d'amplitude avait lieu autour de 200 ms. Des exemples de ces différentes réponses sont visible chez le patient 6 sur les contacts H3-5 (figure A.3 page 232) ou chez le patient 8, électrode T9 (figure A.5 page 235). Des réponses d'amplitude beaucoup plus faible étaient également visibles dans plusieurs autres régions corticales à partir de 70 ms.

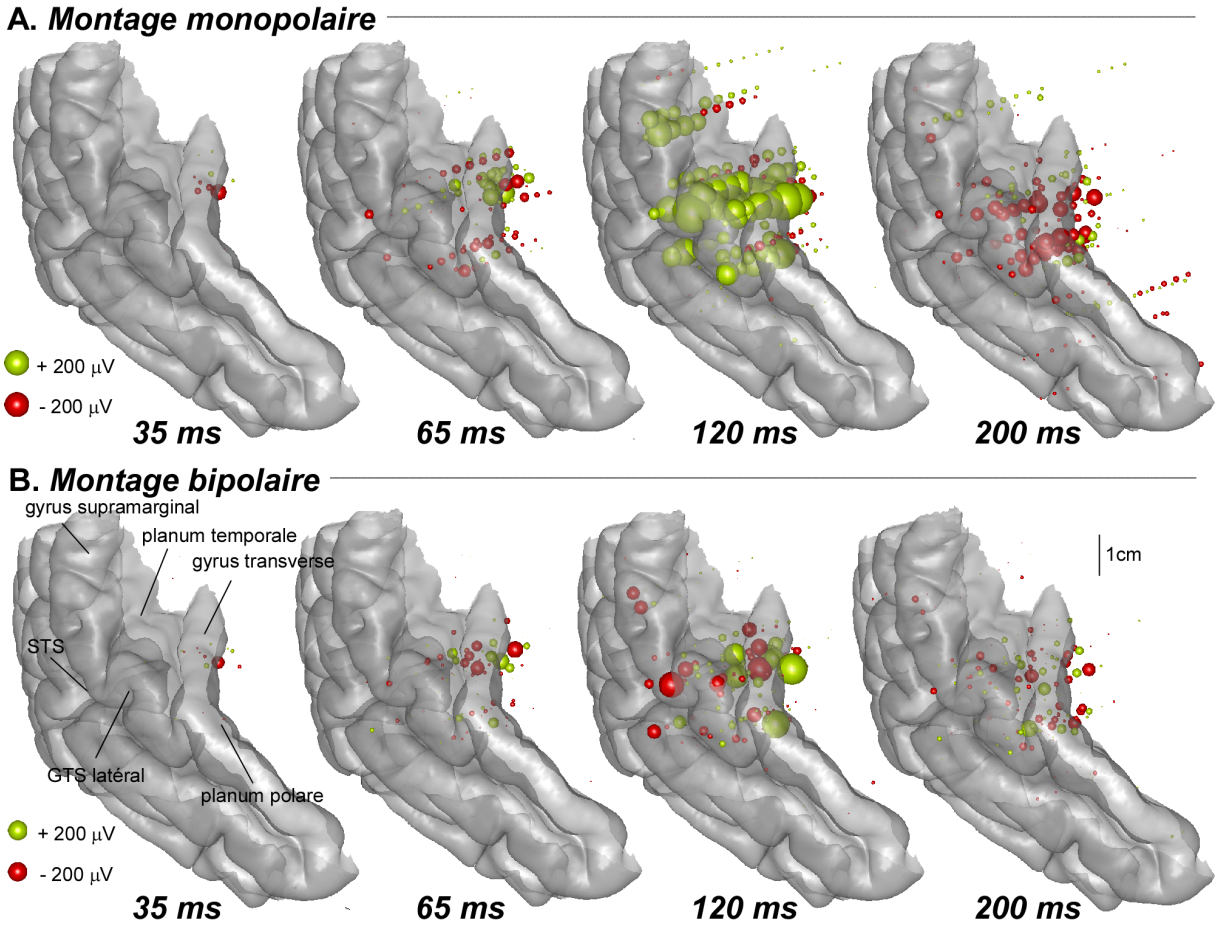

FIG. 10.4 – Réponses auditives de l'ensemble des patients enregistrées dans le cortex temporal, présentés sur une représentation 3D du lobe temporal droit du cerveau du MNI. Les activités enregistrées dans l'hémisphère gauche et droit ont été reportées sur un même hémiphère. Chaque sphère représente la différence de potentiel enregistrée à un contact en montage monopolaire (A) ou bipolaire (B). Le diamètre de la sphère est proportionnel à l'amplitude du potentiel évoqué et la couleur code la polarité. Les coordonnées des contacts des différents patients ont été normalisées et converties dans le repère du cerveau du MNI.

10.3.3 Réponses évoquées visuelles

Les réponses visuelles au mouvement des articulateurs étaient d'amplitude plus faible et avait un caractère moins transitoire que la réponse auditive aux syllabes. Cette différence peut s'expliquer par plusieurs facteurs : d'une part aucune électrode n'explorait les zones visuelles sensorielles (en tous cas pas primaires), d'autre part les stimuli employés dans la modalité visuelle présentaient un départ beaucoup moins abrupt que ceux utilisées dans la modalité auditive, ce qui ne facilite pas l'obtention de réponses élémentaires synchronisées permettant l'observation d'un potentiel évoqué net. Cette différence rappelle celle obtenue en PE de scalp dans l'expérience précédente.

La table A.1 (pages 224–226) rapporte l'ensemble des activités significatives enregistrées en réponse aux mouvements labiaux présentés seuls, que nous avons regroupé par type d'activité présentant des caractéristiques temporelles, spatiales et fonctionnelles communes

Patient	Région explorée	Côté	Latence de début (ms)	Nom des contacts	Coordonnées de Talairach X	Y	Z
5	Gyrus transverse antérieur médial	G	14	H'2	-35	-24	7
10	Gyrus transverse antérieur médial	D	17	H10	43	-19	7
4	Gyrus transverse antérieur médial	G	19	H'6-7	-33	-28	10
7	Gyrus transverse antérieur/planum temporale	G	23	H'6	-48	-22	9
9	Gyrus transverse antérieur médial	G	23	H'6-8	-35	-22	6
6	Gyrus transverse antérieur médial	D	23	H4	39	-20	5
10	Gyrus transverse antérieur médial	D	23	H6	30	-19	7
10	Gyrus transverse antérieur médial	D	23	H8	36	-19	7
4	Gyrus transverse antérieur/planum temporale	G	25	H'8-9	-41	-28	10
8	Gyrus transverse antérieur latéral	D	25	T3-5	42	-11	7
6	Gyrus transverse antérieur médial	D	27	H3	36	-20	5
8	Gyrus transverse antérieur médial	G	27	H'8	-39	-23	7
7	Gyrus transverse antérieur	G	29	H'2-3	-36	-22	9

TAB. 10.3 – Coordonnées, localisation et latence des réponses auditives commençant avant 30 ms chez les différents patients. Les réponses sont classées par latence. Les structures traversées ont été déterminées visuellement sur l'IRM anatomique de chaque patient. Le nom des contacts est constitué de la lettre désignant l'électrode (localisation sur les figures 10.1 page 132 et 10.2 page 133) et du numéro du contact, les nombres les plus petits indiquant les contacts les plus profonds.

et ayant été trouvé chez au moins 3 patients.

On se contentera ici de décrire les 3 premiers types de réponses, les plus précoces, enregistrées dans le MTG postérieur et le STG. D'autres réponses visuelles, généralement plus tardives (à partir de 80 ms après le début du son) ont été enregistrées dans de nombreuses régions. Les régions trouvées chez au moins trois patients étaient : le gyrus supramarginal, le STS antérieur et postérieur, l'opercule post-central, l'insula, le gyrus cingulaire postérieur, l'opercule pré-central/gyrus frontal inférieur (pouvant correspondre à l'aire de Broca), l'hippocampe/ gyrus parahippocampique.

Rappelons que le mouvement des lèvres commençait à partir de 240 ms préstimulus (le temps zéro correspondant au début de la syllabe auditive). Il ne faut donc pas s'étonner que les réponses les plus précoces apparaissaient dès 120 ms pré-stimulus. Ces réponses ont été enregistrées d'une part au niveau de la jonction occipito-temporale et du GTM postérieur et d'autre part au niveau du GTS sur des électrodes explorant aussi bien le gyrus transverse, le planum temporale, le planum polare, le GTS latéral et le bord supérieur du STS.

Concernant la zone occipito-temporale, une réponse y a été enregistrée chez tous les patients dont l'implantation était aussi postérieure. Cette réponse était spécifique à la condition visuelle, et lorsqu'une réponse auditive était enregistrée plus tard dans la même zone, son profil spatio-temporel était clairement différent.

Concernant la partie supérieure du lobe temporal, des réponses visuelles ont été enregistrées sur les mêmes contacts que ceux sur lesquels ont été observés les potentiels évoqués auditifs sensoriels entre 50 et 200 ms. L'un des buts de cette étude étant de vérifier si on peut enregistrer une réponse aux mouvements articulatoires dans le cortex auditif, il nous faut nous assurer que ces réponses proviennent bien du plan supérieur du GTS et non

du STS. En effet, la localisation des contacts ne suffit pas puisque l'activité enregistrée, même en montage bipolaire peut correspondre à la diffusion des potentiels dans le milieu extracellulaire. Cette ambigüité est clairement illustrée dans l'implantation du patient 3 (figure A.2 page 230) : l'électrode H' passe entre le bord supérieur du STS et le planum temporale : il est impossible de dire si une activité enregistrée sur un des contacts de l'électrode H' provient du cortex situé en-dessous ou au-dessus des contacts. Pour répondre à cette question, nous avons comparé le profil spatiale des réponses visuelles à celui des premières réponses auditives transitoires. Il est en effet bien établi que ces réponses auditives précoces sont générées dans le cortex auditif (Liégeois-Chauvel, Musolino, Badier, Marquis & Chauvel, 1994 ; Yvert, Fischer, Bertrand & Pernier, 2005), comme le montre également la figure 10.4 page 138. Si nous pouvons montrer que les réponses visuelles enregistrées dans le lobe supérieur temporal possèdent le même gradient spatial que cette réponse auditive, on pourra en conclure qu'elle est bien générée dans le cortex auditif.

Nous avons classé les différents types de réponse visuelle enregistrées dans le cortex temporal supérieur en fonction de leur ressemblance spatiale avec la réponse auditive transitoire. Sur 12 sites répartis parmi 5 patients, le gradient spatial de la réponse visuelle ressemblait à celui d'une réponse auditive générée à partir de 50 ms, donc à une réponse dont l'origine dans le cortex auditif ne fait guère de doute (type 2 dans la table A.1 page 226 (pages 224–226). On peut voir des exemples d'une telle réponse chez le patient 3 (figure A.2 page 230) au niveau des contacts T4-5 et T7-9 (correspondant respectivement au bord supérieur du STS et au gyrus transverse latéral), chez le patient 8 (figure A.5 page 235), au niveau des contacts H'11-15 et T'8-9 (Planum temporale et STS/GTS latéral). Dans d'autres cas, la ressemblance est plus vague (patient 1, contacts T7-9, gyrus transverse latéral, figure A.1 page 229). Notons que dans le cas du patient 3, la réponse auditive entre 50 et 100 ms était enregistrée avec un gradient plus fort sur le bord supérieur du STS que sur le gyrus transverse, ce qui suggère que le cortex auditif s'étend dans les aires corticales bordant le STS dans le cas de ce patient.

Sur 6 sites répartis sur 4 patients, la réponse visuelle montrait une ressemblance frappante avec une réponse auditive transitoire aux syllabes commençant après 100 ms (type 3 dans la table A.1 pages 224–226). On peut en voir des exemples chez le patient 1 (figure A.1 page 229) sur les contacts H8-10 (gyrus transverse médial/planum temporale) et chez le patient 7 (figure A.4 page 233) au niveau des contacts T'5-7 (planum polare). D'autres sites ne montrent pas le même profil spatial dans les deux modalités, mais l'on observe de forts gradients spatiaux au niveau des mêmes électrodes dans les deux conditions : c'est le cas pour le patient 10 (figure A.6 page 237) au niveau des contacts H7-15 (gyrus transverse médial et planum temporale) et pour le patient 6 (figure A.1 page 229) au niveau des contacts H3-9 (gyrus transverse antérieur médial et postérieur latéral, mais dans ce dernier cas, la réponse visuelle n'était pas significative avec le critère requis).

Au total une telle activation visuelle du cortex auditif a été trouvée chez 7 patients, sur 18 sites. Une telle affirmation n'est pas basée sur une délimitation anatomique du cortex auditif, mais plutôt sur une définition fonctionnelle assez large : le cortex auditif est défini comme la zone du cortex temporal dans laquelle on enregistre une réponse évoquée transitoire à un son ; en effet ces 18 sites comprennent aussi bien le planum polare, le gyrus

transverse, le STG latéral, le bord supérieur du STS que le planum polare jusqu'au gyrus supramarginal.

Un autre argument permettant d'affirmer que cette réponse visuelle venait de la partie supérieure du GTS et non du STS est qu'elle n'était pas enregistrée dans le GTM, ou avec une amplitude beaucoup plus faible, alors que les implantations dans cette région étaient assez nombreuses, comme on peut le voir sur les figures 10.1 à 10.2 pages 132–133 (données non illustrées).

Un autre but de cette étude était de savoir si la réponse visuelle dans le cortex auditif pouvait être générée dans le cortex auditif primaire. Une façon de répondre à cette question est de comparer l'emplacement des sites d'enregistrement de cette réponse visuelle avec la position des sites d'enregistrement des réponses auditives transitoires générées avant 30 ms, probablement dans le cortex auditif primaire. Une telle réponse auditive a été enregistrée sur 13 sites chez 7 patients, exclusivement dans la partie médiale du gyrus transverse, comme c'est illustré dans la figure 10.5 page suivante (aux erreurs de normalisation près). Considérées au niveau individuel, toutes ces réponses étaient enregistrées dans le gyrus transverse médial (voir la table 10.3 page 139). Une comparaison de ces activations auditives primaires avec les réponses visuelles enregistrées dans le cortex auditif (figure 10.5) suggère que les réponses visuelles étaient toujours enregistrées en dehors de la zone définie par les réponses auditives précoces. Cependant, les erreurs de localisation dues à la normalisation des coordonnées et à l'utilisation d'un cerveau standard ne permettent pas d'être catégorique sur ce point.

Si l'on regarde individuellement chaque patient, seuls deux d'entre eux montraient les deux types de réponse sur des contacts voisins : pour le patient 8 (figure A.5 page 235), les foyers étaient clairement différents puisque la réponse visuelle était enregistrée uniquement sur les contacts H'11-15 (planum temporale) alors que la réponse auditive précoce était enregistrée sur H'8-10 (gyrus transverse médial). Quant au patient 10, si le profil spatial de la réponse auditive précoce sur H'7-9 était bel et bien différent de celui de la réponse visuelle, ces deux réponses étaient enregistrées sur les même contacts (c'est également vrai pour le patient 6, contacts H3-4, mais la réponse visuelle émergeait à peine du bruit dans ce cas et n'était pas significative avec le critère requis). L'analyse qualitative de groupe suggère donc que les réponses visuelles dans le cortex auditif sont en général générées hors du cortex auditif primaire. Toutefois, les données d'un (ou peut-être deux) patients suggèrent une activation visuelle du cortex auditif primaire.

10.3.4 Violations du modèle additif

La table A.2 page 227 rapporte les violations significatives de l'additivité des réponses auditives et visuelles, qui signent l'existence d'interactions. Ces violations peuvent être classées en deux catégories selon leur profil spatio-temporel. La figure 10.6 page 143 montre la localisation de ces deux types de violation, qui avaient toutes lieux dans le cortex temporal supérieur. D'autres violations du modèle additif ont été trouvées dans des régions diverses, en dehors du cortex temporal, sans qu'il soit possible d'en dégager une unité fonctionnelle, temporelle ou anatomique (voir la table A.2 pour des détails).

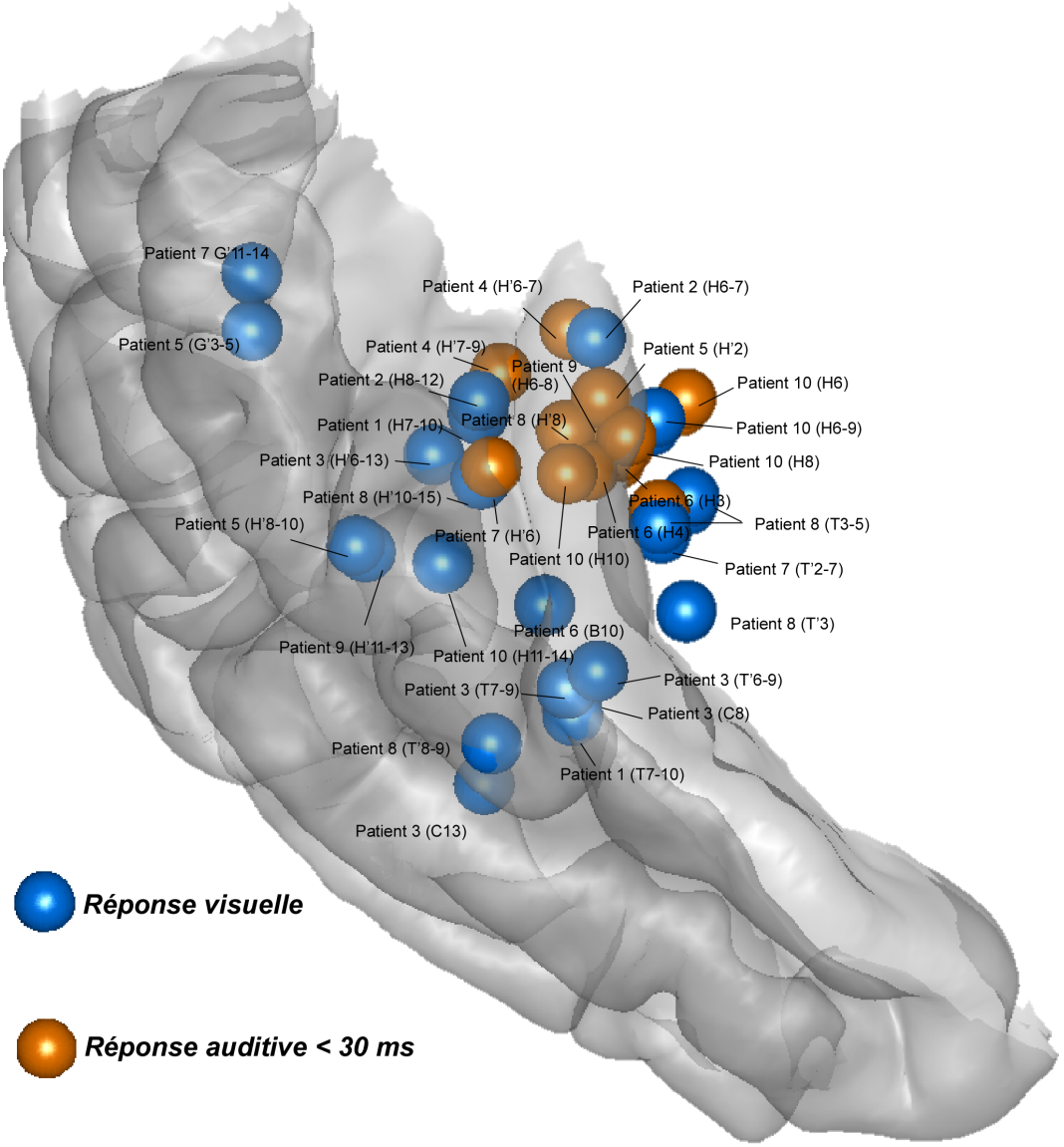

FIG. 10.5 – Sites d'enregistrement des réponses visuelles générées dans le cortex auditif et des réponses auditives précoces générées dans le cortex auditif primaire, présentés sur une représentation 3D du lobe temporal droit du cerveau du MNI. Les activités enregistrées dans l'hémisphère gauche et droit ont été reportées sur un même hémiphère. On considère que la réponse visuelle était générée dans le cortex auditif lorsque le profil spatial de la réponse le long des contacts d'une même électrode était identique à celui d'une réponse auditive transitoire générée entre 50 et 200 ms. On considère qu'une réponse auditive était primaire lorsqu'elle apparaissait avant 30 ms de traitement.

Le premier type de violation du modèle additif a été observé sur 19 sites chez 9 patients. Ces sites étaient tous situés dans la partie supérieure du GTS, dans la région que nous avons définie plus haut comme le cortex auditif au sens large. Ce type de violation de

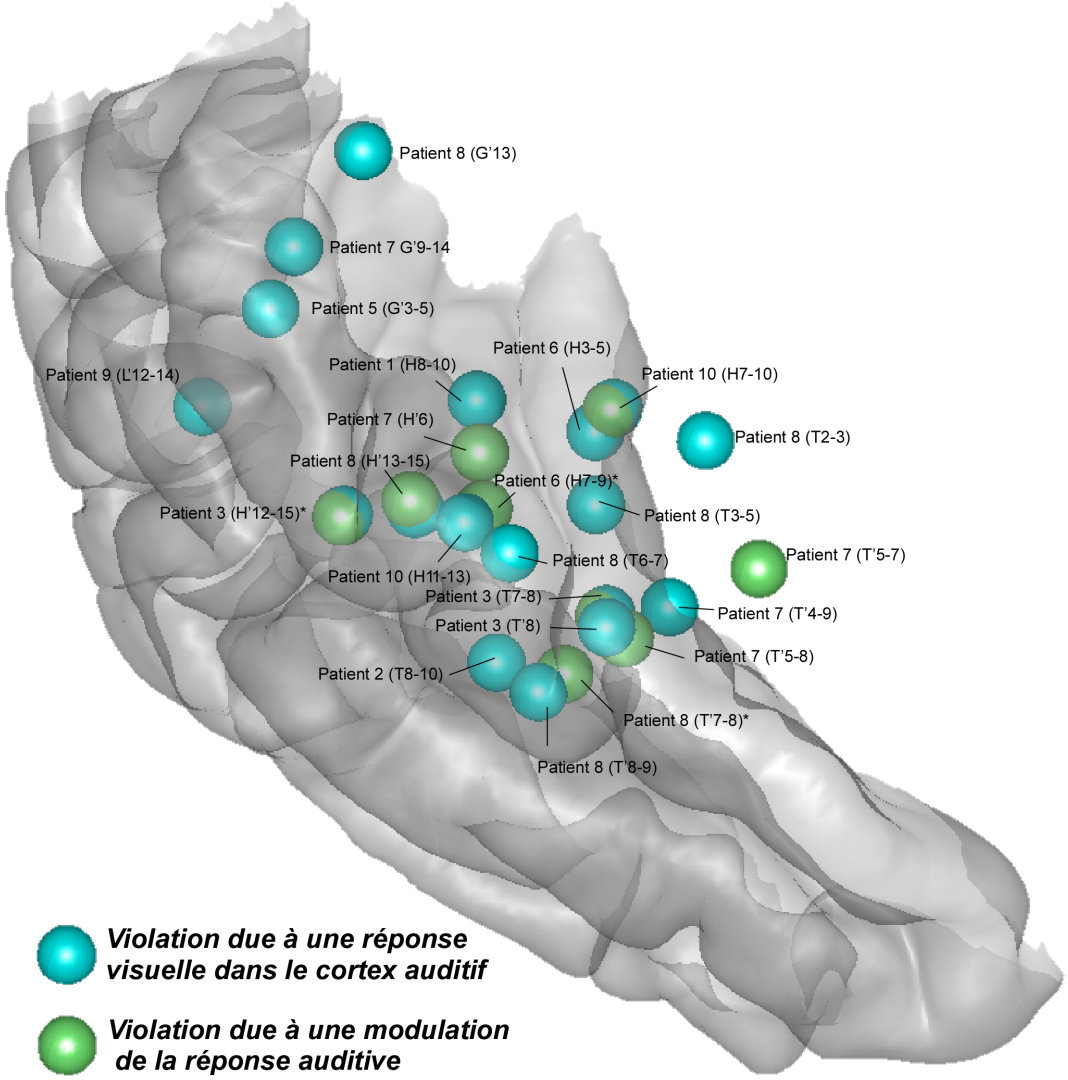

FIG. 10.6 – Deux types principaux de violations du modèle additif commençant avant 200 ms de traitement, présentés sur une représentation 3D du lobe temporal droit du cerveau du MNI. Les activités enregistrées dans l'hémisphère gauche et droit ont été reportées sur un même hémiphère. Les contacts sur lesquels étaient observées ces violations sont indiqués entre parenthèse.

l'additivité est visible chez tous les patients dont les résultats sont illustrés (figures A.1 à A.6 pages 229–237). Ces interactions se présentaient sous la forme suivante : la violation de l'additivité commençait entre 30 et 160 ms après la présentation de la syllabe auditive pour continuer au-delà de la fenêtre d'analyse (200 ms) et souvent au-delà de 600 ms. Le profil spatio-temporel de la violation est exactement celui de la réponse visuelle, mais de polarité opposée. Cela est probablement dû au fait que la réponse en condition audiovisuelle diffère peu de la réponse en condition auditive, autrement dit que la réponse visuelle dans le cortex auditif semble ne pas exister lorsque la stimulation est audiovisuelle, mais seulement lorsque les mouvements articulatoires sont présentés seuls.

Le second type de violation avait lieu entre 40 et 200 ms après la syllabe auditive, au

niveau du gyrus transverse et du planum temporale. Ici, le profil spatio-temporel correspond à celui de la réponse auditive transitoire avec une polarité opposée. Ce type de violation correspond apparemment à une diminution de la réponse auditive transitoire en condition audiovisuelle. On voit clairement cette modulation chez 2 patients.

Chez le patient 8 (figure A.5 page 235), sur le contact H' 11 (planum temporale), en montage bipolaire, on voit clairement un foyer identique à l'activité auditive et audiovisuelle entre 60 et 120 ms, qui n'est pas présent en visuel. La diminution est visible sur les courbes et la violation du modèle additif montre un rebond qui est absent de la réponse visuelle. Chez le patient 10 (figure A.6 page 237), l'activité bipolaire montre une triple inversion de polarité entre 80 et 160 ms aux contacts H6, 7 et 9 (gyrus transverse médial), identique aux inversions observées en conditions auditives et audiovisuelles. À cette latence, on n'observe pas de réponse visuelle dans cette zone. Chez d'autres patients, l'interprétation est plus ambigüe puisque cette forme de violation se superpose au premier type : la violation semble être due à la fois à l'absence de réponse visuelle en condition audiovisuelle et à une diminution de la réponse auditive, à la même latence (patient 7, contacts T'5-7 entre 120 et 200, patient 8, contact H'13 entre 80 et 160 ms).

Enfin, chez certains patients il a fallu augmenter le seuil pour observer cette diminution, tout en conservant l'exigence d'un nombre minimal d'échantillons consécutifs significatifs (patient 7, contacts T'7-8, bord supérieur du STS, entre 60 et 100ms ; patient 8 contacts T'7-8 bord supérieur du STS ; patient 3, contacts H'12-15, GTS latéral entre 50 et 100 ms ; patient 3 contacts T7-8 bord supérieur du STS entre 60 et 120 ms). Notons que pour ces 3 dernières violations, la diminution n'était observée que sur les données monopolaires. L'augmentation du seuil statistique reste raisonnable si l'on considère que ces effets ne pouvaient se produire que sur les contacts sur lesquels étaient enregistrées des réponses transitoires, ce qui réduit en principe le nombre de tests à effectuer (nous reconnaissons le caractère a posteriori de cette affirmation).

La localisation de ce deuxième type d'effet ne diffère guère de celle du premier type, comme on peut le voir sur la figure 10.6 page précédente. Les modulations étaient en fait souvent superposées aux violations dues à l'activation visuelle sur les même contacts décrites plus haut, ce qui rend difficile leur description. Pour la plupart des patients (patients 3, 6, 7, 8 et 10), lorsque l'on compare les courbes de la violation aux courbes de l'activité visuelle, on constate que l'amplitude de la violation est supérieure celle de l'activité visuelle, ce qui suggère que les deux types d'interaction co-existent.

10.3.5 Relations entre réponses auditives, visuelles et interactions audiovisuelles

On peut tenter de décrire les relations existant entre l'activation auditive et visuelle du lobe temporal (supérieur) et les interactions audiovisuelles mises en évidence par l'application du modèle additif, au moins pour les activités communes à plusieurs patients. La table 10.4 page ci-contre donne, pour chaque patient, les latences de début et de fin des 4 principaux effets mis en évidence : l'activation visuelle de la jonction occipito-temporale, l'activation visuelle du cortex auditif, la modulation des ondes audiovisuelles transitoires en condition audiovisuelle et la violation du modèle additif due à la disparition de la réponse

Patient	Réponse V GTM post. JOT		Réponse V Cortex Auditif		Modulation réponse auditive		Disparition réponse V cortex auditif	
	début	fin	**début**	fin	**début**	fin	**début**	fin
1	**-**	-	**-20**	600+	**-**	-	**110**	250
2	**-**	-	**-120**	450	**-**	-	**40**	110
3	**-**	-	**-120**	600+	**50**	120	**80**	600+
4	**-**	-	**-**	-	**-**	-	**-**	-
5	**-**	-	**0**	600+	**-**	-	**130**	250
6	**-80**	350	**-**	-	**40**	120	**30**	600+
7	**-**	-	**-20**	450	**60**	200	**70**	600+
8	**-80**	400	**-70**	600+	**50**	120	**70**	500
9	**-100**	160	**-**	-	**-**	-	**120**	250
10	**-40**	600+	**-30**	600+	**80**	160	**80**	600+

TAB. 10.4 – Latence de début (en gras) et de fin des 4 types d'effets mis en évidence, chez chaque sujet. Réponse V : réponse visuelle significativement différente de la ligne de base. GTM post. : gyrus temporal moyen postérieur. JOT : jonction occipito-temporale. Modulation réponse auditive : violation significative du modèle additif due à une diminution d'une onde auditive transitoire en condition auditive. Disparition réponse V cortex auditif : violation significative du modèle additif due à la disparition de la réponse visuelle du cortex auditif en condition audiovisuel. 600+ : l'effet se prolonge au-delà de 600 ms.

visuelle du cortex auditif en condition audiovisuelle.

Malgré la variabilité des latences, l'enchainement des différentes activations se vérifient chez chacun des patients : lors d'une stimulation audiovisuelle, les indices visuels, qui sont disponibles plus tôt, activent tout d'abord les régions autour de la jonction occipito-temporale (patients 6, 8, 9 et 10), puis immédiatement après le cortex auditif (patients 8 et 10). Cette activation du cortex auditif peut commencer jusqu'à 100 ms avant la présentation de la syllabe auditive (patients 2 et 3). Lorsque les indices auditifs sont présentés, ils activent tout d'abord le cortex auditif primaire puis à partir de 50 ms post-stimulus des zones du cortex auditif qui ont déjà été activées par les indices visuels (voir la partie 10.3.2 page 136). C'est à ce moment que prennent place les deux types d'interaction audiovisuelle : l'amplitude de la réponse auditive est diminuée par rapport à la condition auditive seule alors que le cortex a déjà été activé par les indices visuels (patients 3, 7, 8 et 10). Immédiatement après, ou à la même latence, l'activation soutenue et faible du cortex auditif observée en modalité visuelle seule prend fin pour être dominée par le traitement des indices auditifs (patients 3, 6, 7, 8 et 10). Cette chronologie relative se vérifie en particulier chez les 2 patients chez lesquels nous avons observé les 4 effets (patients 8 et 10).

10.4 Discussion

Les données intracérébrales chez les patients épileptiques donnent des informations précieuses sur le fonctionnement du cerveau, mais proviennent de sujets dont on ne sait pas s'ils représentent un bon modèle du fonctionnement cognitif normal étant donné leur

pathologie. Nous avons donc privilégié, dans notre description des résultats, ceux qui pouvaient être caractérisés de manière fonctionnelle, anatomique et/ou temporelle de la même manière chez plusieurs patients.

10.4.1 Activité du cortex auditif en réponse aux indices visuels de parole

La vision des mouvements articulatoires active de nombreuses aires cérébrales dont la jonction occipito-temporale, le GTS (gyrus transverse, planum temporale, planum polare, GTS latéral), le STS antérieur et postérieur, le gyrus supra-marginal, le STS postérieur, l'opercule post-central, l'opercule pré-central, le gyrus frontal inférieur postérieur, l'insula, l'hippocampe ou le gyrus para-hippocampique. La liste n'est bien évidemment pas exhaustive, d'autant plus que nombre d'aires cérébrales n'étaient pas explorées. Parmi ces aires, on peut en particulier distinguer la jonction occipito-temporale et le GTS dont l'activation, bien que la plupart du temps assez soutenue, commençait avant celle des autres aires cérébrales mentionnées (à partir de 100 ms avant le stimulus auditif, c'est-à-dire 140 ms post-stimulus visuel).

La jonction occipito-temporale faisant partie du cortex visuel, il n'est pas étonnant qu'elle soit la première aire que nous voyions activée par un stimulus visuel. En revanche, il est frappant de voir que le GTS est activé presque à la même latence. La comparaison des profils spatiaux de cette activation avec les réponses auditives transitoires montre qu'il s'agit d'une activation visuelle du cortex auditif. Cette activation avait déjà été rapportée par la plupart des études IRMf sur la lecture labiale, mais c'est la première fois à ma connaissance que l'on a accès à sa dimension temporelle. Il semble qu'elle soit donc relativement précoce puisqu'elle suit de très peu les traitements dans le cortex visuel (ce qu'on en voit en tous cas) et il est donc peu probable qu'elle représente un phénomène d'imagerie auditive. L'analyse de groupe suggère cependant que cette activation a en général lieu hors du cortex auditif primaire, contrairement à ce qui a été montré en IRMf par un certain nombre d'auteurs (Calvert et coll., 1997 ; Ludman et coll., 2000 ; MacSweeney et coll., 2001 ; Pekkola et coll., 2005). Une telle activation est cependant observée chez au moins un patient (le patient 10). Ce résultat peut être attribué soit à un défaut de couverture spatiale chez les autres patients, soit à une réponse atypique chez ce patient.

Les autres aires étaient activées en condition visuelle plus tardivement (en général après 50 ms post-stimulus auditif — 300 ms post-stimulus visuel — pour le STS antérieur, après 100 ms post-stimulus auditif , pour le STS postérieur et le gyrus supra-marginal et après 200 ms post-stimulus auditif dans les autres structures ; voir la figure 10.7 page ci-contre). Il est cependant hasardeux d'établir une chronologie étant donné la variabilité importante des latences entre les patients, sans doute due à la variabilité des implantations.

Notre protocole ne nous permet pas de distinguer parmi les activations trouvées celles qui sont propres à la perception visuelle de la parole et celles qui pourraient être évoquée par tout type de mouvements labiaux, contrairement aux expériences en IRMf ayant utilisé comme contrôle des mouvements labiaux non langagiers (Calvert et coll., 1997 ; Campbell et coll., 2001 ; Paulesu et coll., 2003). Les figures 10.7 page suivante et 10.8 page 148 comparent les activations visuelles trouvées dans notre étude aux résultats des études

IRMf sur la lecture labiale.

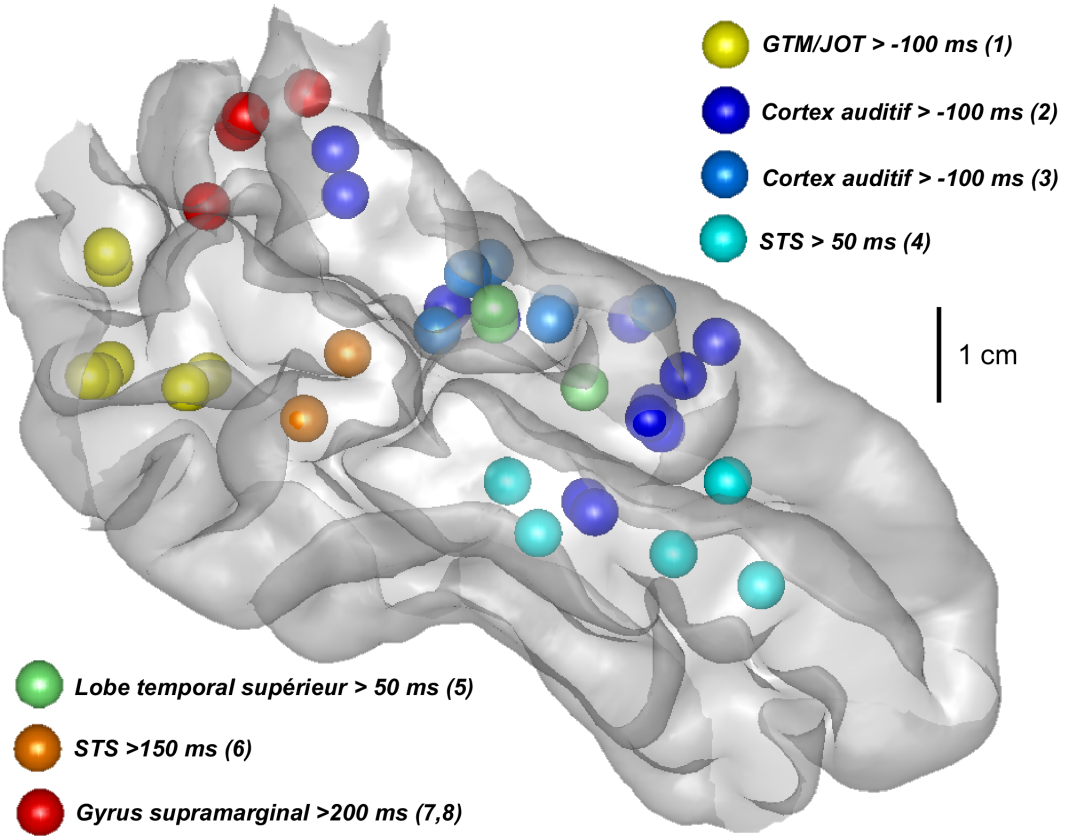

FIG. 10.7 – Activités enregistrées dans le lobe temporal en réponse aux mouvements articulatoires dans la présente étude. Les catégories de réponse correspondent à celles données dans la table A.1 page 226. Les latences sont données par rapport au début de la syllabe auditive. Il existe une discordance entre la localisation indiquée dans la légende et la situation effective sur le cerveau du MNI, due aux erreurs de normalisation. Les activations dans l'hémisphère gauche et droit ont été reportées sur un même hémiphère. JOT : jonction occipito-temporale. STS : sillon temporal supérieur. GTM : gyrus temporal moyen.

Le contraste le plus couramment utilisé dans ces études a pour but d'identifier les zones du cerveau présentant une réponse hémodynamique plus grande pour des mouvements articulatoires langagiers que pour la vision d'une bouche au repos. Il est analogue à la comparaison que nous avons effectuée entre la ligne de base et la réponse au mouvement. La localisation de ces activations en IRMf (sphères de couleur bleu foncé dans la figure 10.8) correspondent grossièrement à celles des activations que nous avons rapporté (figure 10.7), si l'on prend en compte la diffusion de potentiels en sEEG.

Certaines études IRMf ont comparé les activations induites par des mouvements labiaux non langagiers et une bouche au repos. Ces activations sont toutes regroupées au niveau de la jonction occipito-temporale et du GTM postérieur (sphères de couleur turquoise dans la figure 10.8). Il est donc vraisemblable que les premières activations que nous observons au niveau occipito-temporal ne sont pas spécifique de la parole.

Logiquement, les études IRMf qui ont testé un contraste entre mouvements langagiers et non langagiers (sphères de couleur jaune dans la figure 10.8) ont trouvé des activa-

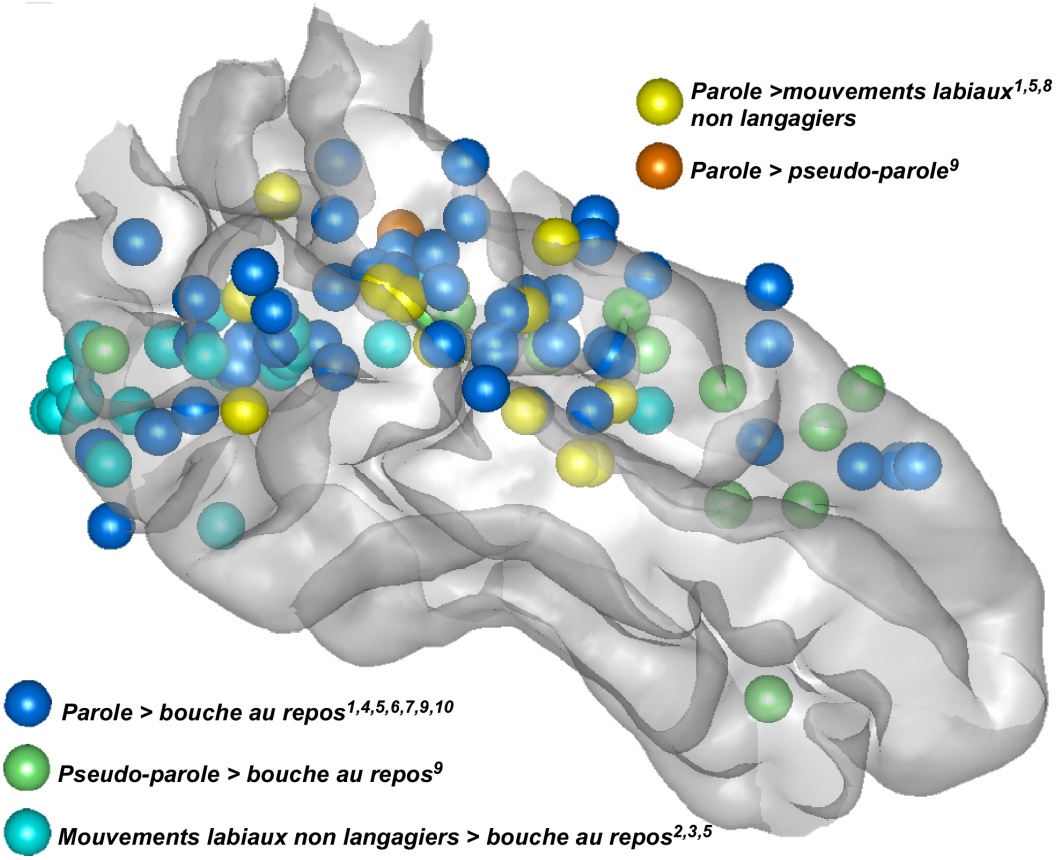

FIG. 10.8 – Activation du lobe temporal en lecture labiale. Les activités reportées proviennent de différentes études en IRMf, dont les résultats étaient reportés en coordonnées de Talairach ou directement en coordonnées du MNI. Les coordonnées de Talairach ont été converties dans le repère du cerveau du MNI. Les chiffres en exposant à côté de chaque contraste indiquent de quelle(s) étude(s) proviennent les activations : 1. Calvert et coll. (1997) 2. Puce et coll. (1998) 3. Puce et Allison (1999) 4. MacSweeney et coll. (2000) 5. Campbell et coll. (2001) 6. MacSweeney et coll. (2001) 7. Olson et coll. (2002) 8. MacSweeney et coll. (2002) 9. Paulesu et coll. (2003) 10. Calvert et Campbell (2003). Les activations dans l'hémisphère gauche et droit ont été reportées sur un même hémiphère.

tions autour du STS, du STG latéral et du planum temporale. Nous pensons donc que nos activations du cortex auditif sont spécifiques au traitement langagier des mouvements articulatoires de la bouche.

En revanche, excepté une activation dans le planum temporale rapporté par Paulesu et coll. (2003, sphère orange dans la figure 10.8), la comparaison entre des mouvements de parole ayant un sens pour le locuteur et des mouvements présentés à rebours (et n'étant donc pas interprétables phonétiquement par le sujet, pseudo-parole) n'active que des zones en dehors du lobe temporal (Calvert et coll., 1997 ; Paulesu et coll., 2003). La pseudo-parole active d'ailleurs largement le cortex auditif (sphères vertes dans la figure 10.8).

N'oublions toutefois pas que les études IRMf n'ont pas accès à la dimension temporelle et que les activations reportées dans la figure 10.8 sont susceptibles de correspondre à des activations plus tardives que celles décrites dans notre étude.

Soulignons enfin une différence fondamentale entre notre expérience et les études IRMf :

dans notre étude, les patients n'avaient pas explicitement à lire sur les lèvres. Cela n'empêche pas que les indices visuels aient eu une certaine pertinence dans la mesure où ils étaient susceptibles d'aider à réaliser la tâche demandée.

Les activations ayant lieu hors des lobes occipitaux et temporaux (opercules pré-central, post-central, le gyrus frontal inférieur, l'insula, l'hippocampe) à des latences plus tardives ont été rapportées de façon récurrente dans les études en IRMf. L'activation de l'opercule pré-central et gyrus frontal inférieur en particulier est intéressante puisqu'elle pourrait correspondre à l'aire de Broca ou à l'aire motrice correspondant aux articulateurs faciaux, dont il a été proposé qu'elle participe au décodage phonologique des sons de parole (Ojanen et coll., 2005 ; K. E. Watkins, Strafella & Paus, 2003 ; Wilson, Saygin, Sereno & Iacoboni, 2004) et/ou lors de la lecture labiale (Blasi et coll., 1999 ; MacSweeney et coll., 2001 ; Paulesu et coll., 2003 ; Sundara, Namasivayam & Chen, 2001). Nous avons cependant peu d'éléments permettant de dire que cette activation était suffisamment précoce pour remplir cette fonction. Certes, chez un patient (patient 6), l'activation de cette région commençait dès 60 ms avant l'arrivée du son. Mais, d'une part, il existe une certaine ambigüité due au fait que l'électrode sur laquelle a été enregistrée cette activité se trouvait juste au dessus du cortex auditif et, d'autre part, chez les autres patients, elle n'avait lieu qu'à partir de 200 ms après le début du son. De plus elle avait lieu dans l'hémisphère droit, alors que l'aire de Broca est censée être fortement latéralisée à gauche.

10.4.2 Interactions audiovisuelles

L'application du modèle additif a révélé de nombreuses violations du modèle additif avant 200 ms de traitement des syllabes, et ce en dépit du fait que les patient n'ont en général pas tiré parti des indices visuels pour améliorer leurs performances. Les violations observées au niveau individuel sans être reproduites chez plusieurs patients ne seront pas discutées plus avant. Ces résultats individuels peuvent être attribués à la fois à la spécificité des implantations des électrodes chez chaque patient et peut-être au caractère idiosyncratique de certaines formes d'interactions audiovisuelles.

Là où l'implantation était la plus fournie, à savoir au niveau du lobe temporal supérieur, nous avons pu mettre en évidence deux formes de violation de l'additivité, qui semblent refléter la non additivité des réponses du cortex auditif aux indices auditifs et visuels de parole. La forme de violation la plus indiscutable semble être due au fait que les indices de chaque modalité active le cortex auditif d'une manière qui lui est propre : logiquement, les activités dues aux indices auditifs sont beaucoup plus nettes, amples et transitoire que celles dues aux indices visuels. Lorsque les indices des deux modalités sont présentés (essais audiovisuels), la réponse visuelle semble complètement s'effacer au profit de la réponse auditive, ce qui résulte en des interactions dont le profil spatio-temporel imite exactement celui de l'activation visuelle avec des polarités opposées. Cette violation de l'additivité représente indubitablement une forme d'intégration des informations auditives et visuelles dans la mesure où, en condition audiovisuelle, l'activation visuelle du cortex auditif semble ne pas continuer dès lors que les même zones sont activées par les indices auditifs. Le traitement visuel semble donc influencé par la présence des indices auditifs dès 30 ms de

traitement auditif. Notons que ce type de violation de l'additivité pourrait être dû à un effet plafond de l'activation du cortex auditif.

Mais ce qui nous intéresse plus encore est de savoir si les indices visuels ont réciproquement une influence sur le traitement des syllabes auditives dans le cortex auditif. Il semble bien que ce soit le cas (bien que l'effet soit moins robuste dans ce cas) : chez 5 patients la violation de l'additivité présente un profil spatio-temporel ressemblant à celui d'une réponse auditive transitoire et ne peut être expliquée par la réponse visuelle sur ces contacts et à cette latence. Chez tous les patients, cette modulation prend place à une latence à laquelle une réponse aux indices visuels a déjà pris place, sur les même contacts. Il parait vraisemblable que la préactivation visuelle est responsable de la diminution de la réponse auditive. On peut imaginer que le traitement des indices auditifs est ici facilité par le traitement déjà réalisé sur les indices visuels. Mais, pas plus qu'en EEG, ces données ne nous permettent de dire si les informations auditives et visuelles intégrées à ce niveau sont de nature phonétique ou non ou si cette facilitation représente un amorçage phonologique ou un effet d'indiçage attentionnel.

De même que les activations visuelles décrites plus haut, ces deux types d'interaction semblent avoir lieu majoritairement dans le cortex auditif secondaire (GTS, Planum temporale, Gyrus transverse latéral, Planum polare). Quant au cortex auditif primaire, on y retrouve logiquement la première forme de violation chez deux patients (6 et 10) qui montraient également une réponse visuelle au niveau du cortex auditif primaire. On observe également une diminution de la réponse auditive transitoire au niveau du cortex auditif primaire chez le patient 10, mais il s'agit d'une réponse transitoire générée entre 80 et 160 ms et non d'une composante auditive précoce. Nous n'avons donc pas d'éléments permettant de dire que le traitement auditif des syllabes peut être modulé par les indices visuels avant 50 ms de traitement auditif.

Le cortex auditif primaire a été impliqué dans plusieurs études IRMf de l'intégration des indices auditifs et visuels de parole. Une expérience de L. M. Miller et D'Esposito (2005) a montré par exemple qu'il était plus activé lorsque la syllabe audiovisuelle était perçue comme un évènement audiovisuel unitaire que lorsque les indices auditifs et visuels n'étaient pas subjectivement fusionnés. Son activité serait également liée à l'amélioration de l'intelligibilité de la parole dans le bruit sous l'influence des indices visuels (Callan et coll., 2003). Cependant nos résultats sont contradictoires avec des données IRMf ayant utilisé un critère de super-additivité (voir la partie 4.5 page 72) pour mettre en évidence une implication du cortex auditif primaire(Calvert et coll., 2000) ou du GTS (Wright et coll., 2003). En effet, les effets observés chez nos deux patients suggèrent plutôt un effet de type sous-additif puisque l'activité visuelle semble disparaitre et que l'activité auditive semble diminuer en condition audiovisuelle. Il se peut que l'activité observée dans les études IRMf correspondent à une activité plus tardive du cortex auditif.

10.4.3 Comparaison avec l'expérience EEG de surface

Comparons maintenant les données obtenues dans cette expérience sEEG à celle obtenues en EEG de scalp. Rappelons que les stimuli étaient identiques dans les deux expériences, à ceci près que les syllabes étaient présentées dans un casque aux patient et en champ ouvert aux sujets de l'expérience EEG.

On peut faire deux constats : la réponse générée dans le cortex auditif par les indices visuels de parole n'a pas été observée en scalp, et les latences des violations de l'additivité dans les deux expériences ne correspondent pas. La réponse visuelle, tout comme les violations du modèle additif provenant du cortex auditif (types 1 et 2), devrait en principe apparaitre sur le scalp comme des inversions de polarité entre les mastoïdes et le vertex. Or, on n'observe pas une telle topographie en EEG dans la condition visuelle seule. Par ailleurs, la violation ne prend la forme d'une inversion de polarité qu'à partir de 120 ms en EEG de scalp alors qu'en sEEG le premier type de violation du modèle apparait dès 30 ms et les modulations de l'activité auditive sont visibles principalement sur des composantes générées entre 50 et 120 ms.

On peut avancer plusieurs explications pour cette divergence de résultats : Tout d'abord, il est possible que les patients épileptiques ne constituent pas un bon modèle du fonctionnement cognitif normal. Cette explication parait cependant insuffisante étant donné d'une part la reproductibilité chez plusieurs patients des résultats rapportés et d'autre part le fait qu'aucun d'entre eux ne présentait de difficulté de compréhension ou de production de la parole.

Une possibilité plus convaincante est que l'EEG de scalp n'accède qu'à une partie des composantes générées dans le cortex auditif, notamment du fait que les activités avant 100 ms sont de polarités variées en montage monopolaire. On peut donc s'attendre à ce que la résultante de ces activations, et donc de leurs modulations par les informations visuelles, aient une amplitude assez faible sur le scalp et n'émergent pas du bruit. De la même façon les réponses visuelles dans le cortex auditif, qui présentaient souvent le même profil spatial que les réponses auditives transitoires avant 100 ms présentaient des polarités variées qui pourraient expliquer qu'elles soient invisibles en EEG de scalp. Étant donné que cette réponse visuelle n'est pas visible en EEG de scalp, cela permet d'exclure que la violation de l'additivité observée dans l'expérience précédente corresponde au premier type de violation observée en sEEG.

En revanche, la composante qui apparait à partir de 70 ms sur une large part du planum temporale et du gyrus transverse médian et qui présente un pic d'activation entre 100 et 130 ms selon les patients pourrait correspondre à l'onde N1, bien que le pic de cette dernière avait lieu vers 135 ms en EEG de scalp. En sEEG, la polarité de cette composante en montage monopolaire, positive sur des contacts situés sous le cortex correspond bien à la polarité de l'onde N1, qui est positive au niveau des mastoïdes en EEG (voir aussi Godey, Schwartz, Graaf, Chauvel & Liégeois-Chauvel, 2001 ; Yvert et coll., 2005). Une modulation de cette composante, visible entre 80 et 200 ms chez trois patients, dont au moins deux sur une composante positive en montage monopolaire, pourrait donc fort bien correspondre à l'effet trouvé en EEG de scalp.

Chapitre 11

Étude comportementale de l'effet d'indiçage temporel des stimuli visuels sur le traitement de la parole

11.1 Introduction

Nous avons montré que voir les mouvements de lèvres accompagnant une syllabe auditive permet de la traiter plus rapidement dans une tâche de discrimination et que cet avantage temporel était associé dans les potentiels évoqués à la diminution de l'onde N1 auditive évoquée par la syllabe plosive. Les données sEEG ont montré, d'une part, que les informations visuelles de parole pouvaient activer le cortex auditif avant la présentation de la syllabe auditive et, d'autre part, que cette activation modifiait l'activation du cortex auditif par la syllabe auditive. Ces résultats ont d'abord été interprétés comme un effet de l'intégration des informations phonétiques visuelles données par la configuration des articulateurs faciaux (ouverture de la bouche notamment) aux informations auditives, permettant de faciliter le traitement phonétique de la syllabe auditive. Il existe cependant d'autres explications plausibles. Elles tiennent principalement au fait que, dans les syllabes plosives utilisées, le mouvement des lèvres précède toujours le son. En effet les lèvres doivent préparer l'explosion du /p/. Bien que ce mouvement soit de faible amplitude par rapport à l'ouverture de la bouche qui accompagne le son et qui donne une véritable information phonétique, il est néanmoins clairement perceptible et commence entre 200 et 100 ms avant l'explosion. Ce mouvement précoce peut donner deux types d'informations :

- il informe le locuteur du moment précis auquel se produira le son.
- par le phénomène de co-articulation, il peut informer le locuteur sur la nature phonétique de la voyelle qui suit.

C'est le premier phénomène qui peut mettre en défaut notre interprétation : en effet si le mouvement des lèvres indique au locuteur que la syllabe arrive, il réduit l'incertitude sur le début de ce son et peut permettre de le traiter plus efficacement. L'effet observé au niveau de l'onde N1 auditive pourrait alors refléter cet effet d'indiçage temporel. Ce phénomène pourrait alors être l'équivalent intermodal et temporel de l'indiçage périphérique dans le

domaine spatial.

De nombreuses études ont montré l'existence d'effets attentionnels exogènes intermodaux en dehors du champ de la parole. Ainsi, il a été montré qu'un indice visuel spatial facilite le traitement d'un stimulus auditif présenté subséquemment au même emplacement (Ward, 1994 ; Ward, McDonald & Lin, 2000). L'existence d'un tel effet attentionnel intermodal a cependant longtemps été controversé (Buchtel & Butter, 1988 ; Spence & Driver, 1997) et semble plus difficile à démontrer expérimentalement que celui d'un indice auditif spatial sur le traitement visuel.

Au niveau des potentiels évoqués de scalp, les bénéfices attentionnels d'un indice visuel sur le traitement auditif se manifestent par une négativité accrue (McDonald et coll., 2001), contrairement à ce que nous avions observé dans l'étude EEG. Cependant dans notre cas, il s'agit non pas d'attention spatiale exogène, mais d'un effet d'alerte du stimulus visuel sur le traitement auditif, et les manifestations de ce type d'attention sur les potentiels évoqués pourraient être différents ce ceux de l'attention spatiale exogène intermodale. Contrairement à l'effet d'alerte d'un stimulus auditif accessoire sur la vitesse de traitement visuel, ce phénomène intersensoriel a été très peu étudié. On dispose de quelques données comportementales sur l'amélioration du seuil de perception auditive (Child & Wendt, 1938 ; Howarth & Treisman, 1958) et sur une diminution de temps de détection de stimuli auditifs par un stimulus accessoire visuel, qui suggèrent qu'un effet d'alerte d'un stimulus auditif sur la vitesse de traitement visuel pourrait exister (L. K. Morrell, 1968a ; Posner et coll., 1976 ; I. H. Bernstein et coll., 1973, expérience 2). Mais il n'existe pas à ma connaissance de données en électrophysiologie.

En ce qui concerne la perception de la parole, il semble bien que l'information temporelle (non phonétique) apportée par la vision des articulateurs puisse être utilisée pour faciliter le traitement de l'information auditive. Ainsi Grant et Seitz (2000) ont montré que les informations visuelles permettaient de diminuer le seuil de perception d'une phrase dans le bruit. Ils avancent que cela est dû à la corrélation temporelle existant entre la variation de la surface d'ouverture de la bouche et l'enveloppe du signal auditif. Cependant il se pourrait que les zones de fortes corrélations correspondent aux zones temporelles où le visage donne le plus d'informations, auquel cas l'effet ne serait pas dû à un effet d'indiçage temporel mais à une intégration des informations phonétiques auditives et visuelles.

Schwartz et coll. (2004) ont tenté d'isoler la contribution des indices visuels temporels d'un possible effet des informations visuelles phonétiques sur l'amélioration de l'intelligibilité de la parole. Dans leur expérience, ils utilisaient 10 syllabes différant soit sur leur lieu d'articulation, soit sur leur mode, soit sur leur voyelle (/gy/, /gu/, /dy/, /du/, /ty/, /tu/, /ky/, /ku/, /y/, /u/). Ces 10 syllabes présentent toutes un mouvement articulatoire identique si bien qu'elles sont impossible à distinguer visuellement. La tâche des sujets consistait, à chaque essai, à identifier la syllabe présentée dans le bruit (un bruit de foule), accompagnée ou non des indices visuels. Les résultats montrent que les indices visuels, bien que non discriminants, améliorent l'intelligibilité du voisement dans le bruit, mais pas des autres traits phonétiques (dans leur expérience 3, la même vidéo était artificiellement montée sur les 10 syllabes auditives pour s'assurer que les indices visuels ne fournissent aucune information phonétique pour la réalisation de la tâche). C'est donc que l'information tem-

porelle portée par le mouvement a facilité la détection du pré-voisement, dont la présence ou l'absence détermine la nature voisée ou non voisée de la syllabe. Il s'agit donc d'un pur effet d'indiçage temporel par le mouvement de lèvres. Cette facilitation semble toutefois être spécifique aux indices visuels de parole, puisque lorsque la bouche est remplacée par un rectangle de surface variant proportionnellement à la surface d'ouverture de la bouche, cet effet disparait.

La question se pose alors de savoir quels sont les corrélats neurophysiologiques de cet effet d'indiçage temporel. Se manifestent-ils de la même manière que les interactions audiovisuelles que nous avons mises en évidence en EEG et en sEEG ? Si tel était le cas, les effets observés dans ces expériences pourraient refléter cet effet d'indiçage intermodal et ne pourraient plus être considérés comme un corrélat de l'intégration audiovisuels d'informations phonétiques auditives et visuelles. Une question intéressante est alors de savoir si cet effet d'indiçage est spécifique à la parole ou peut s'observer avec n'importe quel indice temporel visuel.

Nous avons donc voulu explorer par une méthode électrophysiologique les mécanismes à l'œuvre dans cet effet d'indiçage temporel. Notre projet était à l'origine de réaliser une expérience en MEG en utilisant les stimuli de Schwartz et coll. (2004), présentés dans les modalités auditive, visuelle et audiovisuelle et d'utiliser le modèle additif pour mettre en évidence d'éventuels effets d'interaction audiovisuelle associés à cet effet d'indiçage. Les expériences comportementales présentées dans cette thèse étaient destinées à voir comment on peut adapter l'expérience de Schwartz et coll. (2004) à une étude MEG, afin de mettre en évidence à la fois l'effet comportemental de facilitation et des interactions audiovisuelles. L'expérience en MEG n'a pu être réalisée, faute de temps.

11.2 Expérience comportementale 1

L'application du modèle additif en électrophysiologie nécessite un nombre d'essais important avec des stimuli identiques présentés dans trois conditions (auditive, visuelle et audiovisuelle). Or, dans le protocole de Schwartz et coll. (2004), les sujets devaient identifier 12 syllabes assez différentes d'un point de vue acoustique. Il fallait donc limiter le nombre de syllabes différentes présentées aux sujets. Le résultat principal de leur étude étant que l'indiçage visuel temporel facilite la discrimination du voisement, nous avons décidé de n'utiliser qu'une paire de syllabes différant sur leur voisement (par exemple /du/-/tu/), la tâche étant de simplement discriminer ces deux syllabes. Ainsi, le processus de discrimination sur lequel influe la modalité visuelle reste présent et devrait engager à peu près les mêmes processus sensoriels dans un protocole plus simple et adapté à la MEG.

Pour optimiser le temps d'expérience et réduire les problèmes liés à la réponse motrice, l'idéal est d'utiliser une des deux syllabes comme stimulus non-cible fréquent et l'autre comme stimulus cible rare. Nous avions donc besoin de savoir si l'influence des informations visuelles sur la discrimination s'exerce sur l'un, l'autre ou les deux types des syllabes afin de choisir quelles seraient la syllabe cible et la syllabe non-cible.

Un autre problème de la MEG/EEG est que le rapport signal/bruit des réponses cérébrales doit être le plus grand possible. On doit donc éviter de présenter les stimuli dans le bruit car celui-ci risque de rajouter un bruit neuronal à l'activité MEG de fond dont on tente de se débarrasser en moyennant les essais individuels. Or, à supposer que l'on observe des résultats analogues à ceux de Schwartz et coll. (2004) au même niveau de bruit, il n'est pas garanti qu'ils seraient toujours observés sans bruit car la performance dans ce cas atteint un plafond, d'autant qu'une tâche de discrimination entre deux syllabes est plus facile que la tâche d'identification parmi 12 syllabes. Nous avons donc testé 3 conditions de bruit (pas de bruit, un niveau de bruit de équivalent à celui utilisé dans le protocole original et un niveau intermédiaire) et nous avons mesuré à la fois les performances dans la tâche de discrimination et les TR de discrimination, car l'effet de facilitation était plus susceptible de s'exprimer sur les TR dans les conditions où le bruit était plus faible.

De plus nous voulions savoir si les effets éventuellement mis en évidence étaient spécifiques aux mouvements des lèvres ou s'ils pouvaient exister si les lèvres étaient remplacées par le mouvement d'un rectangle donnant les même informations temporelles.

On a donc une expérience manipulant 4 facteurs : le voisement, le niveau de bruit, la modalité et la nature de l'information visuelle (lèvres ou rectangle). Nous avons émis l'hypothèse que l'on devrait observer un taux d'erreurs moins important dans la condition audiovisuelle que dans la condition auditive seule, mais seulement lorsque les informations temporelles étaient données par les lèvres, et non par les rectangles. Cette configuration d'effets devrait être observé au moins dans la condition la plus bruité. En ce qui concerne les temps de discrimination, ils devraient être plus courts dans la condition audiovisuelle que dans la condition auditive, et cet effet devrait être plus important pour la bouche que pour le rectangle.

Ces deux effets devraient interagir avec le niveau de bruit puisqu'il est connu que l'influence des informations visuelles est d'autant plus important que le rapport signal sur bruit est faible. On espère cependant qu'ils seront toujours présents dans la modalité sans bruit, contrairement au taux d'erreurs.

11.2.1 Méthodes

Sujets

Onze sujets droitiers (dont 8 de sexe féminin), d'une moyenne d'âge de 27,7 ans (écart-type : 4 ans) ont passé cette expérience. Aucun ne souffrait de troubles auditifs ou visuels.

Stimuli

Les vidéos utilisées dans cette expérience ont été adaptées de celles utilisées par Schwartz et coll. (2004). Les syllabes étaient prononcées par un homme de langue maternelle française aux lèvres peintes en bleu (pour une raison indépendante de notre volonté), dont seule la partie inférieure du visage était visible. La taille de la bouche correspondait à 2,2° d'angle visuel. Une séquence visuelle commençait par une l'image fixe d'une bouche au repos et se terminait par la même image fixe. Les mouvements labiaux présentés étaient

identiques, quelle que soit l'identité de la syllabe auditive et consistaient en une suite de 20 images d'une durée de 33 millisecondes chacune. Dans la condition "rectangle", le visage était remplacé par un rectangle rouge dont la surface variait de façon inversement proportionnelle à l'aire d'ouverture de la bouche. La largeur de ce rectangle était identique à celle de la bouche, sa hauteur minimale était de 0,12° et sa hauteur maximale de 0,52° d'angle visuel.

Les stimuli visuels était présentés dans les mêmes conditions que notre première étude en EEG. En prévision de l'étude MEG, dans laquelle on utilise un vidéo projecteur ayant une fréquence de rafraichissement, non modifiable, de 60 Hz, nous avons dû présenter chaque image à une cadence de 30 images par seconde, alors qu'elles avaient été enregistrées à 25 images par seconde. La vitesse était donc accélérée d'un facteur 6/5 par rapport aux mouvements naturels présentés dans l'étude originale. En conséquence, les syllabes auditives ont dû être compressées d'un facteur équivalent afin de conserver la synchronisation des indices auditifs et visuels, tout en conservant le spectre fréquentiel du signal acoustique original. Cette compression temporelle a été réalisée grâce au logiciel Soundforge. Les syllabes résultant de cette transformation semblaient tout aussi naturelles que les syllabes originales, aussi bien sur le plan visuel qu'auditif.

Nous avons utilisé 4 couples de syllabes (/gu/-/ku/, /gy/-/ky/, /du/-/tu/, /dy/-/ty/) qui étaient toujours présentés dans des blocs expérimentaux différents, dans le but de conserver pour l'expérience MEG uniquement le couple de syllabes montrant l'effet comportemental le plus net. Chacune des 8 syllabes présentait une structure audiovisuelle différente, mais pour chaque paire de syllabe, le son de la syllabe voisée commençait toujours systématiquement plus tôt par rapport au début du mouvement des lèvres que celui de la syllabe non voisée, en raison du pré-voisement. Le schéma temporel des stimulations est illustré pour les syllabes /ku/ et /gu/ dans la figure 11.1. L'intensité de chacune des syllabes était ajusté de façon à ce que la puissance acoustique moyenne de la partie stationnaire du signal, correspondant à la voyelle, soit la même.

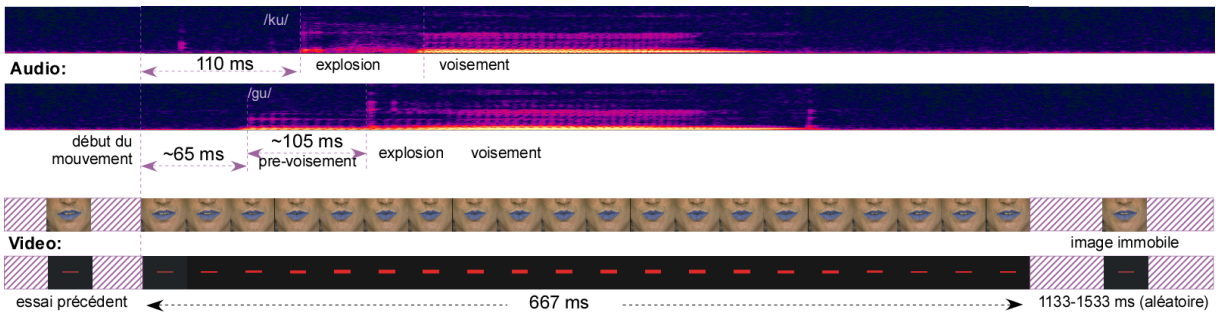

FIG. 11.1 – Structure temporelle des syllabes /ku/ et /gu/. Audio : variation temporelle du spectre fréquentiel entre 0 et 10000 Hz. Pour tous les couples de syllabes, le pré-voisement de la syllabe voisée commençait toujours avant l'explosion de la consonne de la syllabe non voisée. Les délais temporels indiqués sont les valeurs moyennes sur l'ensemble de 4 syllabes voisées et des 4 syllabes non voisées.

Dans les deux conditions bruitées, un bruit de foule continu était présenté pendant tout le bloc de stimulation. Le rapport signal (syllabe) sur bruit (foule) était calculé comme le rapport de la puissance moyenne pendant la partie stationnaire, correspondant à la voyelle,

sur la puissance moyenne du bruit. Dans la condition la plus bruitée, le rapport signal sur bruit était de -9 dB, dans la condition intermédiaire de 0 dB et dans la conditions sans bruit aucun bruit n'était présenté. Contrairement à nos premières études comportementales, les sons ont été présentés dans un casque à écouteurs afin d'imiter les conditions de stimulation dans la MEG.

Procédure

Dans tous les blocs expérimentaux, un essai commence avec la présentation d'un visage (ou un rectangle) au repos. Avec un intervalle interstimulus variant aléatoirement entre 1800 et 2200 ms, il entend une syllabe parmi deux syllabes possibles (variable "Voisement" : voisée ou non voisée). Cette syllabe est accompagnée ou non de l'articulation visuelle (variable "Modalité" : auditif et audiovisuel). Donc, en condition auditive seule, le sujet voit un visage (ou un rectangle) immobile.

La tâche du sujet consiste à cliquer le plus rapidement possible sur l'un des 2 boutons de la souris, chacun des boutons correspondant à une des 2 syllabes, ce qui revient à discriminer le voisement, sans que cela soit explicitement dit au sujet. Les sujets n'étaient pas informés que les mouvements labiaux ne donnaient aucune information sur l'identité de la syllabe et il leur était seulement demandé de fixer la bouche pendant toute l'expérience, sans préciser s'il fallait ou non se servir des indices visuels. Les associations bouton/voisement étaient constantes pour tous les couples de syllabes pour un sujet donné, mais contrebalancées entre les sujets.

Chaque bloc expérimental contenait 40 stimuli (10 syllabes de chacune des conditions suivantes : voisée auditive, voisée audiovisuelle, non voisée auditive et non voisée audiovisuelle)

En plus de ces 2 variables intrabloc, on manipulait 3 variables interbloc :

- le niveau de Bruit (-9dB, 0dB, sans bruit).
- la Nature de l'information visuelle (visage ou rectangle).
- les couples de syllabes voisée/non voisée (/gu/-/ku/, /gy/-/ky/, /du/-/tu/ ou /dy/-/ty/). Cette dernière variable n'entrait pas dans l'analyse statistique et les performances et TR étaient moyennés à travers les 4 couples.

Chaque sujet était donc soumis à 24 blocs de stimuli, dont l'ordre était aléatoire et différent pour chaque sujet.

Analyses

Deux ANOVA avec, pour facteurs, le niveau de bruit, le voisement, la modalité et la nature des informations visuelles on été réalisées, l'une sur le pourcentage d'erreurs moyen sur l'ensemble des 4 couples de syllabes et l'autres sur le TR moyen dans les essais justes. Les degrés de libertés ont été corrigés selon la méthode de Greenhouse-Geisser pour prendre en compte la non homogénéité éventuelle des variances. Lorsqu'une interaction était significative, des ANOVA était réalisées sur chacune des modalités de l'un des facteurs impliqués dans l'interaction, pour tester l'effet des autres facteurs impliqués, et ceci jusqu'à aboutir à des ANOVA à un seul facteur, où jusqu'à ce qu'aucune interaction ne soit significative.

11.2.2 Résultats

Performances

La figure 11.2 montre les performances de sujets en fonction des 4 facteurs expérimentaux. Comme on aurait pu le prédire, le pourcentage d'erreurs augmente significativement avec le niveau de bruit ($p < 0,001$). On observe un effet significatif du voisement sur le pourcentage d'erreur ($p < 0,04$), les sujets se trompant plus souvent sur les non-voisées que sur les voisées. Enfin, on observe une interaction significative entre les facteurs voisement et modalité ($p < 0,04$) indiquant que si les informations visuelles améliorent les performances pour les syllabes non-voisées, elles les dégradent pour les voisées. Mais si on teste maintenant l'effet de la modalité séparément pour les syllabes voisées et non voisées, il n'est significatif pour aucun des 2 types de syllabe. Aucun autre effet ou interaction n'est significatif.

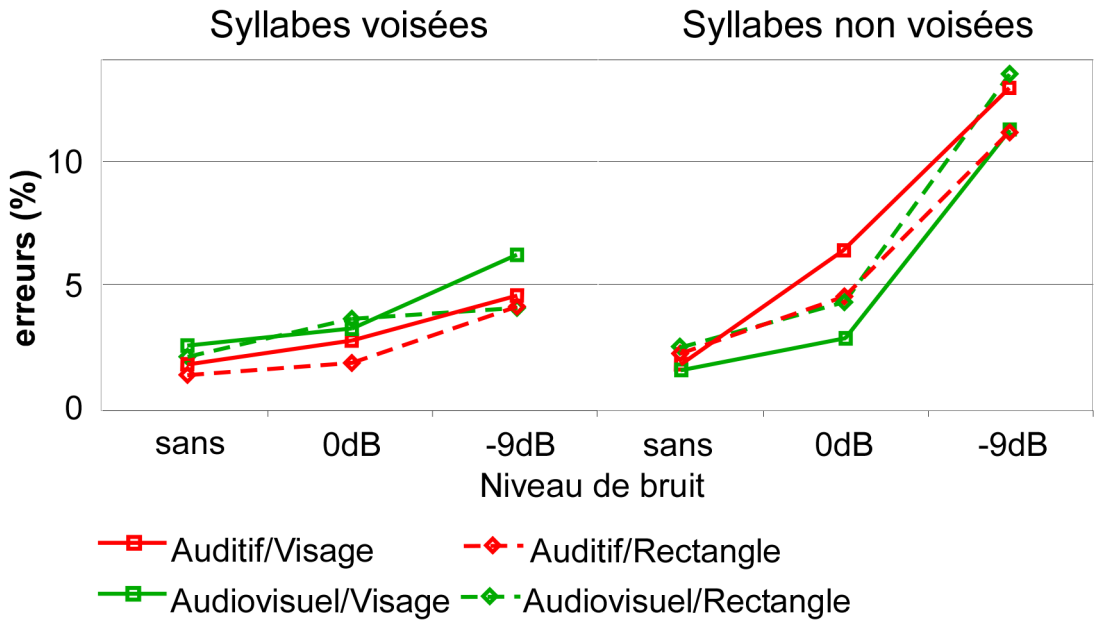

FIG. 11.2 – Pourcentage d'erreur dans la tâche de discrimination du voisement, en fonction du voisement, du niveau bruit, de la modalité de présentation et de la nature des informations visuelles.

Temps de réaction

La figure 11.3 page suivante présente les TR moyens pour les 24 conditions expérimentales testées. Comme nous l'avions prédit, on trouve un effet très significatif du bruit ($p < 0,0001$) sur le temps de traitement des syllabes qui augmente avec le niveau de bruit. L'effet du voisement est également présent ($p = 0,008$), les voisées donnant lieu à des temps de réaction plus courts, comme c'était prédictible étant donné que le début du son commençait plus tôt par rapport à l'instant où est mesuré le TR (le début du mouvement des lèvres), dans ces syllabes.

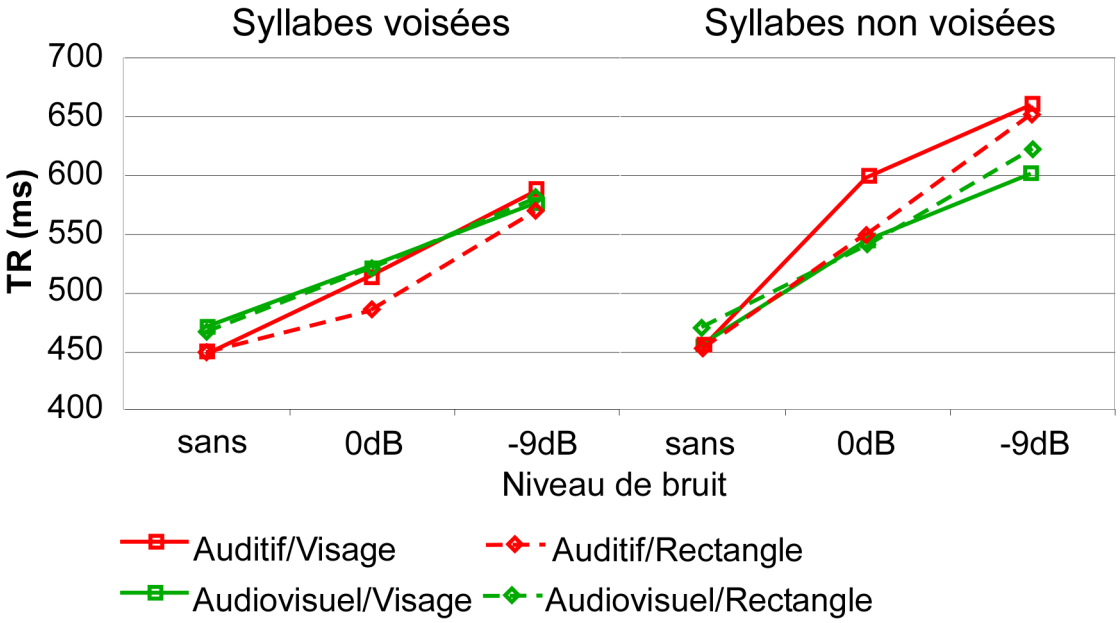

FIG. 11.3 – Temps de réaction dans la tâche de discrimination du voisement, en fonction du voisement, du niveau bruit, de la modalité de présentation et de la nature des informations visuelles.

Contrairement à nos hypothèse, l'effet principal de la modalité de présentation n'est pas significatif. Mais il semble, si l'on examine la figure 11.3, que cela soit dû au fait que l'effet de la modalité était différent selon le type de syllabe et le niveau de bruit. De fait, la triple interaction Voisement × Bruit × Modalité était marginalement significative ($p < 0,06$)

La figure 11.4 page ci-contre décrit cette interaction. Dans les 2 conditions bruitées, l'interaction entre les variables Modalité et Voisement est significative (-9dB : $p < 0,02$; 0dB : $p < 0,0004$). Dans la condition la plus bruitée, cette interaction indique un effet bénéfique des informations visuelles temporelles sur le temps de traitement, présent pour les syllabes non voisées ($p < 0,005$), mais pas pour les voisées. Dans la condition de bruit intermédiaire, l'interaction peut se décrire comme un effet opposé de la modalité sur les syllabes voisées et non voisées : on observe une diminution du TR avec les informations visuelles pour les syllabes non voisées ($p < 0,005$) et une augmentation du TR pour les syllabes voisées ($p < 0,06$).

Dans la condition sans bruit, l'interaction entre les facteurs Voisement et Modalité n'est pas significative, mais on observe un effet principal de la modalité se traduisant par une augmentation du TR dans la condition audiovisuelle par rapport à la condition auditive ($p < 0,04$). On n'observe en revanche pas d'effet significatif du voisement dans cette condition sans bruit.

Concernant l'interaction entre la présence d'informations visuelles temporelles et la nature de ces informations, nous avions prédit, sur la base des résultats antérieurs de Schwartz et coll., que la diminution de TR devrait être plus forte pour le visage que pour le rectangle, et que cette relation pouvait évoluer en fonction du niveau de bruit. La triple interaction Bruit × Modalité × Nature de informations était marginalement significative ($p < 0,07$).

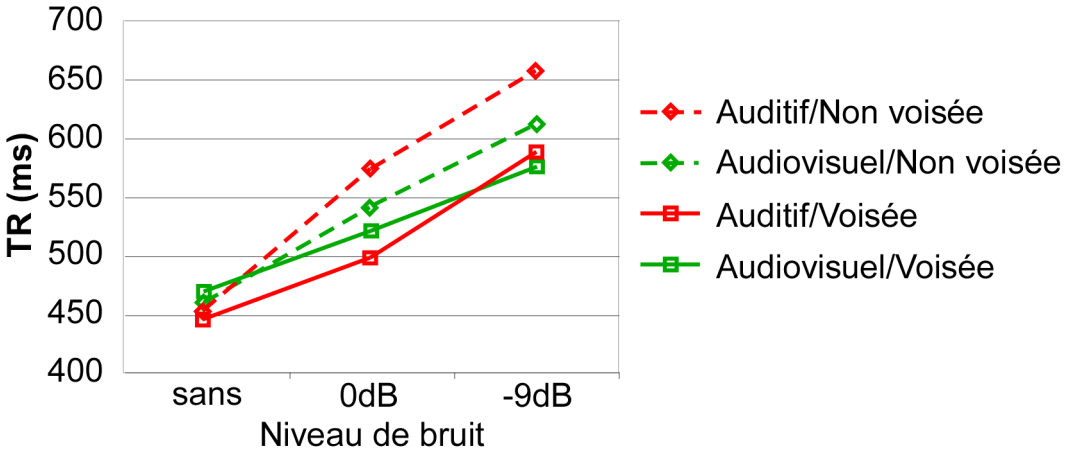

FIG. 11.4 – Temps de réaction dans la tâche de discrimination du voisement, en fonction du voisement, du niveau bruit et de la modalité de présentation.

La représentation graphique de cette interaction (figure 11.5) suggère en effet que le schéma d'interaction entre la présence et la nature de informations visuelles variait en fonction du niveau de bruit, mais d'un manière différente de celle à laquelle on aurait pu s'attendre.

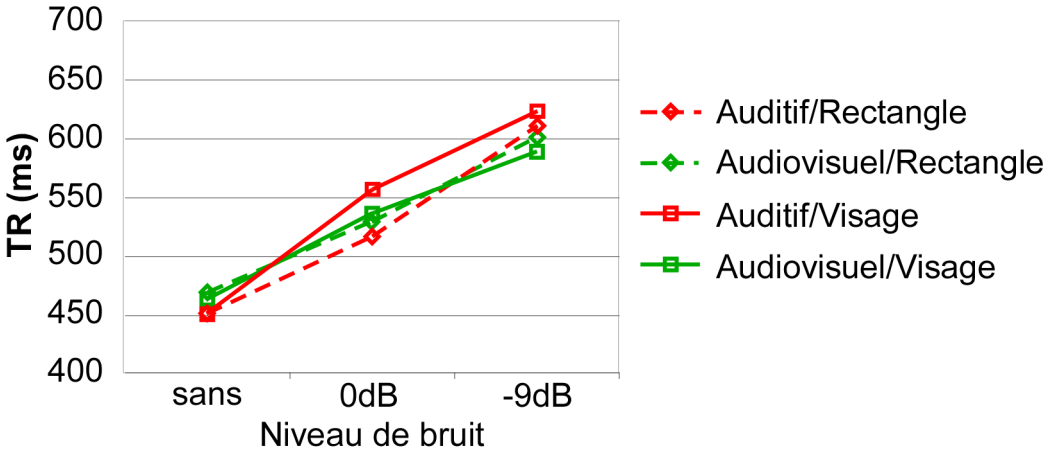

FIG. 11.5 – Temps de réaction dans la tâche de discrimination du voisement, en fonction du niveau bruit, de la modalité et de la nature des informations visuelles.

Dans les deux conditions de bruit, on trouve une interaction significative entre les facteurs Modalité et Nature des informations (-9dB : $p < 0,05$; 0dB : $p < 0,002$). Dans les deux cas, cette interaction va dans le sens prédit puisque le TR en audiovisuel est significativement inférieur au TR auditif dans le cas du visage (-9dB : $p < 0,02$; 0dB : $p < 0,03$) mais pas dans le cas du rectangle. Cependant, un aspect troublant de l'interaction est que dans la condition 0dB, l'effet semble être dû plus à une différence entre les temps de traitement des syllabes auditives (c'est-à-dire accompagnées par un rectangle ou un visage immobile $p < 0,02$) que par une facilitation plus forte du visage que du rectangle en condition audiovisuelle. Il est probable que ces différences entre conditions auditives seules

aient contribué en grande partie à la présence d'une interaction entre les facteurs Modalité et Nature des informations pour ces deux conditions de bruit.

Dans la condition sans bruit, on n'observait pas d'interaction entre les facteurs Modalité et Nature des informations, ni d'effet principal de la nature des informations visuelles, mais un effet principal de la modalité, sous la forme d'un cout de la condition audiovisuel ($p < 0,04$), déjà décrit plus haut.

Aucune autre interaction ou effet principal que ceux décrits n'étaient significatif. Jamais nous n'avons observé d'interaction entre les facteurs Voisement et Nature de l'information visuelle.

11.2.3 Discussion

L'analyse des performances n'indique que des effets faibles et peu significatifs de la modalité de présentation. Cette quasi-absence d'effet de la modalité pourrait s'expliquer par la variabilité intersujet importante du taux d'erreur.

En tout état de cause, le pourcentage d'erreur moyen observé était plus faible que celui trouvé par Schwartz et coll. (2004) avec les même stimuli et pourrait refléter la différence de tâche demandée au sujet. Discriminer entre deux syllabes est en effet plus facile qu'identifier une syllabe parmi 12, à niveau de bruit équivalent, et la simplicité de notre tâche pourrait être une seconde raison pour laquelle on n'a pas observé pas de facilitation de la performance avec l'apport d'information visuelle temporelle.

L'aide apportée par les informations visuelles temporelles a en revanche été répliquée sur les TR, mais uniquement pour les syllabes non voisées dans les deux conditions de bruit. De plus, dans ces deux conditions, on trouvait une interaction entre la présence d'informations visuelles temporelles et la nature de ces informations, mais cet effet semblait autant venir d'une diminution du TR pour le visage en mouvement par rapport au visage immobile que d'une augmentation du TR pour le rectangle en mouvement par rapport au rectangle immobile.

Par ailleurs, l'effet des informations visuelles temporelles change selon le niveau de bruit. De manière générale, il semble rester vrai que plus le niveau de bruit est important, plus les informations visuelles sont utiles, mais ces effets s'expriment différemment pour les syllabes voisées et non voisées. Pour les syllabes non voisées, en augmentant le niveau de bruit, on passe d'une situation où les indices visuels n'aident pas à une situation où ils diminuent le TR. À l'inverse, pour les syllabes voisées, en augmentant le niveau de bruit, on passe d'une situation où les indices visuels augmentent le TR à une situation où le TR pour les syllabes auditives et audiovisuelles est équivalent. Peut-être en augmentant encore le niveau de bruit, observerait-on une amélioration du TR pour les syllabes voisées également.

Cette triple interaction peut avoir plusieurs explications : d'une part les syllabes voisées ont une puissance spectrale totale plus importante que celle des syllabes non-voisées (la zone stationnaire du signal dure plus longtemps), ce qui peut expliquer pourquoi elles sont plus facilement détectables dans le bruit, comme on peut le constater au niveau des

performances. De ce fait il est possible que leur traitement bénéficie moins de la présence des indices visuels. D'autre part, le délai séparant le début des indices visuels et auditifs est différent pour les voisées et les non voisées. Or plusieurs études ont montré des effets d'intégration multisensorielle différents selon le délai séparant les informations des deux modalités (Ghazanfar, Maier, Hoffman & Logothetis, 2005 ; Lakatos, Chen, O'Connell, Mills & Schroeder, 2007).

Enfin, la triple interaction entre voisement, bruit et modalité se traduit également par une convergence des TR des différentes combinaisons voisement/modalité dans la condition sans bruit : cet effet pourrait s'expliquer soit par un effet plancher, soit une différence de stratégie. Selon la seconde explication, les mécanismes de discrimination du voisement seraient des plus efficaces dans la condition sans bruit et, par conséquent, le traitement des voisées et non voisées prendrait des temps équivalents tout en laissant peu l'occasion aux mécanismes d'intégration de se manifester. Dans les conditions bruitées, au contraire, la discrimination du voisement reposerait beaucoup plus sur la détection de la présence ou de l'absence d'un prévoisement, qui pourrait être plus sensible à la présence d'informations visuelles temporelles.

Deux aspects des données jettent toutefois le doute sur l'interprétation des résultats obtenus. Il s'agit d'une part du fait que dans certaines conditions (dans la condition sans bruit et, pour les syllabes voisées, dans la condition de bruit intermédiaire), on observait une augmentation des TR dans la condition audiovisuelle par rapport à la condition auditive, et d'autre part, de la différence de TR observée entre les deux conditions auditives seules.

Ces effets suggèrent que les conditions auditives choisies n'étaient pas de bons contrôles, dans la mesure où le type d'informations visuelles présentes à l'écran semble influer sur le temps de traitement de la syllabe bien qu'il ne donne aucune information sur le voisement, pas même une information temporelle.

Cette différence entre les conditions rectangle et visage pourrait être due à des différences de stratégie : en effet la variable Nature des informations est une variable interbloc et il est tout à fait envisageable que les sujets aient traité différemment les stimuli (A et AV) selon que le contexte était celui d'un visage ou celui d'un rectangle. Un visage qui prononce une syllabe une fois en remuant les lèvres, une fois sans les bouger n'a pas le même sens que des syllabes accompagnées ou non du mouvement d'un rectangle. L'interaction entre la présence d'informations visuelles et leur nature est donc difficile à interpréter du fait de la présence possible d'un effet de bloc. Afin de mieux étudier l'effet de la nature des informations visuelles sur l'effet de la modalité et de confirmer la présence d'un cout de l'ajout d'informations temporelles visuelles, nous avons mené une nouvelle expérience comportementale.

11.3 Expérience comportementale 2

Dans l'expérience précédente, la présence d'un cout audiovisuel et d'un effet de la nature des informations statiques sur le temps de traitement des syllabes "auditives" nous a incité à la prudence quant à nos conclusions.

En effet, dans la mesure où les conditions auditives montraient des différences significatives entre les conditions visage et rectangle, on peut mettre en doute l'interprétation des effets en termes de bénéfices ou de cout des informations temporelles visuelles. Il se pourrait en effet que la simple présence d'un visage au repos, même immobile, accélère la discrimination du voisement. Il nous fallait donc trouver un meilleur contrôle auditif seul. Nous avons ajouté une condition auditive seule dans laquelle l'écran était totalement vide pendant la présentation de la syllabe. Et, pour éviter les effets de blocs, nous avons présenté les 5 conditions dans un même bloc expérimental : la condition auditive seule, les deux conditions audiovisuelles statiques dans laquelle seule une bouche ou un rectangle au repos était présenté pendant la stimulation auditive (conditions auditives de l'expérience précédente) et les deux conditions audiovisuelles dynamiques dans lesquelles le mouvement du visage ou du rectangle donnaient une information temporelle sur la syllabe.

11.3.1 Méthodes

Sujets

Neuf sujets droitiers (dont 5 de sexe féminin) âgés en moyenne de 27,5 ans (écart-type : 4 ans) ont passé cette expérience. Huit de ces sujets avait passé l'expérience 1 deux mois auparavant.

Stimuli

Les stimuli utilisés étaient identiques à ceux de l'expérience 1, excepté que nous n'avons employé qu'un seul couple couple de syllabe (les syllabes /ku/ et /gu/), afin d'éliminer une source de variabilité des TR. Ce couple a été choisi pour la ressemblance des effets sur les TR présentés par ce seul couple avec les effets estimés sur la moyenne des 4 couples de syllabes dans l'expérience précédente. Pour des raisons qui seront exposées ci-dessous, le rectangle rouge était présenté sur un fond gris au lieu d'un fond noir. Dans la condition auditive seule, la syllabe auditive était présentée avec un fond visuel gris uni.

Procédure

Le sujet devait donc réaliser la tâche de discrimination des syllabes voisées et non voisées dans 5 conditions visuelles mélangées aléatoirement : écran gris (auditif seul), visage statique, rectangle statique, visage dynamique, rectangle dynamique.

Contrairement à la première expérience, le changement de condition au sein d'un bloc nécessitait l'apparition et la disparition brusque des stimuli visuels (passage d'un essai visage à un essai rectangle ou auditif seul, par exemple). Pour éviter que le début d'un essai donne plus d'informations temporelles dans une condition que dans une autre, les essais au sein d'un bloc étaient séparés par un écran noir pendant 150 ms. Ainsi, la prédictibilité du stimulus auditif était identique pour les cinq conditions : au moment ou l'écran noir disparait, apparait soit un visage, soit un rectangle sur fond gris, soit un fond gris seul. Cependant, on ne voulait pas que cette information temporelle à elle seule aide le sujet à détecter le voisement. Quel intérêt y aurait-il alors à exploiter les informations visuelles dynamiques ? Afin de limiter la prédictibilité temporelle de la syllabe et de favoriser la

capacité du mouvement visuel (que ce soit celui du rectangle ou du visage) à fournir de l'information temporelle, nous avons introduit une période aléatoire (variant entre 300 et 750 ms) entre l'apparition de l'image immobile et le début de la syllabe auditive et/ou du mouvement articulatoire. Un essai se terminait par un période aléatoire et l'intervalle interstimulus moyen était de 2000 ms. La structure temporelle d'un essai est illustrée dans la difugre 11.6.

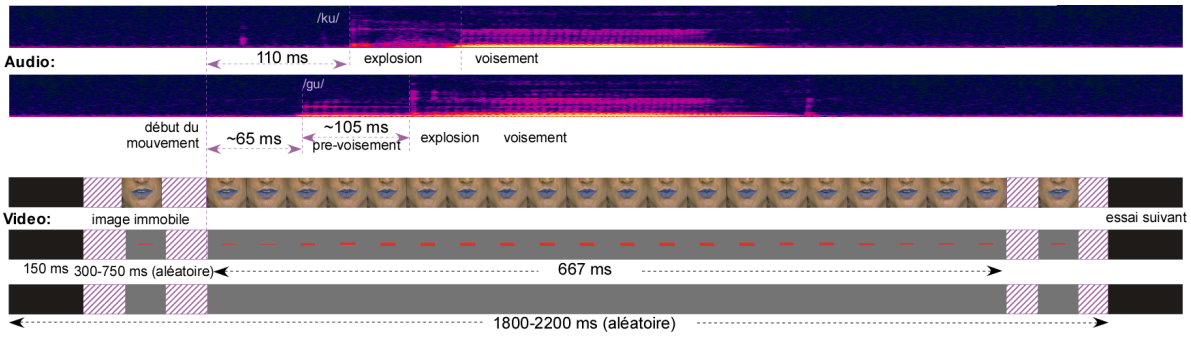

FIG. 11.6 – Structure temporelle des syllabes /ku/ et /gu/. Audio : variation temporelle du spectre fréquentiel entre 0 et 10000 Hz. Pour tous les couples de syllabes, le pré-voisement de la syllabe voisée commençait toujours avant l'explosion de la consonne de la syllabe non voisée. Vidéo : stimuli visuels des conditions audiovisuelles dynamiques (visage ou rectangle) et de la condition auditive seule. dans tous les cas, un écran noir précédait la présentation du visage, du rectangle ou de l'écran gris. Les délais temporels indiqués sont les valeurs moyennes sur l'ensemble de 4 syllabes voisées et des 4 syllabes non voisées.

Afin d'étudier plus finement la variation des effets avec le niveau de bruit, nous avons utilisé 5 niveaux de bruit : sans bruit, 0dB, -4,5dB, -9dB et -13,5dB. Le niveau de bruit le plus fort devrait permettre d'observer une facilitation de TR pour les syllabes voisées. Les différents niveaux de bruit étaient présentés dans des blocs différents.

Chaque sujet passait 20 blocs de stimulation, soit 4 blocs de chaque niveau de bruit. Un bloc comprenait 5 syllabes voisées (/gu/) et 5 syllabes non voisées (/ku/) dans chacune des 5 conditions de présentation, pour un total de 50 syllabes.

Analyses

Pour cette expérience nous n'avons analysé que les TR, dans les essais où les sujets n'avaient pas commis d'erreur. On a effectué deux types d'analyse sur les temps de réaction.

- on a analysé les données des 4 conditions déjà présentes dans l'expérience 1 avec la même ANOVA à 4 facteurs : Bruit × Voisement × Modalité × Nature des informations visuelles, sans prendre en compte les essais auditifs seuls. Cela permet d'évaluer l'effet de la présentation aléatoire par rapport à la présentation par bloc des rectangles et des visages. Notons tout de même quelques différences supplémentaires entre les 2 protocoles : utilisation d'un seul couple de syllabes, présence de 5 niveaux de bruit et sujets plus familiers avec les stimuli et la tâche (les mêmes sujets ont en effet en majorité participé aux deux expérience).
- Afin d'évaluer l'existence de bénéfices et éventuellement de couts dans les conditions audiovisuelles dynamiques et statiques, on a testé la significativité de la différence

entre chacune des 4 combinaisons Modalité × Nature des informations et la condition auditive seule, ainsi que l'interaction de cet effet avec les variables bruit et voisement. On a donc réalisé, pour chacune des conditions visage dynamique, visage statique, rectangle dynamique et rectangle statique, une ANOVA Présence d'informations visuelles (statique ou dynamique) × Bruit × Voisement.

Tous les tests ont été corrigés pour la non sphéricité des données par la méthode de Greenhouse-Geisser.

11.3.2 Résultats

ANOVA Bruit × Voisement × Modalité × Nature

On retrouve l'effet attendu du bruit sur les temps de réaction ($p < 0,0001$), ainsi que l'effet du voisement, les voisées donnant lieu à des TR plus rapides que les non voisées (p=0,0007). Ces deux effets et leur interaction sont décrits dans la figure 11.7.

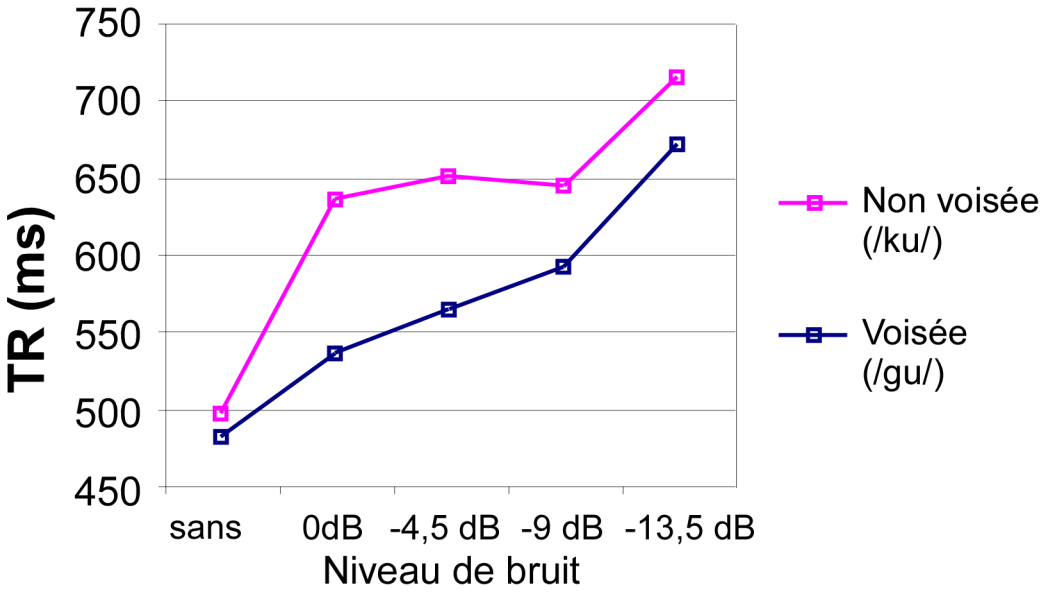

FIG. 11.7 – Temps de réaction dans la tâche de discrimination du voisement, en fonction du voisement et du niveau bruit.

L'interaction entre ces deux facteurs est significative ($p < 0,0001$) et semble s'expliquer par le fait que la différence entre voisées n'existe que pour les conditions bruitées (0dB : $p < 0,0001$; 4,5dB : $p = 0,0004$; 9dB : $p < 0,02$; 13,5 dB : $p < 0,02$).

Contrairement à l'expérience 1, le facteur voisement n'interagissait avec aucun autre facteur de l'analyse.

Par contre, comme dans l'expérience 1, la triple interaction Modalité × Nature × Bruit était marginalement significative ($p < 0,08$). Cette interaction est décrite dans la figure 11.8 page ci-contre.

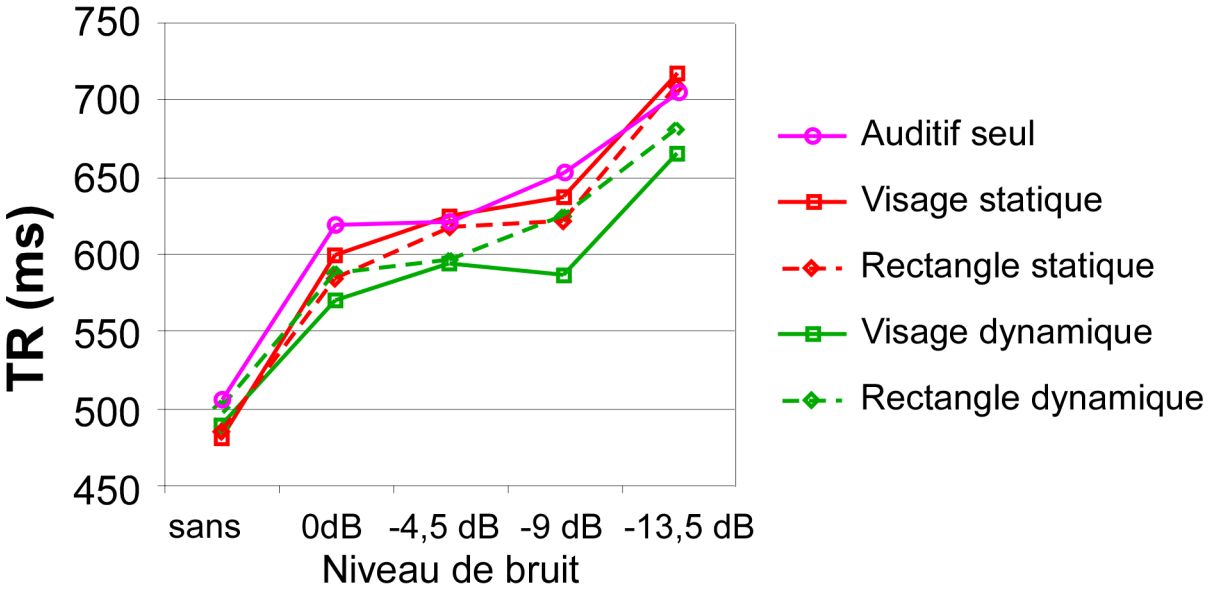

FIG. 11.8 – Temps de réaction dans la tâche de discrimination du voisement, en fonction du niveau bruit, de la modalité et de la nature des informations visuelles.

Nous avons testé l'interaction Modalité × Nature des informations dans chacune des conditions de bruit. Dans la condition sans bruit, l'interaction n'est pas significative. L'effet de la modalité est marginalement significatif et s'exprime par une augmentation du TR pour les conditions audiovisuelles dynamiques par rapport aux conditions audiovisuelles statiques ($p < 0,08$).

Dans la condition 0dB, l'interaction significative (p=0,01) se manifeste autant par un cout significatif du visage immobile par rapport au rectangle immobile ($p < 0,02$) que par un gain du visage en mouvement par rapport au rectangle en mouvement ($p = 0,05$).

Dans la condition 4,5dB, l'interaction n'est pas significative et on trouve un effet principal de la modalité de présentation qui s'exprime par une diminution des TR avec les informations visuelles dynamiques ($p < 0,02$).

Dans la condition 9dB, l'interaction est significative ($p < 0,01$) et se traduit par un cout marginalement significatif du visage immobile ($p < 0,07$) et un gain très significatif du visage mobile (p=0,004) par rapport au rectangle.

Enfin dans la condition 13,5 dB, l'interaction était marginalement significative ($p < 0,07$) et s'expliquait par un avantage marginalement significatif du visage dynamique par rapport au rectangle dynamique ($p < 0,04$), le cout pour le visage immobile par rapport au rectangle immobile n'étant pas significatif.

Test des couts et bénéfices

Dans chacune des 4 conditions visage dynamique, visage statique, rectangle dynamique et rectangle statique, on a retrouvé les effets significatifs du bruit, du voisement ainsi que leur interaction, déjà décrits. De même que dans l'analyse précédente, le voisement n'interagissait jamais avec le facteur Présence d'information visuelle, dans aucune des 4

conditions. La figure 11.9 présente donc la différence de TR entre les 4 conditions audiovisuelles et la condition auditive seule en fonction du niveau de bruit, moyennée sur le type de voisement.

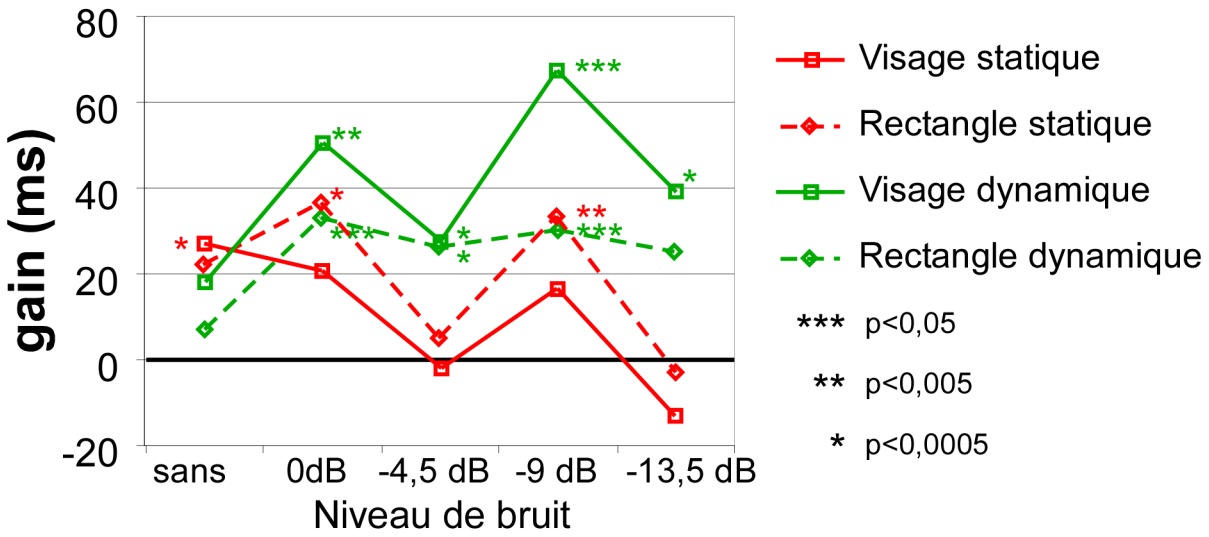

FIG. 11.9 – Bénéfices et cout du TR par rapport à la condition auditive seule dans la tâche de discrimination du voisement, en fonction du niveau bruit, de la modalité et de la nature des informations visuelles. Les étoiles indiquent les conditions dans lesquelles l'effet de la présence d'un stimulus visuel (statique ou dynamique) est significatif.

On peut constater que, dans presque toutes les conditions, cette différence prenait l'aspect d'un bénéfice. Pour la condition visage dynamique, même si l'effet de la présence d'informations visuelles était très significatif ($p < 0,0001$), il interagissait significativement avec le niveau de bruit ($p < 0,04$) : le bénéfice apporté par les informations visuelles était significatif pour toutes les conditions de bruit mais pas dans la conditions sans bruit (0dB : $p < 0,003$; 4,5dB : $p < 0,04$; 9dB : $p < 0,0001$; 13,5dB : $p < 0,008$).

Pour la condition rectangle dynamique, l'interaction entre la présence d'information visuelle et le bruit n'était pas significative et l'effet global de l'information visuelle était très significatif ($p < 0,0001$). Donc excepté dans la condition sans bruit pour le visage, la présence d'informations visuelles temporelles dynamiques diminuait le TR par rapport à une condition auditive seule.

Dans les deux conditions où l'information visuelle consistait en la simple présence d'une image immobile, l'interaction entre la présence d'information visuelle et le niveau de bruit était significative (visage $p < 0,02$; rectangle : $p < 0,008$).
Dans la condition visage statique, le bénéfice était plus ou moins significatif selon le niveau de bruit (sans bruit : $p < 0,04$; 0dB : $p < 0,10$; 9dB : $p < 0,10$).
Dans la condition rectangle statique, on observe exactement le même schéma d'interaction, en plus significatif (sans bruit : $p < 0,06$; 0dB : $p < 0,03$; 9dB : $p < 0,003$).

Ajoutons enfin qu'aucune de ces 20 conditions Modalité × Nature × Bruit ne montre de cout de la condition audiovisuelle (statique ou dynamique) par rapport à la condition auditive seule.

11.3.3 Discussion

En dépit du fait que toutes les conditions audiovisuelles (visage, rectangle, statique, dynamique) étaient mélangées dans cette expérience, on continue à observer d'une part des TR plus rapides pour les conditions audiovisuelles statiques que pour les conditions audiovisuelles dynamiques dans la condition sans bruit et d'autre part, des TR globalement plus rapides lorsque les syllabes sont présentées associées à un rectangle statique que lorsqu'elles sont présentées avec un visage statique.

Ces effets sont toutefois moins significatifs que dans l'expérience 1, peut-être parce que le nombre de sujets est moins élevé et/ou parce qu'ils sont atténués par le mélange des conditions visage et rectangle.

Cependant la condition auditive seule nous a permis de montrer que l'ajout d'informations visuelles, qu'elles soient temporellement informatives (dynamiques) ou (statiques), ne se traduit jamais par un cout en termes de temps de traitement. Si l'on ne considère que les conditions audiovisuelles dynamiques et la condition auditive seule, nous avons donc montré une diminution du temps de réaction dans la discrimination du voisement lorsque les sujets disposent d'informations temporelles pouvant les aider à détecter le prévoisement par rapport à une condition où aucun stimulus visuel n'est présenté. Cette effet existe aussi bien pour des informations temporelles dynamiques fournies par un rectangle que par un visage, mais uniquement lorsque la discrimination est rendue plus difficile par la présence de bruit. Cependant lorsque ces informations temporelles sont fournies par un visage, la diminution du temps de réaction est plus importante que lorsqu'elles sont fournies par un rectangle, au moins dans deux conditions de bruit[1]. Cet effet semble donc être en partie spécifique aux indices visuels de parole et représente un réplication de l'effet mis en évidence par Schwartz et coll. (2004) sur les performances.

Toutefois on observe également une diminution du TR par rapport à la condition auditive seule lorsque l'on ajoute un visage ou un rectangle statique (les anciennes conditions "auditives" de l'expérience 1), au moins dans les conditions les moins bruitées. On peut en conclure, d'une part, que ces conditions ne constituaient vraisemblablement pas de bon contrôles pour étudier l'effet d'indiçage temporel et, d'autre part, que cette diminution du TR n'est pas de aux informations temporelles. En effet les deux conditions audiovisuelles statiques donnaient exactement les mêmes informations temporelles que la condition auditive seule. On peut donc exclure qu'il s'agisse de quelque effet d'indiçage temporel.

De plus ce bénéfice inattendu des stimuli visuels statiques semble être plus important pour les rectangles que pour les visages. Il ne s'agit donc pas simplement d'un effet attentionnel non spécifique, ou alors il faudrait expliquer pourquoi cet effet est plus fort pour un rectangle qu'une visage. Il se pourrait que cet effet représente la conjonction d'un effet attentionnel non spécifique qui aurait tendance à diminuer le TR et d'un effet d'incongruité des stimuli auditifs et visuels qui aurait tendance à augmenter le TR, l'incongruité d'un

[1] Curieusement, cet effet d'interaction n'est observé que pour les conditions de bruit 0db et -9dB, alors qu'il n'existe pas dans la condition -4,5dB et n'est pas significatif dans la condition -13,5dB. Dans ces deux dernières conditions, le bénéfice associé à la présences d'informations dynamiques est d'ailleurs plus faibles que dans les deux autres. La seule différence entre ces conditions était que les sujet avaient déjà été confrontés au niveaux de bruit 0dB et -9dB dans la première expérience.

visage immobile et d'un son de parole étant plus forte que celle d'un rectangle et d'un son de parole.

En tout état de cause, ce bénéfice semble diminuer avec le niveau de bruit, au contraire du bénéfice dû aux informations visuelles dynamiques, ce qui suggère que les indices visuels dynamiques améliorent spécifiquement la détection du prévoisement dans le bruit, alors que l'effet de la présence d'un stimulus visuel statique influencerait plutôt des processus plus généraux et non liés à la perception de la parole.

Une autre différence entre l'expérience 1 et l'expérience 2 est la disparition de l'effet d'interaction entre le voisement et la présence d'informations visuelles : cette disparation peut être due à une perte de puissance statistique due au nombre moins important de sujets, mais également au fait que les TR ont été mesurés pour un seul couple de syllabe.

11.4 Discussion générale

L'objectif initial des ces expériences comportementales étaient d'adapter le protocole de Schwartz et coll. (2004) à une expérience électrophysiologique. Nous avons montré que l'effet d'indiçage temporel des mouvements pré-phonatoires sur la perception du voisement pouvait être mis en évidence sur les temps de réaction dans une tâche de discrimination entre une syllabe voisée et une syllabe non voisée. Ce paradigme, plus simple, pourrait permettre d'étudier les corrélats électrophysiologiques à l'origine de cet effet, en enregistrant les potentiels évoqués par les mouvements articulatoires, une syllabe voisée et une syllabe voisée accompagnée des mouvements articulatoires.

On pourrait, à l'aide du modèle additif, étudier l'influence des informations visuelles temporelles non phonétiques sur le potentiel évoqué par le prévoisement dans le cortex auditif ou d'autres structures temporales. Si cet effet se traduit par une diminution de l'onde N1 auditive, on aurait un argument pour dire que l'effet observé dans notre première expérience électrophysiologique représenterait plutôt un effet d'indiçage temporel qu'une véritable intégration audiovisuelle phonétique. Dans le cas contraire, il serait plus difficile de conclure, étant donné la différence de structure audiovisuelle et acoustique des stimuli utilisés dans les deux paradigmes. L'expérience MEG que nous avions prévue au départ n'a malheureusement pas pu être réalisée, faute de temps.

Néanmoins, nos résultats comportementaux suggèrent que l'effet de pur indiçage temporel ne s'observe que lorsque la tâche des sujets consiste à détecter le pré-voisement dans le bruit et non lorsqu'il s'agit de discriminer le voisement dans de bonnes conditions acoustiques. À ce stade de nos investigations, c'est un argument supplémentaire pour dire que la diminution du TR observée dans nos expériences d'EEG était bien due à une intégration audiovisuelle phonétique et non à cet effet d'indiçage temporel, car notre expérience électrophysiologique était réalisée sans bruit acoustique et montrait néanmoins une diminution robuste du TR pour la discrimination des syllabes audiovisuelles par rapport aux syllabes auditives. À l'appui de cette affirmation, Callan et coll. (2004) ont montré en IRMf des effets d'interaction audiovisuelle dans la perception de la parole dans le STG/STS spécifiques aux informations visuelles de haute fréquence spatiale et qui ne sont pas trouvées pour des informations visuelles basse-fréquence, qui pourtant donnent une information tem-

porelle. À l'inverse, certains effets d'interaction audiovisuels très précoces sur les potentiels évoqués auditifs du tronc cérébral (Musacchia, Sams, Nicol & Kraus, 2006), similaires à des effets attentionnels, peuvent difficilement s'expliquer par une intégration phonétique et sont probablement dus à l'avance temporelle des informations visuelles sur les informations auditives.

Un résultat frappant et inattendu de nos deux expériences comportementales est que la simple présentation d'un stimulus visuel, ne fournissant aucune information pertinente, même temporelle, pour la tâche auditive à réaliser, semble diminuer le TR pour effectuer cette tâche. Cet effet n'est pas sans rappeler l'effet d'un stimulus accessoire sur le temps de traitement d'un stimulus dans une autre modalité (voir partie 2.3.2 page 34). Il était néanmoins assez faible et nécessiterait d'être répliqué et étudié plus en détail. Tout ce qu'on peut en dire pour l'instant c'est qu'il constitue une nouvelle preuve de l'interdépendance des traitements auditifs et visuels.

Quatrième partie

Interactions audiovisuelles en mémoire sensorielle

Chapitre 12

Introduction générale

12.1 MMN Auditive

La négativité de discordance (Mismatch Negativity, MMN) est une onde des potentiels évoqués auditifs, observée en réponse à tout changement sonore dans un environnement de stimulation répétitive. On peut l'observer dans un protocole dit *oddball* : on présente au sujet une suite de sons identiques ("standards") dans lesquels on introduit occasionnellement des sons "déviants" (Näätänen, Gaillard & Mantysalo, 1978). La MMN est observée quelle que soit la nature du trait acoustique déviant par rapport aux standards (la hauteur tonale, la durée, l'intensité, la localisation, etc...), aussi bien lorsque le sujet prête attention aux stimuli que lorsque son attention est dirigée vers une autre tâche ou une autre modalité sensorielle. La détection par le cerveau d'un changement dans l'environnement implique la conservation d'une trace physiologique du stimuli précédents. La MMN reflèterait donc un processus automatique de discordance neuronale entre cette trace mnésique des stimuli passé et l'entrée d'un nouveau stimulus implique. La MMN est en partie générée dans le cortex auditif secondaire (par exemple Kropotov et coll., 2000).

Plusieurs autres interprétations non mnésiques de la MMN ont été exclues, par exemple que la différence de traitement du son standard et du son déviant provienne de la différence physique entre les stimuli et donc de l'activation de populations de neurones partiellement différentes. Cette interprétation peut facilement être rejetée en comparant le potentiel évoqué par le même stimulus dans un contexte où il est standard et dans un contexte où il est déviant : la différence entre ces deux conditions révèle toujours l'existence d'une MMN.

Une autre hypothèse qui n'implique pas l'existence d'une trace mnésique est que la MMN reflèterait la différence de fréquence d'apparition des stimuli standards et déviants. Ainsi, si la population de neurones répondant au stimulus répond d'autant moins que le stimulus est présenté souvent, en raison par exemple de l'existence d'une période réfractaire, la moyenne des réponses au stimulus déviant devrait être différente de la moyenne des réponse au stimulus standard, même si ces sons sont identiques. Cette hypothèse de *refractoriness* peut être rejetée en comparant la réponse au même son, dans le cas où il est déviant parmi des sons standards et dans une condition appelée équiprobable dans laquelle il est présenté, avec la même probabilité, parmi plusieurs stimuli différents ayant la même fréquence de présentation (Schröger & Wolff, 1996). Dans ce cas, on continue à observer

une MMN. Donc le même stimulus, présenté avec la même fréquence d'apparition, mais dans un cas où il brise une régularité (lorsqu'il est présenté dans une suite de standards) et dans un cas où il ne brise aucune régularité (la condition équiprobable) donne lieu à des traitements différents qui ne peuvent être attribués qu'à l'effet de l'organisation des autres stimuli de la séquence, en l'occurrence la répétition des sons standards.

Ainsi contrôlée, l'observation d'une MMN implique donc l'existence d'une représentation mnésique du son standard à laquelle le son déviant est comparé. Cette représentation mnésique est souvent assimilée à la mémoire sensorielle ou échoïque, mise en évidence de façon comportementale dans l'effet de récence lors d'une tâche de rappel ou l'effet de masquage auditif (Hawkins & Presson, 1986) et des tentatives ont été faites de lier la représentation indexée par la MMN et la mémoire échoïque (Cowan, Winkler, Teder & Näätänen, 1993 ; Winkler, Reinikainen & Näätänen, 1993). Il existe toutefois d'autres candidats électrophysiologiques à la corrélation avec la mémoire échoïque. Certains auteurs proposent ainsi que l'existence de périodes réfractaires, en particulier dans le cas de l'onde N1, peut être interprété comme un phénomène mnésique et sous-tendre la mémoire échoïque (Lu, Williamson & Kaufman, 1992b, 1992a ; McEvoy, Levänen & Loveless, 1997). Cette question étant loin d'être tranchée, on utilisera donc le terme de mémoire sensorielle auditive au sens de "ce qui est indexé par la MMN", sans faire d'hypothèse sur une correspondance avec la mémoire sensorielle mise en évidence avec des techniques comportementales.

12.2 Rappel de la problématique

Récemment, l'interprétation de la nature des représentations mnésiques reflétées par la MMN a été révisée par certains auteurs. En effet, de nombreuses études ont montré que la MMN n'est pas générée uniquement lorsque standards et déviants diffèrent sur un ou plusieurs traits acoustiques élémentaires, mais également lors de violations de régularités acoustiques plus complexes impliquant des relations entre plusieurs stimuli auditifs (par exemple : Horvath, Czigler, Sussman & Winkler, 2001 ; Korzyukov, Winkler, Gumenyuk & Alho, 2003 ; Tervaniemi, Maury & Näätänen, 1994) ou plusieurs traits élémentaires d'un même stimulus (violation d'une conjonction de 2 traits, voir la partie 16.1 page 205 ; Paavilainen, Simola, Jaramillo, Näätänen & Winkler, 2001). Ces données, entres autres, ont mené à l'idée que la mémoire sensorielle auditive indexée par la MMN a pour fonction de représenter toute régularité dans un environnement sonore complexe. Le rôle fonctionnel de cette représentation serait de détecter n'importe quelle anomalie de cet environnement sonore pouvant représenter une menace ou intérêt pour l'organisme (Winkler, Karmos & Näätänen, 1996).

Dans l'introduction de cette thèse, nous avons avancé divers résultats neuro-anatomiques, comportementaux et neurophysiologiques suggérant que les informations visuelles pouvaient influencer des traitements spécifiques à la modalité auditive. De même, dans la première partie expérimentale de cette thèse, nous avons montré que la vision pouvait moduler l'activité auditive à des étapes relativement précoces du traitement dans le cortex auditif, dans le cas particulier, il est vrai, de la perception de la parole. Puisque des traitements, censés être purement auditifs, sont en réalité influencés par la vision, cette influence pourrait avoir des répercussions sur la représentation en mémoire sensorielle de

l'environnement sonore et en particulier de ses régularités. La question que nous posons dans cette deuxième partie expérimentale est la suivante : les régularités audiovisuelles sont-elles représentées en mémoire sensorielle auditive ? Autrement dit, si un stimulus auditif est constamment associé à un stimulus visuel, cette composante visuelle va-t-elle être incluse dans la représentation du son en mémoire sensorielle auditive ?

Il existe plusieurs façons d'aborder cette question : notre première approche sera comportementale et exploitera le lien qui existe entre la mémoire sensorielle et la détection de la déviance. Dans les trois expériences suivantes, nous étudierons la question de la représentation d'une régularité audiovisuelle en étudiant diverses influences visuelles possibles sur le marqueur électrophysiologique de la mémoire sensorielle auditive : la MMN.

Chapitre 13

Détection d'une déviance audiovisuelle : étude comportementale

13.1 Introduction

Il a été montré à plusieurs reprises que les performances dans une tâche de détection d'un stimulus déviant présenté parmi des stimuli standards, étaient corrélées aux caractéristiques de la MMN automatiquement évoquée par ces déviants lorsque les sujets n'y prêtent pas attention. Ainsi, Tiitinen, May, Reinikainen et Näätänen (1994) ont montré, d'une part, que la latence de la MMN à une déviance fréquentielle et le temps de détection des mêmes déviants décroissaient avec l'amplitude de la déviance de manière identique et, d'autre part, étaient fortement corrélés. Par ailleurs, Novitski, Tervaniemi, Huotilainen et Näätänen (2004) ont montré que l'amplitude et la latence de la MMN à une déviation fréquentielle étaient corrélées à la fois au temps de détection et au taux de détection de cette déviation, la MMN étant d'autant plus grande que les performances sont bonnes. Ces résultats suggèrent que les performances comportementales dans la détection d'un stimulus déviant présenté parmi des stimuli distracteurs standards sont directement liées aux processus indexés par la MMN (voir Schröger, 1997, pour une revue). Une façon d'étudier si ces processus, qui, on l'a vu, mettent en jeu la mémoire sensorielle auditive, peut être influencé par des informations visuelles, est de comparer les temps de détection d'une déviation auditive et d'une déviation audiovisuelle d'un évènement audiovisuel standard. Si le temps de détection d'un déviation audiovisuelle est plus rapide que celui d'une déviation auditive, c'est que la dimension visuelle du stimulus entre en compte dans le processus de comparaison aboutissant à la détection de la déviance.

Deux études ont montré que la déviance occasionnelle d'un stimulus bimodal sur ses deux dimensions auditive et visuelle simultanément, était détectée plus rapidement qu'une déviance uniquement sur sa dimension auditive ou sur sa dimension visuelle (Squires et coll., 1977 ; Teder-Sälejärvi et coll., 2002). Cependant ce résultat pourrait s'expliquer, tout comme l'effet du stimulus redondant, par un phénomène de facilitation statistique dans un modèle d'activations séparées : si le temps de détection du premier processus de détection auditif ou visuel arrivé à son terme détermine le temps de détection d'un essai donné, alors le temps de détection de deux déviances simultanées sera en moyenne inférieur au temps

de détection d'une seule déviance sans que l'on n'ait besoin de postuler d'interactions entre les processus auditifs et visuels de détection de déviance (voir la partie 7.1 page 99).

Pour exclure cette possibilité, il faut tester l'inégalité de Miller sur la distribution des temps de détection des déviances auditive, visuelle et audiovisuelle d'un évènement audiovisuel standard. Si cette inégalité est falsifiée et les modèles d'activations séparées rejetés, alors on pourra supposer que les processus auditif et visuel de détection de la déviance ont interagi. Dans la mesure où le processus de détection de la déviance auditive est lié à la comparaison du déviant auditif avec la représentation présente en mémoire sensorielle auditive, ce résultat serait compatible avec la mise en jeu de la dimension visuelle dans cette comparaison.

Mais d'autres explications (non exclusives) sont possibles puisque la diminution du TR pourrait aussi bien refléter une influence des informations auditives dans le processus analogue de détection de la déviance visuelle. Par ailleurs les interactions audiovisuelles pourraient concerner des étapes de traitement en aval de la comparaison à la trace mnésique, comme, par exemple, ceux impliqués dans la détection consciente de la déviance ou dans la réponse motrice.

L'inégalité de Miller a été testée par Schröger et Widmann (1998) dans une telle tâche de détection de stimuli audiovisuels déviants sur leur localisation spatiale, soit dans la dimension visuelle, soit dans la dimension auditive, soit dans les deux dimensions : les temps de détection des déviants audiovisuels étaient significativement plus rapides que ceux prédits par les modèles d'activations séparées. Il semble donc que les processus auditifs et visuels de détection de la déviance interagissent. Nous avons voulu tester l'inégalité de Miller avec d'autres types de stimuli standards et déviants. Ces stimuli sont ceux qui seront utilisés dans les expériences suivantes. Une violation de l'inégalité de Miller permettrait d'établir que les stimuli utilisés sont susceptibles de donner lieu à des interactions audiovisuelles au niveau de la mémoire sensorielle auditive.

13.2 Méthodes

13.2.1 Sujets

Quinze sujets droitiers (dont 8 de sexe féminin) âgés en moyenne de 23,1 ans ont passé cette expérience. Aucun sujet ne souffrait de troubles neurologiques. Ils avaient tous une audition normale et une vision normale ou corrigée.

13.2.2 Stimuli

Les stimuli utilisés étaient inspirés de ceux utilisés par notre équipe dans des expériences précédentes et qui ont permis de mettre en évidence des interactions audiovisuelles précoces (revue dans Fort & Giard, 2004, et dans la partiediscrimination PE). Nous avons utilisé 4 types de stimuli audiovisuels A_1V_1, A_1V_2, A_2V_1 et A_2V_2, représentés dans la figure 13.1 page suivante.

Les composantes visuelles de ces stimuli consistaient en une déformation horizontale (V_1) ou verticale (V_2) transitoire d'un cercle jaune sur fond noir, ayant un diamètre de 2°

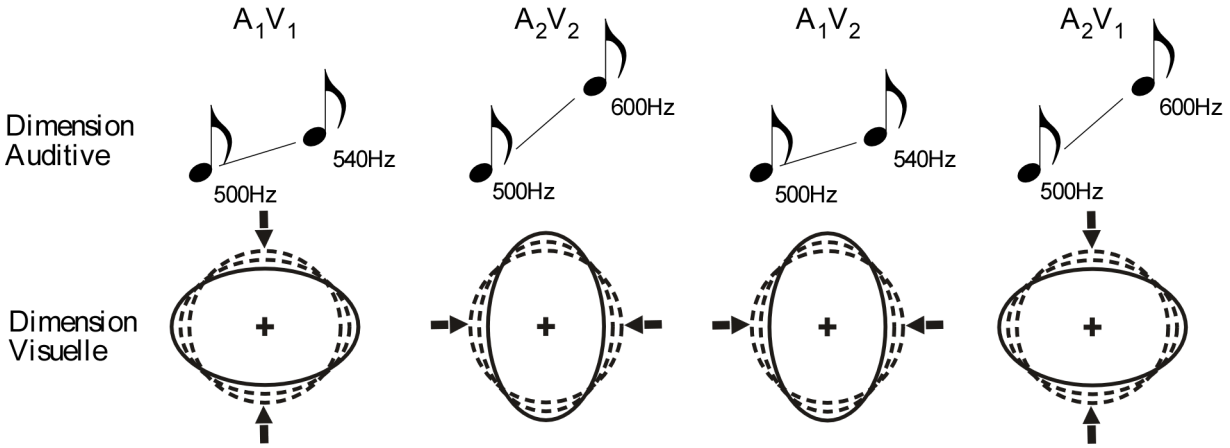

FIG. 13.1 – Stimuli audiovisuels utilisés dans les expériences comportementales et électrophysiologiques sur la mémoire sensorielle. Chaque stimulus était constitué d'une composante auditive A_1 ou A_2 et d'une composante visuelle V_1 ou V_2.

d'angle visuel. La déformation avait une durée totale de 140 ms incluant le retour du cercle à son état initial. L'amplitude de la déformation du cercle à son maximum représentait 33% du diamètre du cercle de départ.

Les composantes auditives des stimuli consistaient en un son pur enrichi des deux premières harmoniques paires dont la fréquence fondamentale variait linéairement soit de 500Hz à 540Hz (A_1), soit de 500 Hz à 600 Hz (A_2) sur une durée de 140 ms (montée/descente : 14 ms).

La taille des déviances auditives et visuelles a été choisie de façon à ce que, sur un groupe de sujets, le TR pour discriminer le stimulus A_1 du stimulus A_2 soit équivalent au TR pour discriminer le stimulus V_1 du stimulus V_2. Nous avons choisi d'équilibrer la discriminabilité des composantes auditives et visuelles car plusieurs études ont monté que la diminution du TR en condition audiovisuelle est maximale dans ces conditions (par exemple Squires et coll., 1977).

Dans la moitié des blocs expérimentaux, le stimulus A_1V_1 était présenté avec une probabilité de 76% (standard) et les stimuli A_1V_2, A_2V_1 et A_2V_2 (respectivement déviants visuel, auditif et audiovisuel) étaient présentés avec une probabilité de 8% chacun. Dans l'autre moitié des blocs, le stimulus A_2V_2 était statndard et les stimuli A_2V_1, A_1V_2 et A_1V_1 étaient déviants. Les stimuli audiovisuels standards seront désormais notés AV, et les déviants auditifs, visuels et audiovisuels A'V, AV' et A'V' respectivement.

13.2.3 Procédure

Un bloc de stimuli commençait par la présentation du cercle sur l'écran, qui restait présent pendant toute la durée d'un bloc.
Les stimuli standards et déviants étaient présentés de façon pseudo-aléatoire avec, pour contrainte, qu'un bloc commençait nécessairement par au moins 3 stimuli standards et que deux stimuli déviants étaient séparés par au moins 1 stimulus standard.
La tâche du sujet consistait à fixer la croix de fixation (centre du cercle) et à cliquer le

plus vite possible avec l'index sur le bouton gauche de la souris à chaque apparition d'un stimulus déviant, que la composante déviante soit auditive, visuelle ou audiovisuelle.
Un total de 1000 stimulations (dont 80 déviants de chaque type) a été présenté en 4 blocs expérimentaux d'une durée approximative de 2 minutes 20 secondes chacun. L'intervalle interstimulus était de 560 ms.
Les blocs ayant pour standard les stimuli A1V1 et A2V2 était présentés dans un ordre aléatoire et différent pour chacun des sujets.

13.2.4 Analyses

Seuls les TR supérieurs à 150 ms et inférieurs à 1500 ms étaient pris en compte, les autres étant considérés comme des fausses alarmes. Les temps de détection ont été analysés conformément aux méthodes exposées dans la partie 9.2.5 page 122.

13.3 Résultats

Les temps moyens pour détecter les déviants auditifs, visuels et audiovisuels étaient respectivement 446, 429 et 356 ms (écarts-types : 52, 51 et 38 ms). Les taux d'erreurs (cibles manquées) étaient de 6,67%, 23,30%, et 3,33 % dans les conditions auditive, visuelle et audiovisuelle respectivement (écarts-types : 4,67%, 15,01% et 3,36%). La figure 13.2 montre les fonctions de répartition des temps de détection pour les déviants auditifs, visuels et audiovisuels ainsi que la somme des fonctions de répartition auditive et visuelle.

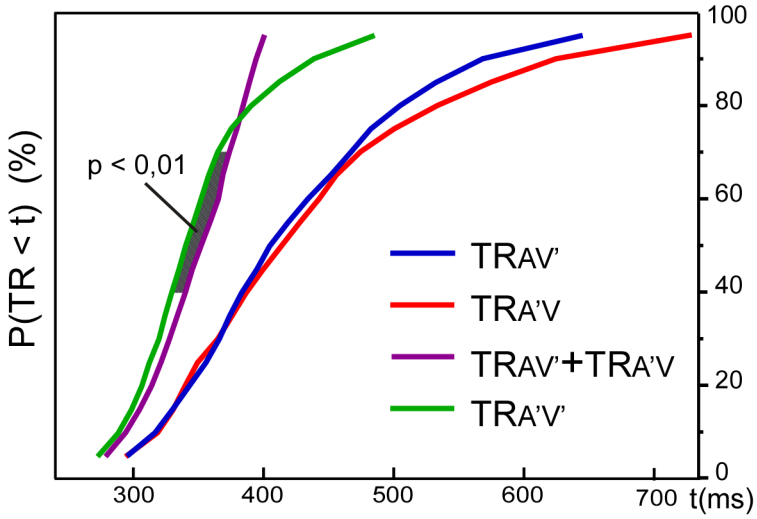

FIG. 13.2 – Application de l'inégalité de Miller. $TR_{AV'}$: fonction de répartition des temps de détection du déviant visuel ; $TR_{A'V}$: fonction de répartition des temps de détection du déviant auditif ; $TR_{AV'}+TR_{A'V}$: somme des 2 fonctions de répartition unimodales ; $TR_{A'V'}$: fonction de répartition des temps de de détection du déviant audiovisuel. La partie hachurée désigne les zones où les fractiles correspondants des deux fonctions de répartition sont significativement différents.

Comme on peut le constater, les temps de détection des déviants audiovisuels étaient plus courts que ceux prédits par les modèles d'activations séparées. Cette différence était

significative au seuil $p < 0,05$ pour les fractiles $t(0,2)$ à $t(0,7)$ et significative au seuil $p < 0,01$ pour les fractiles $t(0,4)$ à $t(0,7)$.

13.4 Discussion

Commençons par noter que le taux d'erreur était plus élevé pour les déviants visuels que pour les déviants auditifs ou audiovisuels. Il est probable que les sujets ont sacrifié l'exactitude pour garder des TR équivalents pour les déviants auditif et visuel. Nos résultats confirment ceux de Schröger et Widmann (1998) : la déviance audiovisuelle d'un évènement audiovisuel est détectée plus rapidement qu'une déviance sur une des deux dimensions sensorielles seulement. Le rejet des modèles d'activations séparées a, depuis lors, été répliqué par Teder-Sälejärvi, Di Russo, McDonald et Hillyard (2005) pour des déviances auditives et visuelles sur l'intensité (sonie et brillance), sous l'hypothèse d'indépendance des distributions de TR unimodales, puis par Gondan et coll. (2005) pour des déviants qui consistaient en une répétition du stimulus standard, dans le cas général des modèles d'activations séparées. Ces résultats et les nôtres suggèrent une coactivation entre les processus auditif et visuel de détection de la déviance. Comme nous l'avons souligné dans l'introduction, cela ne garantit pas l'existence d'une dimension visuelle de la représentation de l'évènement audiovisuel en mémoire sensorielle auditive, mais cela montre l'existence d'interactions audiovisuelles dans un processus mettant vraisemblablement en jeu cette mémoire sensorielle.

Dans l'étude de Schröger et Widmann (1998), les potentiels évoqués par les trois déviants et par le stimulus standard avaient été enregistrés. Lorsque l'on calcule la différence entre déviants et standards auditifs dans ce type de protocole, où le sujet a pour tâche de détecter les déviants, on observe en plus de la MMN, des ondes plus tardives telles que la N2b et la P3, vraisemblablement associées au traitement conscient de la déviance. En appliquant le modèle additif, ils ont pu établir que des interactions audiovisuelles prenaient place à partir de 180 ms au niveau de l'onde N2b et de l'onde P3, mais pas au niveau de la MMN. Leur conclusion était donc que ce qui expliquait le gain de temps de réaction était attribuable à une co-activation au niveau des processus conscients de détection de la déviance plutôt qu'à la comparaison automatique des traces en mémoire sensorielle.

Toutefois, la MMN et la N2b sont deux ondes qui se recouvrent partiellement, et il est possible que des interactions audiovisuelles prennent place vers la fin du processus indexé par la MMN et soient superposées à des interactions au niveau de l'onde N2b. Dans l'expérience suivante, nous allons donc appliquer le modèle additif aux différences entre déviants et standards dans une situation où le sujet ignore les stimuli et où les processus étudiés (indexés par la MMN) sont automatiques.

Chapitre 14

Additivité des MMNs auditives et visuelles

Cette étude ayant fait l'objet d'une publication (Besle, Fort & Giard, 2005), elle ne sera que brièvement présentée ici. Les détails en sont décrits dans l'article, intégrée au manuscrit en annexe (page 257).

14.1 Introduction

Pour les raisons exposées dans l'introduction générale de cette partie, nous pensons que la trace en mémoire sensorielle est susceptible d'incorporer des régularités visuelles lorsqu'elles sont associées à des régularités auditives puisque des interactions audiovisuelles ont probablement lieu avant les processus responsables de la construction de la trace (Giard & Peronnet, 1999). Une incorporation de la sorte devrait nécessairement se traduire par une différence entre les MMN générées par une déviance auditive et une double déviance auditive et visuelle d'un évènement audiovisuel. En effet plusieurs expériences suggèrent que les MMN générées par des déviants différant d'un son standard sur différents traits acoustiques sont générées dans différentes parties du cortex auditif (Giard et coll., 1995 ; Rosburg, 2003). Par ailleurs, l'amplitude de la MMN générée par une déviance sur le même trait acoustique augmente avec l'amplitude de la déviance (Novitski et coll., 2004 ; Tiitinen et coll., 1994). Si les caractéristiques visuelles d'un évènement audiovisuel sont intégrées à la trace en mémoire sensorielle auditive, alors des déviances différentes (auditives et audiovisuelles) d'un évènement audiovisuel devraient générer des MMN différentes.

Cependant, on ne peut se contenter de comparer la MMN générée par un déviant auditif et un déviant audiovisuel car on doit tenir compte de l'éventuelle existence de processus de détection automatique de la déviance visuelle. Des études récentes ont mis en évidence une onde analogue à la MMN dans la modalité visuelle (revue dans Pazo-Alvarez et coll., 2003) et ont montré que cette MMN visuelle possède certaines des caractéristiques d'un marqueur des processus de comparaison automatique à une trace mnésique (en mémoire sensorielle visuelle) : indépendance à l'attention (Heslenfeld, 2003), exclusion de l'hypothèse de *refractoriness* (Czigler, Balazs & Winkler, 2002 ; Pazo-Alvarez, Amenedo & Cadaveira,

2004, voir cependant Kenemans, Jong & Verbaten, 2003 pour une autre hypothèse non mnésique). La MMN visuelle semble être générée dans les aires occipitales (Berti & Schröger, 2004), mais certaines études ont décrit une composante additionnelle plus antérieure dans la MMN visuelle(Czigler et coll., 2002 ; Heslenfeld, 2003). Pour comparer la MMN auditive à une déviance audiovisuelle à celle générée par une déviance audiovisuelle, il faudra donc corriger pour l'existence éventuelle de la MMN visuelle, ce qui revient à tester l'additivité des ondes générées par des déviances auditives, visuelles et audiovisuelles d'un même standard audiovisuel.

L'étude de Schröger et Widmann (1998) semble indiquer que la violation d'additivité concerne les processus en aval de ceux indexés par la MMN, mais il n'était pas possible de séparer dans cette étude les violations d'additivité dues à la MMN de celles dues à l'onde N2b. Une autre étude plus ancienne avait testé l'additivité des MMN auditives et visuelles (Nyman et coll., 1990) et n'était pas parvenue à mettre en évidence des processus visuels de détection automatique de la déviance. Les auteurs avaient donc conclu à la spécificité auditive de la MMN auditive. La violation de l'additivité n'avait cependant pas été testée statistiquement dans cette étude.

Par ailleurs, plusieurs études, déjà mentionnées, ont rapporté l'existence d'une MMN auditive évoquée par une déviance visuelle d'un évènement audiovisuel, par exemple dans le cas de l'illusion McGurk (Colin et coll., 2004 ; Colin, Radeau, Soquet, Demolin et coll., 2002 ; Möttönen et coll., 2002 ; Sams et coll., 1991), de l'illusion de ventriloquie (Colin, Radeau, Soquet, Dachy & Deltenre, 2002 ; Stekelenburg et coll., 2004), ainsi que dans le cas d'un biais visuel dans la perception d'émotions portées par une voix (de Gelder, Bocker, Tuomainen, Hensen & Vroomen, 1999). Dans tous les cas, sauf le dernier, l'existence d'une illusion irrépressible préservait la possibilité que les informations visuelles aient été converties sous forme auditive et que le processus de comparaison des traces aboutissant à la MMN ait été indépendant de toute interaction audiovisuelle. Quoiqu'il en soit, aucune de ces études n'a envisagé que la MMN enregistrée dans ces conditions, c'est-à-dire la différence entre les réponses à l'évènement audiovisuel standard et à l'évènement audiovisuel déviant sur sa composante visuelle, pouvait refléter en réalité un processus visuel de comparaison du stimulus à une trace (indexant une supposée mémoire sensorielle visuelle) ou tout autre processus visuel automatique dû à la présence d'une déviance visuelle.

Tester l'additivité des MMN auditive et visuelle peut donc permettre de répondre à plusieurs questions : Observe-t-on une MMN à une déviance visuelle d'un évènement audiovisuel ? Et si oui, cette MMN reflète-t-elle une influence visuelle sur un processus auditif, même en l'absence d'une illusion audiovisuelle, ou un processus visuel automatique de détection de la déviance (la MMN visuelle) ? Observe-t-on une modulation de la MMN auditive par la présence d'une déviance visuelle qui pourrait refléter le fait que le processus auditif de comparaison du déviant à une trace en mémoire sensorielle auditive est influencé par les informations visuelles ? Pour répondre à ces différentes questions, il est important d'étudier la topographie de la violation de l'additivité, le cas échéant : une topographie auditive peut signifier soit que la déviance visuelle d'un évènement audiovisuel provoque une MMN auditive, comme dans le cas des illusions (McGurk ou ventriloquie), soit que la

MMN auditive évoquée par le déviant audiovisuel a été influencée par la présence d'informations visuelles. En revanche, une topographie visuelle suggèrerait que c'est le processus de détection automatique de la déviance visuelle qui est influencé par les informations auditives.

Dans ce dernier cas cependant une ambigüité peut provenir du fait qu'on connait mal la topographie de la MMN visuelle et de l'existence possible d'une composante antérieure de la MMN visuelle. Afin de résoudre cette ambigüité, le cas échéant, nous avons également enregistré la MMN visuelle évoquée par nos stimulations dans une condition visuelle seule.

14.2 Méthodes

14.2.1 Sujets

Les sujets étaient les mêmes que ceux ayant participé à l'expérience précédente. En réalité l'expérience électrophysiologique a été réalisée avant l'expérience comportementale, le même jour.

14.2.2 Stimuli

Les stimuli étaient identiques à ceux utilisés dans l'expérience comportementale.

14.2.3 Procédure

Puisqu'il s'agissait de mesurer dans cette expérience des processus automatiques, il fallait s'assurer que les sujets portent leur attention ailleurs que sur les évènements audiovisuels. À cette fin, la tâche du sujet était, dans cette expérience, de répondre le plus rapidement possible lorsque la croix de fixation disparaissait. Cette disparition avait une durée de 120 ms et avait une probabilité d'occurrence de 13%. Elle était cependant désynchronisée par rapport aux évènements audiovisuels et ne pouvait se produire que pendant un essai standard (pour éviter de rejeter trop d'essais déviants dans le calcul des potentiels évoqués), et n'avait jamais lieu dans un essai précédent un déviant (pour éviter que les potentiels évoqués par les déviants ne soient contaminés par des processus liés à la réponse, étant donné l'intervalle inter-stimulus relativement faible). Ainsi, le sujet devait regarder l'écran sur lequel étaient présentés les évènements audiovisuels, avec son attention dirigée vers une autre tâche. Il avait, de plus, pour consigne d'ignorer le cercle et les sons.

Pour le test de l'additivité des MMNs, un total de 3200 stimulations ont été présentées (dont 8%, c'est-à-dire 256 déviants, de chaque type). Les stimuli étaient mélangés aléatoirement avec des contraintes identiques à l'expérience comportementale et répartis en 12 blocs d'un durée approximative de 2 minutes 30. Dans la moitié de ces blocs, le stimulus A_1V_1 était le standard, dans la deuxième moitié, c'était le stimulus A_2V_2.

Pour la condition visuelle seule, les séquences de stimuli étaient du même type que dans les conditions audiovisuelles, excepté qu'aucun son n'était présenté : la probabilité d'occurrence d'un déviant visuel (V') était donc de 16% et celle d'un stimulus standard

(V) de 84%. Un total de 1600 stimulations (dont 512 déviants) a été présenté, réparties en 6 blocs. Dans la moitié de ces blocs, le stimulus V1 était le standard.

Les blocs audiovisuels et visuels seuls étaient présentés dans un ordre aléatoire, différent d'un sujet à un autre.

14.2.4 Analyses

Pour le calcul des PE standards moyens, tous les essais ayant inclu une cible, ainsi que les essais suivant immédiatement un déviant ont été exclus. Après rejet des artéfacts d'enregistrement, le nombre moyen d'essais par sujet pour le calcul des PE étaient de 1299, 649 et 204 respectivement pour les standards audiovisuels, les standards visuels et chacun des 4 types de déviants (auditif, visuel, audiovisuel et visuel seul). La ligne de base était prise entre 100 ms et 0 ms avant la stimulation.

Pour le paradigme audiovisuel, les MMN auditive ($MMN_{A'V}$), visuelle ($MMN_{AV'}$) et audiovisuelle ($MMN_{A'V'}$) ont été calculées respectivement comme la différence, point par point, entre les potentiels évoqués par les déviants A'V, AV' et A'V' et le potentiel évoqué par le standard AV. Chaque PE déviant ou standard était donc lui-même une moyenne des potentiels évoqués par deux stimuli différents dans un rôle particulier (le potentiel évoqué déviant audiovisuel était par exemple la moyenne du potentiel évoqué par le stimulus A_1V_1 dans son rôle de déviant et la moyenne du potentiel évoqué par le stimulus A_2V_2 dans son rôle de déviant).

Dans le paradigme visuel seul, la MMN visuelle ($MMN_{V'}$) a été calculée comme la différence entre les potentiels évoqués par le déviant V' et le standard V.

Tous les tests statistiques étaient des tests de Student appariés. Pour éviter le problème des tests multiples, nous n'avons effectué chaque test qu'à un échantillon correspondant au pic maximum de la MMN concernée, sur une valeur moyennée sur une fenêtre de 40 ms autour de la latence de ce pic.

Pour le test de la violation du modèle additif, nous avons arbitrairement choisi la latence du maximum d'amplitude de la $MMN_{A'V}$ car l'objectif premier était de montrer une modulation visuelle de la MMN auditive.

14.3 Résultats

Les TR pour la tâche distractive dans les blocs audiovisuels et visuels étaient de 404 et 409 ms respectivement (écart-types : 51 et 52 ms). Les taux de cibles manquées étaient respectivement de 3,51 et 3,24% (écarts-types : 3,13 et 3,11%). Aucune des deux mesures n'était significativement différente entre les deux conditions.

Les MMN A'V, AV' et A'V' du paradigme audiovisuel sont illustrées dans les figures 14.1 page suivante et 14.2 page 190.

La $MMN_{A'V}$ (courbe rouge sur la figure 14.1) a son pic vers 198 ms et présente la topographie fronto-centrale habituelle (figure 14.2.A) avec inversion de polarité aux mastoïdes, typique des activités générées dans le cortex auditif. Les tests statistiques à la latence du maximum sont très significatifs sur l'ensemble du scalp.

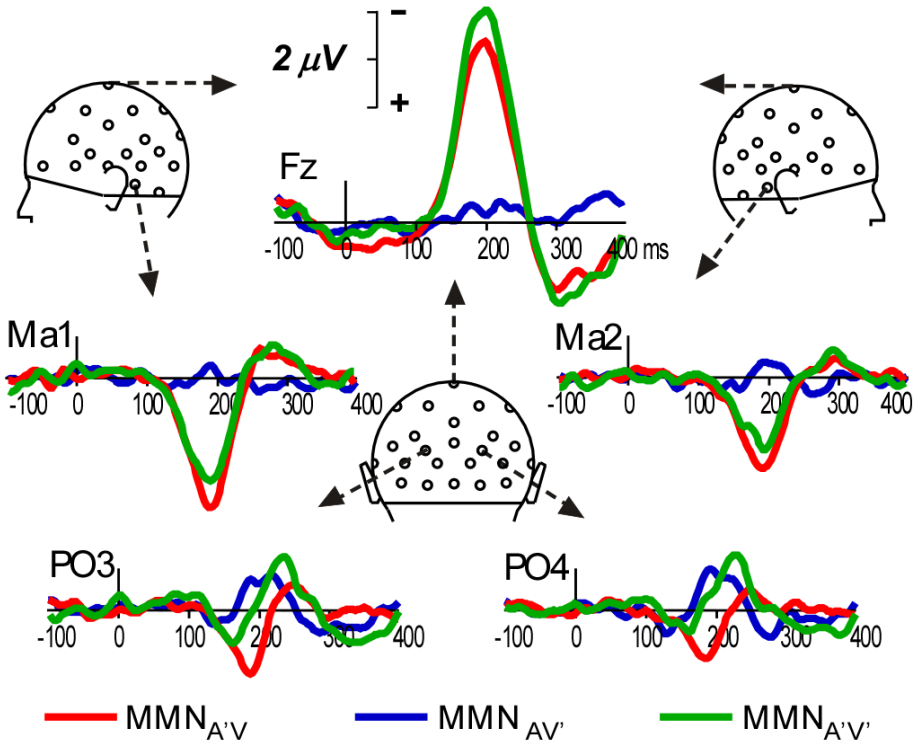

FIG. 14.1 – MMN générées par les déviants A'V, AV' et A'V' dans le paradigme audiovisuel, sur un sous-ensemble d'électrodes.

La $\mathrm{MMN}_{\mathrm{AV'}}$ présente une topographie bilatérale occipitale (figure 14.2.B), avec deux pics, l'un vers 192 ms et le suivant vers 215 ms (courbe bleue sur la figure 14.1). Sa topographie postérieure suggère qu'elle est générée dans le cortex visuel. On n'a pas observé d'activité plus antérieure ou typique d'activations du cortex auditif. Les tests de Student menés à la latence du premier pic indiquent des potentiels significativement différents de 0 sur un grand nombre d'électrodes occipitales.

La $\mathrm{MMN}_{\mathrm{A'V'}}$ (courbe verte sur la figure 14.1) ressemble fort à la $\mathrm{MMN}_{\mathrm{A'V}}$, avec un pic d'amplitude à la même latence (199 ms). Si on regarde cependant plus attentivement les électrodes occipitales PO3 et PO4, on constate qu'elle se rapproche de la $\mathrm{MMN}_{\mathrm{AV'}}$. Au niveau de la topographie des potentiels (figure 14.2.C), il est très difficile de la distinguer de celle de la $\mathrm{MMN}_{\mathrm{A'V}}$. Mais la topographie des densités radiales de courant permet de distinguer clairement des générateurs temporaux, identiques à ceux de la MMN auditive, et des générateurs occipitaux.

Nous avons comparé l'amplitude de la $\mathrm{MMN}_{\mathrm{A'V'}}$ à la somme des amplitudes des $\mathrm{MMN}_{\mathrm{A'V}}$ et $\mathrm{MMN}_{\mathrm{AV'}}$, à la latence du pic de la MMN auditive : l'additivité est significativement violée sur 12 électrodes situées pour la plupart sur l'hémiscalp gauche (figure 14.3.A page 191). La topographie de la violation du modèle additif est centrée sur une zone pariéto-occipitale gauche et ne ressemble ni à la topographie auditive, ni à la topographie visuelle.

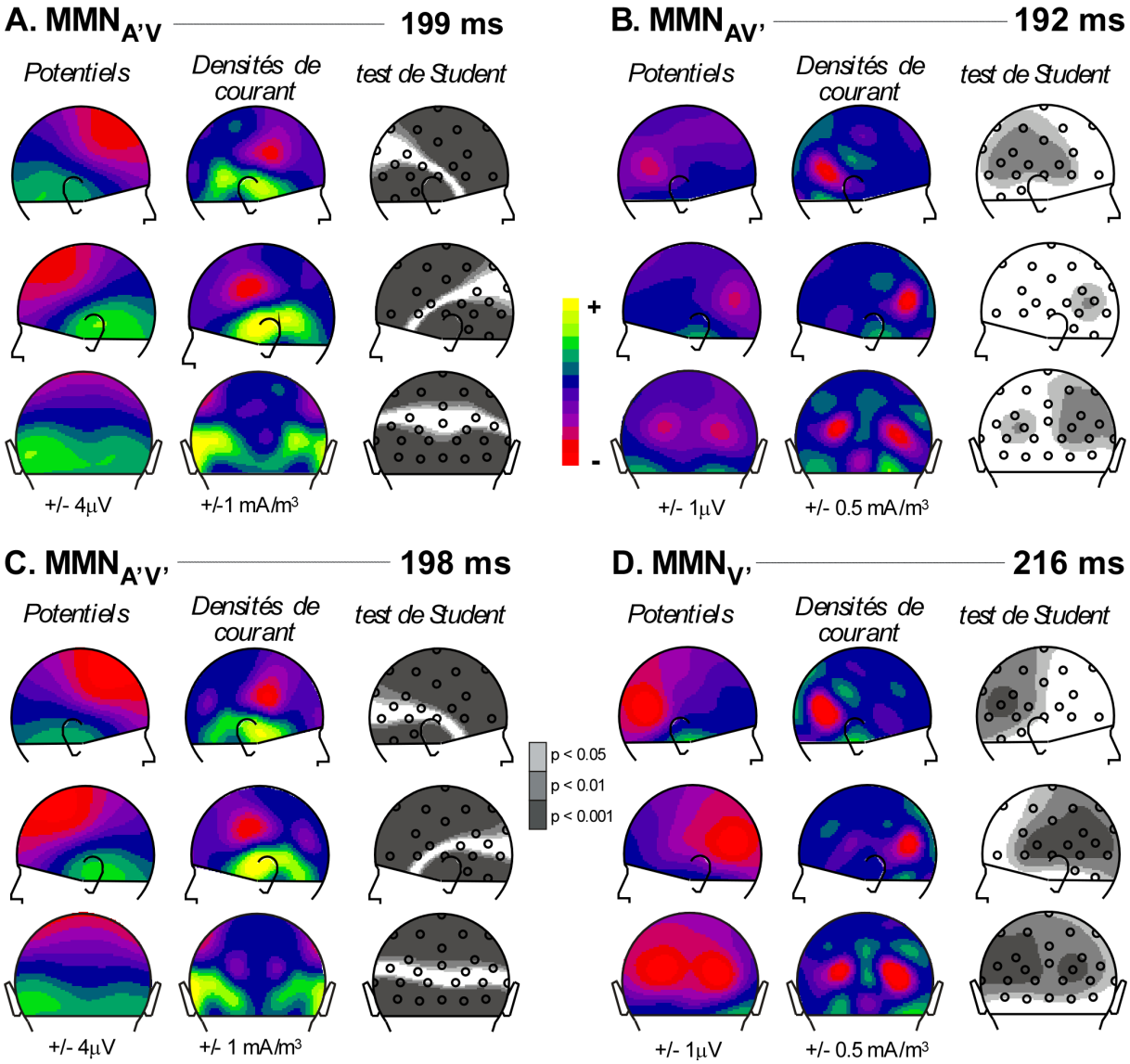

FIG. 14.2 – Topographies des MMN générées par les déviants A'V (A), AV' (B) et A'V' (C) dans le paradigme audiovisuel, et par le déviant V' dans le paradigme visuel seul (D), à la latence de leurs pics d'amplitude respectifs. Le maximum de l'échelle de couleur est indiqué sous chaque ensemble de cartes. Les cartes des tests de Student en niveaux de gris indiquent la significativité des amplitudes par rapport à la ligne de base.

La MMN visuelle générée en contexte unimodal ($MMN_{V'}$) est illustrée dans la figure 14.4 page 192 et comparée à la MMN visuelle générée en contexte bimodal ($MMN_{AV'}$). Les deux MMNs sont très ressemblantes, comme on peut le constater également sur la topographies des potentiels et des densités radiales de courant (figures 14.2.B et 14.2.D de la présente page). La $MMN_{V'}$ semble cependant ne posséder qu'un pic d'amplitude vers 216 ms. La différence d'amplitude entre les MMN visuelles générées dans les deux contextes à cette latence est significative sur 8 électrodes (figure 14.3.B page ci-contre). La topographie de la différence est difficile à interpréter mais suggère que cette différence d'amplitude n'est

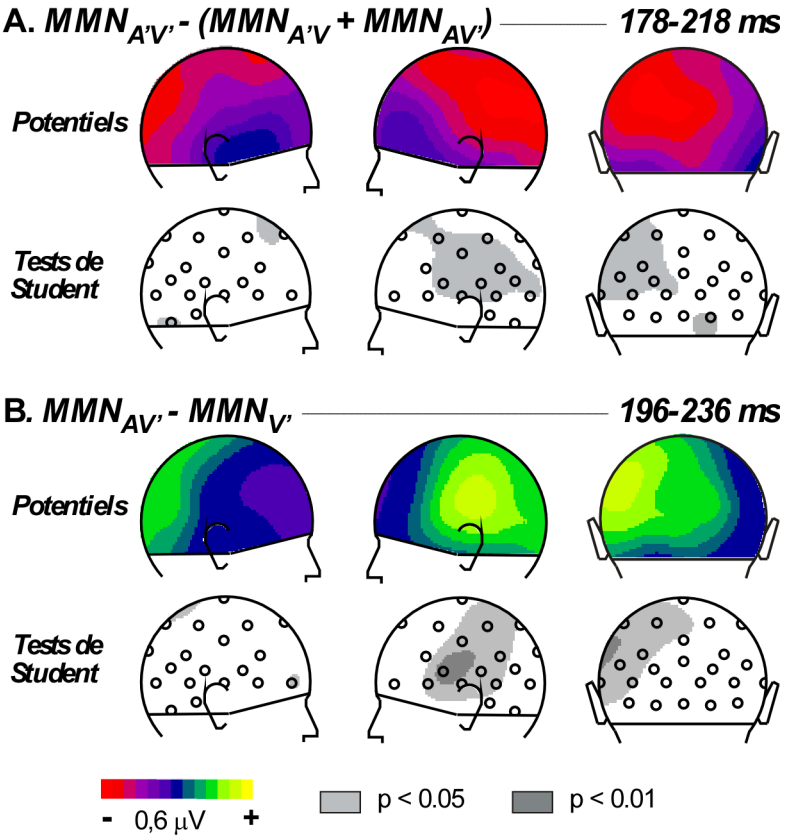

FIG. 14.3 – A. Topographie de la violation du modèle additif [$\mathrm{MMN_{A'V'}} - (\mathrm{MMN_{A'V}} + \mathrm{MMN_{AV'}})$] à la latence du maximum des MMNs auditives (198 ms), dans le paradigme audiovisuel. B. Topographie de la différence entre les MMN visuelles en contexte unimodal ($\mathrm{MMN_{V'}}$) et bimodal ($\mathrm{MMN_{AV'}}$) à la latence du second pic de la $\mathrm{MMN_{AV'}}$. L'échelle est commune à toutes les cartes de potentiels. Les cartes de Student indiquent la significativité des différences.

pas due à une modulation d'amplitude des générateurs de la MMN visuelle.

14.4 Discussion

Une MMN générée par une déviation audiovisuelle d'un évènement audiovisuel présente donc les deux caractéristiques suivantes : elle est composée d'un générateur supra-temporal et d'un générateur occipital, ce qui indique qu'elle met en jeu à la fois les aires sensorielles auditives et les aires sensorielles visuelles ; mais elle n'est pas strictement égale à la somme des MMN générées d'une part par une déviance auditive et d'autre part par une déviance visuelle du même évènement audiovisuel. Les processus indexés par les MMN visuelle et auditive semblent donc n'être pas totalement indépendants. Contrairement aux conclusions de Schröger et Widmann (1998), la coactivation qui facilite le temps de détection des déviants audiovisuels semble commencer dès l'étape de détection automatique de la déviance, qui repose sur l'existence d'une représentation des sons (et des images) standards en mémoire sensorielle.

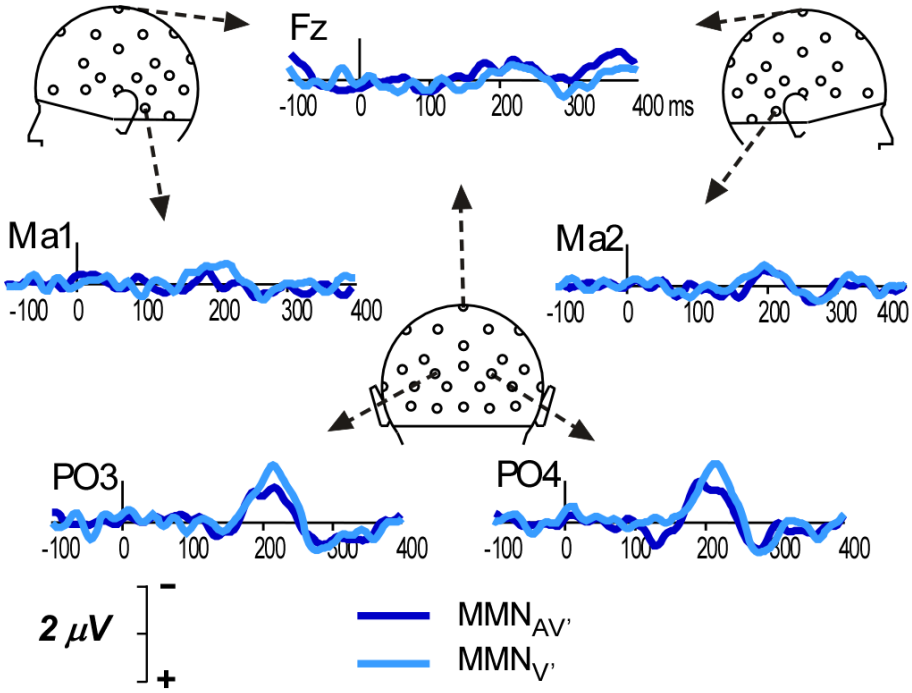

FIG. 14.4 – MMNs visuelles générées en contexte unimodal (MMN$_{V'}$) et bimodal (MMN$_{AV'}$), à un sous-ensemble d'électrodes.

La violation de l'additivité semble corroborer notre hypothèse de l'existence d'une trace audiovisuelle en mémoire sensorielle auditive. Il existe cependant d'autres explications de la violation de l'additivité à considérer.

D'abord, la non-additivité pourrait avoir été provoquée par la présence d'une MMN auditive générée par la déviance visuelle, comme dans le cas de l'illusion McGurk (Colin et coll., 2004 ; Colin, Radeau, Soquet, Demolin et coll., 2002 ; Möttönen et coll., 2002 ; Sams et coll., 1991) et de la ventriloquie (Colin, Radeau, Soquet, Dachy & Deltenre, 2002 ; Stekelenburg et coll., 2004). Cette explication semble cependant ici peu plausible car il est peu probable que la présentation d'un déviant visuel dans notre protocole ait pu modifier la perception auditive du standard auditif comme dans le cas des illusions. Récemment, Saint-Amour, De Sanctis, Molholm, Ritter et Foxe (2007) ont montré que la MMN aux syllabes McGurk déviant sur leur dimension visuelle provenait effectivement du cortex auditif et non d'une détection de la déviance visuelle.

D'autres études récentes ont toutefois montré que le même phénomène était observable lorsque les stimuli auditifs et visuels entretenaient des liens étroits sans pour autant créer une illusion audiovisuelle. Ainsi des stimuli audiovisuels écologiques tels que l'action d'un marteau sur un clou peut provoquer une activité auditive ressemblant à une MMN lorsque sa dimension visuelle est déviante (Ullsperger, Erdmann, Freude & Dehoff, 2006). Il en est de même pour des association audiovisuelles arbitraires stockées en mémoire à long terme telles que les associations graphème/phonème (Yumoto et coll., 2005) et pour des association apprises pour les besoins de l'expérimentation (associations symboliques arbi-

traires : Widmann, Kujala, Tervaniemi, Kujala & Schröger, 2004, ou physique : Aoyama, Endo, Honda & Takeda, 2006). Notons que dans tous ces études, les informations visuelles étaient disponibles avant le stimulus auditif (dans l'étude de Yumoto et coll., 2005, l'effet n'était plus observé lorsque le délai était trop réduit) si bien qu'il est possible que la MMN auditive ait pu être générée parce que le son présenté violait une attente créée par les informations visuelles. Dans notre étude, au contraire, les informations auditives et visuelles étaient disponibles au même moment, étaient associées de manière arbitraire sans être apprises avant l'expérience.

Quoiqu'il en soit, la violation observée dans notre expérience ne présente pas la topographie typique des activités générées dans le cortex auditif, ce qui rend peu probable cette explication.

D'autres explications de la non additivité semblent plus plausibles. Par exemple, dans la mesure où l'on observait un MMN visuelle d'origine occipitale en réponse à un déviant visuel, il est possible que la trace visuelle indexée par cette MMN ait été modifiée par la présence d'informations auditives, à l'inverse de notre hypothèse de départ. La topographie de la violation du modèle additif ne nous permet pas de conclure en faveur de l'une ou l'autre des hypothèses car elle ne présente ni les caractéristiques d'une activité générée dans le cortex visuel, ni celles d'une activité générée dans le cortex auditif.

La différence inattendue entre les MMN visuelles générées en contexte audiovisuel et en contexte visuel suggère néanmoins que la trace en mémoire sensorielle visuelle (sous l'hypothèse que la MMN visuelle a une origine mnésique, voir Czigler, sous presse, pour une revue) a intégré des informations sur la régularité auditive. En effet, la seule différence entre les deux protocoles était que dans le cas audiovisuel, les stimuli visuels étaient toujours associés à un stimulus auditif, et en particulier, que le standard visuel était associé au standard auditif dans 76% des essais. Étant donné que la déviance qui génère la MMN visuelle était la même dans les deux conditions, et que les traitements associés au stimulus auditif doivent disparaître dans la différence entre les PE standards et déviants, une interprétation tentante est que la trace en mémoire sensorielle visuelle a enregistré l'association régulière des standards auditifs et visuels. Cette explication n'est bien sûr pas incompatible avec notre hypothèe initiale : les deux processus de détection automatique de la déviance pourraient être influencés chacun par les informations de l'autre modalité sensorielle.

Une autre hypothèse à considérer pour expliquer la violation de l'additivité est que les informations auditives et visuelles n'ont interagi que dans le traitement des déviants, sans que les traces auditive et visuelle n'aient elles-mêmes été influencées par les informations de l'autre modalité sensorielle. Nos résultats montrent sans ambigüité que les traitements des déviants auditifs et visuels ont interagi avant 200 ms de traitement. En effet, si tel n'était pas le cas, les MMN auditive et visuelle auraient dû être additives, même si les traces auditives et/ou visuelles intègrent des informations intersensorielles. Par contre la violation de l'additivité pourrait s'expliquer uniquement par une coactivation en aval du processus de comparaison, tout en préservant le caractère modalité-spécifique des traces mnésiques.

Dans le domaine auditif, par exemple, plusieurs études ont montré une violation de

l'additivité des MMN à la déviance simultanée sur deux traits acoustiques (Czigler & Winkler, 1996 ; Winkler, Czigler, Jaramillo, Paavilainen & Näätänen, 1998). Dans les deux cas, la MMN à la double déviance avait une amplitude inférieure à la somme des MMN aux déviances simples, comme si la détection d'une des deux déviances diminuait l'importance de l'autre déviance, suggérant l'existence de processus communs déclenchés par les deux déviances. De la même façon il est possible que la détection d'une déviance dans une modalité ait diminué le traitement de l'autre déviance, provoquant une violation de l'additivité. Cette explication est néanmoins insuffisante au moins pour la MMN visuelle puisqu'on trouvait une différence entre les MMN visuelles dans les contextes audiovisuel et visuel, qui peut difficilement s'expliquer par une différence de traitement des déviances.

En résumé, cette expérience ne nous a pas permis de vérifier sans ambigüité notre hypothèse de départ, à savoir que la représentation de l'évènement en mémoire sensorielle auditive inclut des informations sur la régularité visuelle.

Ajoutons que nos données confirment l'origine occipitale de la MMN visuelle, dont la seule représentation topographique disponible était jusqu'à présent celle de Berti et Schröger (2004) dans une étude ou l'attention des sujets étaient portée sur les stimuli, mais où la dimension de la déviance n'était pas pertinente pour la tâche à réaliser. Nos données ne suggèrent en revanche pas l'existence d'un composante antérieure de la MMN visuelle. Notons que nous n'avons pas contrôlé l'hypothèse de *refractoriness* dans notre expérience et qu'on ne peut donc formellement conclure que notre MMN visuelle est le marqueur d'une mémoire sensorielle visuelle. Plus généralement, nous ne pouvons exclure que la violation de l'additivité des MMN auditive et visuelle, résulte d'un phénomène de *refractoriness*. Cela impliquerait cependant l'existence de populations neuronales sensibles à l'association de stimuli auditifs et visuels particuliers.

Chapitre 15

Représentation d'une régularité audiovisuelle en mémoire sensorielle auditive

15.1 Introduction

L'expérience précédente n'a pas permis de montrer formellement qu'une régularité audiovisuelle est codée en mémoire sensorielle auditive. En revanche, nos données suggèrent que la représentation en mémoire sensorielle visuelle, si elle existe, inclut des informations auditives puisque la MMN visuelle unimodale était différente de celle générée par une déviance visuelle d'une évènement audiovisuel. Pour montrer que la mémoire sensorielle auditive inclut des éléments visuels, il nous faut donc montrer, réciproquement, que la MMN générée par la déviance auditive d'un évènement audiovisuel est différente de celle générée par la même déviance en contexte unimodal auditif.

Pour cela nous allons présenter dans un bloc expérimental unimodale un son standard pouvant dévier occasionnellement sur sa fréquence et dans un autre bloc audiovisuel, les mêmes sons standards et déviants mais associés à un stimulus visuel standard. Ainsi, dans le bloc audiovisuel, les traitements évoqués par les stimuli visuels devraient disparaitre dans le calcul de la MMN. Si une différence subsiste entre les MMNs évoquées dans les deux blocs, elle devrait être due à des différences dans les processus de détection automatique de la déviance auditive. Mais puisque les déviances sont identiques dans les deux blocs, la différence devrait provenir de la différence existant dans la mémoire sensorielle auditive entre la trace d'un évènement standard auditif et la trace d'un évènement standard audiovisuel.

Nous allons comparer la même MMN auditive générée dans deux contextes différents : un contexte audiovisuel et un contexte auditif seul. Il reste donc toujours la possibilité que la simple présence d'informations visuelles module la MMN. Il a été montré par exemple que la MMN auditive peut être influencée par la présence de stimuli visuels émotionnels (Surakka, Tenhunen-Eskelinen, Hietanen & Sams, 1998), par la charge attentionnelle visuelle (Otten, Alain & Picton, 2000 ; Valtonen, May, Makinen & Tiitinen, 2003 ; Zhang, Chen, Yuan, Zhang & He, 2006) ou la direction de l'attention sélective vers la modalité visuelle ou

auditive (Alho, 1992 ; Dittmann-Balcar, Thienel & Schall, 1999 ; Muller-Gass, Stelmack & Campbell, 2006 ; Woods, Alho & Algazi, 1992).

Le fait que la tâche distractive visuelle soit la même dans les blocs auditif et audiovisuel devrait être un contrôle suffisant pour exclure ces effets attentionnels dans la mesure où elle devrait équilibrer l'attention visuelle soutenue de la même façon dans les deux blocs. Cependant, des stimuli visuels distracteurs peuvent provoquer des déplacements involontaires de l'attention visuelle (spatiale) et avoir une influence sur l'amplitude de la MMN (Mathiak, Hertrich, Zvyagintsev, Lutzenberger & Ackermann, 2005). De plus il est difficile de dire si la simple présence d'un stimulus visuel, même hors du focus attentionnel, pourrait influencer de manière non spécifique la MMN auditive, car cela n'a jamais été testé.

Un meilleur contrôle serait donc de montrer que l'effet des informations visuelles sur la MMN auditive a lieu lorsque les stimuli audiovisuels constituent une véritable régularité audiovisuelle, c'est-à-dire lorsque les mêmes évènements auditifs et visuels sont associés de manière régulière, mais pas lorsque l'association audiovisuelle standard varie d'un essai à l'autre. On pourrait ainsi séparer l'effet non spécifique de la présence de stimuli visuels sur la MMN auditive de la construction d'une véritable représentation de l'évènement audiovisuel régulier en mémoire sensorielle auditive. Nous avons donc ajouté une condition de stimulation que nous avons appelé "audiovisuelle équiprobable" dans laquelle des sons standards et déviants identiques à ceux des autres conditions pouvaient être associés de manière équiprobable à quatre stimuli visuels différents.

Notre hypothèse est donc que la MMN générée par une même déviance auditive devrait être différente dans le cas où elle dévie par rapport à une régularité auditive (condition auditive unimodale), une régularité audiovisuelle (condition audiovisuelle) ou une régularité auditive accompagnée d'informations visuelles ne constituant pas une régularité (condition audiovisuelle équiprobable). En particulier, nous prédisons que la MMN auditive dans la condition audiovisuelle devrait se différencier à la fois de la MMN auditive dans la condition unimodale et de celle générée dans la condition audiovisuelle équiprobable, ces deux dernières devant être identiques, si la simple présence d'informations visuelles n'a pas d'effet sur la MMN auditive.

15.2 Méthodes

15.2.1 Sujets

Seize sujets droitiers (dont 9 de sexe féminin) âgés en moyenne de 24 ans (écart-type : 2,5 ans) ont participé à cette expérience. Aucun sujet ne souffrait de troubles neurologiques. Ils avaient tous une audition normale et une vision normale ou corrigée.

15.2.2 Stimuli

Les stimuli utilisés étaient identiques à ceux des deux expériences précédentes excepté dans la condition audiovisuelle équiprobable, où deux types de composantes visuelles supplémentaires ont été ajoutés. Il s'agissait de déformations du cercle dans deux directions

obliques (V_3 et V_4), montrés dans la figure 15.1.

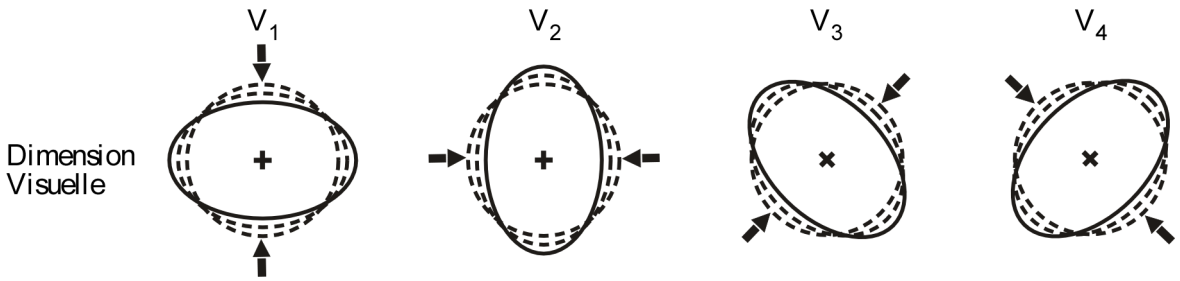

FIG. 15.1 – Composantes visuelles des stimuli audiovisuels utilisées dans la condition audiovisuelle équiprobable

15.2.3 Procédure

Chacune des trois conditions de stimulation comprenait 1600 évènements, dont 320 (20%) étaient des déviants auditifs, répartis en 8 blocs de 200 stimuli d'une durée d'environ 1 minutes 50.

Dans la condition auditive unimodale, le stimulus standard (A) était le stimulus A_1 et le déviant (A') était le stimulus A_2, dans la moitié des blocs. Dans l'autre moitié, les rôles de A_1 et A_2 étaient inversés.

Dans la condition audiovisuelle, les stimuli auditifs standard A et déviant A' d'un même bloc étaient toujours accompagnés du même stimulus visuel V. Dans la moitié des blocs, le stimulus standard (AV) était le stimulus A_1V_1 et le stimulus déviant (A'V) était le stimulus A_2V_1. Dans l'autre moitié, les stimuli standard et déviant étaient les stimuli A_2V_2 et A_1V_2.

Dans la condition audiovisuelle équiprobable, les sons standard et déviant pouvaient indifféremment être associés à l'un des quatre stimuli visuels. Dans la moitié des blocs, il y avait donc 4 standards audiovisuels A_1V_1, A_1V_2, A_1V_3 et A_1V_4, présentés chacun dans 20% des essais et 4 déviants audiovisuels A_2V_1, A_2V_2, A_2V_3 et A_2V_4 présentés chacun dans 5% des essais. Dans l'autre moitié des blocs, les probabilité d'occurrence étaient inversées entre les stimuli constitués du son A1 et ceux constitués du son A2. Les standards et les déviants dans cette condition seront désormais nommés AVeq et AV'eq.

Les 24 blocs de stimulations étaient présentés dans un ordre aléatoire, différent d'un sujet à l'autre. La tâche distractive et la probabilité d'occurrence de la disparition du point de fixation étaient identiques à celles utilisées dans l'expérience précédente.
Les contraintes de succession appliquées aux déviants et aux standards étaient identiques à celles des deux expériences précédentes.

15.2.4 Analyses

Après rejet des artéfacts d'enregistrement et l'exclusion des essais standards ayant contenu une cible ou suivant un déviant, le nombre moyen d'essais par sujet pour le calcul des potentiels évoqués moyens étaient de 741, 755, 826, 279, 284 et 317 respectivement

pour les standards A, AV, AVeq et les déviants A', A'V et A'Veq. La ligne de base était prise entre 100 ms et 0 ms avant la stimulation.

Les MMN auditives $\mathrm{MMN_{A'}}$, $\mathrm{MMN_{A'V}}$ et $\mathrm{MMN_{A'Veq}}$ dans chacune des 3 conditions étaient calculées comme la différence, point par point, entre les potentiels évoqués par chacun des déviants et les potentiels évoqués par chacun des standards.

Pour les tests statistiques, nous avons voulu limiter les hypothèse faites sur la latence des effets, tout en limitant le risque de première espèce global à 5%. Nous avons donc testé la différence entre les MMN à toutes les latences dans une fenêtre 150-250 ms (correspondant à la latence de la MMN auditive dans l'expérience précédente) grâce à un test bilatéral de permutation des conditions appariées (Efron & Tibshirani, 1993, p212 ; $2^{16} = 65536$ permutations). Les tests multiples ont été pris en compte au niveau de chaque électrode par la méthode du minimum d'échantillons significatifs successifs, avec un risque local $\alpha = 0,05$ et un risque global $\alpha_{global} = 0,05$ (voir la partie 8.1 page 112).

15.3 Résultats

Les TR moyens dans la tâche distractive étaient respectivement de 334, 345 et 348 ms dans les conditions auditive, audiovisuelle et audiovisuelle équiprobable (écarts-types : 52, 53 et 50 ms). Les TR dans les trois conditions étaient significativement différents ($p < 0,00007$) et cette différence était due au fait que le TR dans la condition auditive était plus rapide que dans les deux conditions audiovisuelles (auditive contre audiovisuel : $p < 0,003$; auditive contre équiprobable : $p < 0,0001$). Les taux de cibles manquées pour les 3 conditions étaient respectivement de 1,17%, 1,00% et 1, 25% (écarts-types : 1,25%, 1,12% et 1,64%). Ils n'étaient pas significativement différents

Les figures 15.2 page suivante et 15.3 page 200 présentent les potentiels évoqués par les évènements standards et déviants dans les 3 conditions de présentation. Dans la condition auditive (figure 15.2.A), les sons standards et déviants évoquaient une série d'ondes fronto-centrales, caractéristiques du traitement d'un stimulus auditif, visibles notamment sur Cz : une P50 avec un pic d'amplitude vers 60 ms et une inversion de polarité dont le maximum se situe aux mastoïdes, puis une minuscule N100 avec un pic d'amplitude à 100 ms, dont la faible amplitude est probablement due à l'intervalle inter-stimulus relativement rapide. À partir d'environ 120 ms, les potentiels évoqués par les standards et les déviants se séparent et les déviants évoquent une onde d'amplitude importante (la MMN) dont le pic négatif se situe vers 200 ms sur les électrode fronto-centrales et qui présente une inversion de polarité aux mastoïdes

Dans les deux conditions audiovisuelles (figures 15.2.B et 15.2.C), les potentiels évoqués par les standards sont des agrégats complexes de réponses sensorielles auditives et visuelles. Concernant la modalité visuelle, on peut remarquer sur les électrodes occipitales (O1 et O2 sur la figure), superposées aux réponses auditives, d'abord une onde positive avec un pic d'amplitude vers 130 ms et une onde négative dont le pic d'amplitude se trouve vers 170 ms. Ces ondes sont aussi bien évoquées par les standards que par les déviants. Comme dans la condition auditive, les réponses évoquées par les standards et déviants commencent à différer vers 120 ms.

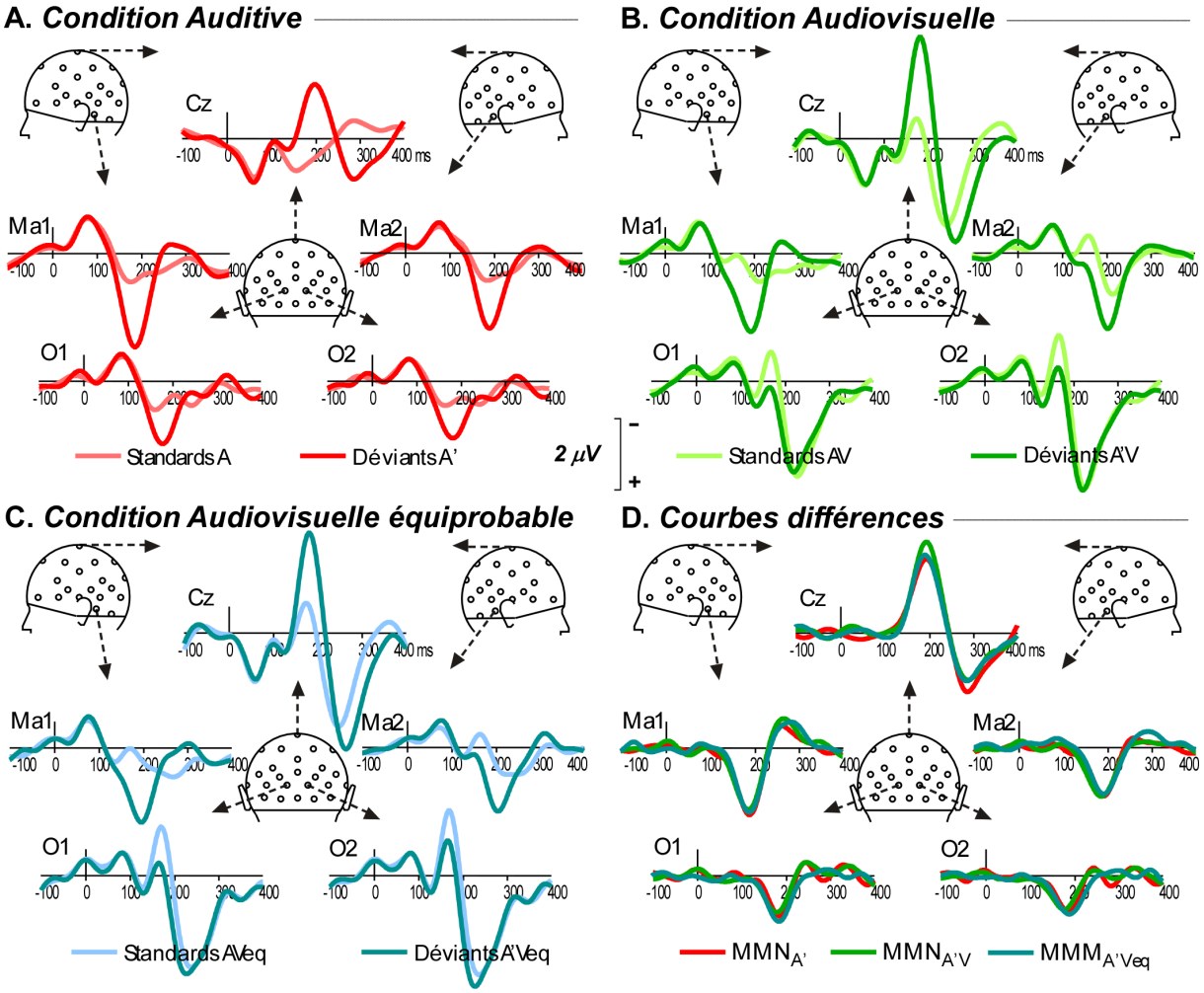

FIG. 15.2 – PE standards et déviants générés dans les conditions auditive unimodale (A), audiovisuelle (B) et audiovisuelles équiprobable (C). D. MMN auditives générées dans les trois conditions.

Lorsque l'on calcule la différence entre les réponses aux déviants et standards, on obtient des courbes très similaires dans les trois conditions (figure 15.2.D), ce qui correspond au fait que la déviance était identique dans ces conditions. La $\mathrm{MMN}_{\mathrm{A'}}$ avait son pic sur l'électrode Fz à 192 ms ($-2,764\mu V$), la $\mathrm{MMN}_{\mathrm{A'V}}$ à 194 ms ($-3,021\mu V$) et $\mathrm{MMN}_{\mathrm{A'Veq}}$ à 192 ms ($-2,714\mu V$).

Comme on peut le constater sur la figure 15.3 page suivante, les topographies des 3 MMN sont très similaires, aussi bien au niveau des potentiels que des densités radiales de courant. L'amplitude du pic négatif de la $\mathrm{MMN}_{\mathrm{A'V}}$ semble cependant plus importante. Les tests de permutation de la différence entre les $\mathrm{MMN}_{\mathrm{A'}}$ et $\mathrm{MMN}_{\mathrm{A'V}}$ (figure 15.4 page suivante) montrent en effet que l'amplitude des deux MMN est significativement différente sur plusieurs électrodes pariéto-centrales entre 180 et 205 ms. Seule la différence sur l'électrode CP1 subsiste lorsque les test multiples sont pris en compte.

Concernant le test de la comparaison entre la condition audiovisuelle et la condition audiovisuelle équiprobable (figure 15.5 page 201), la différence d'amplitude entre les deux

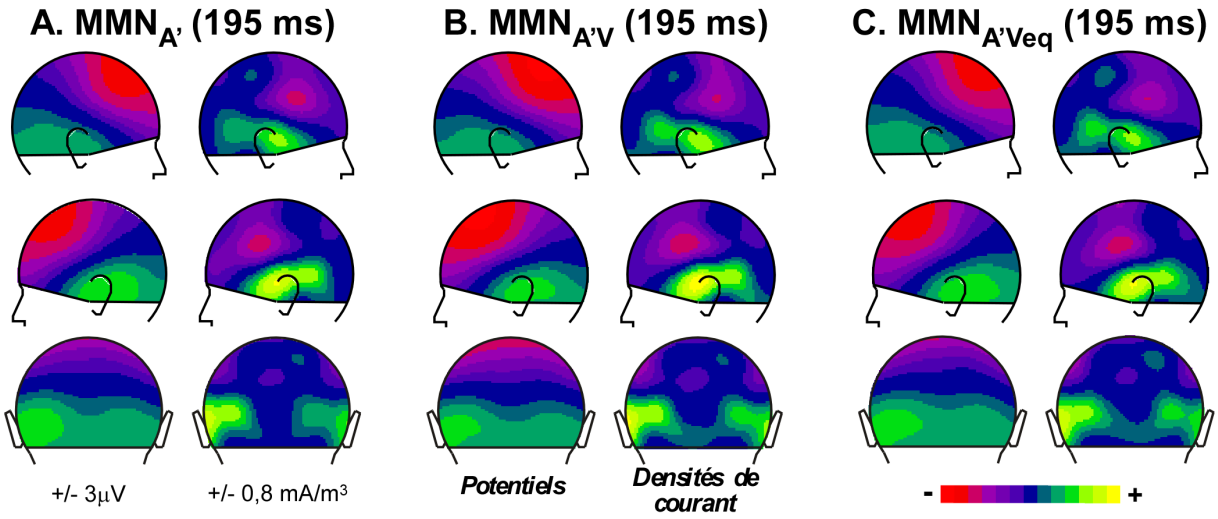

FIG. 15.3 – Topographies des MMN auditives générées dans les conditions auditive unimodale (A), audiovisuelle (B) et audiovisuelles équiprobable (C) à 195 ms.

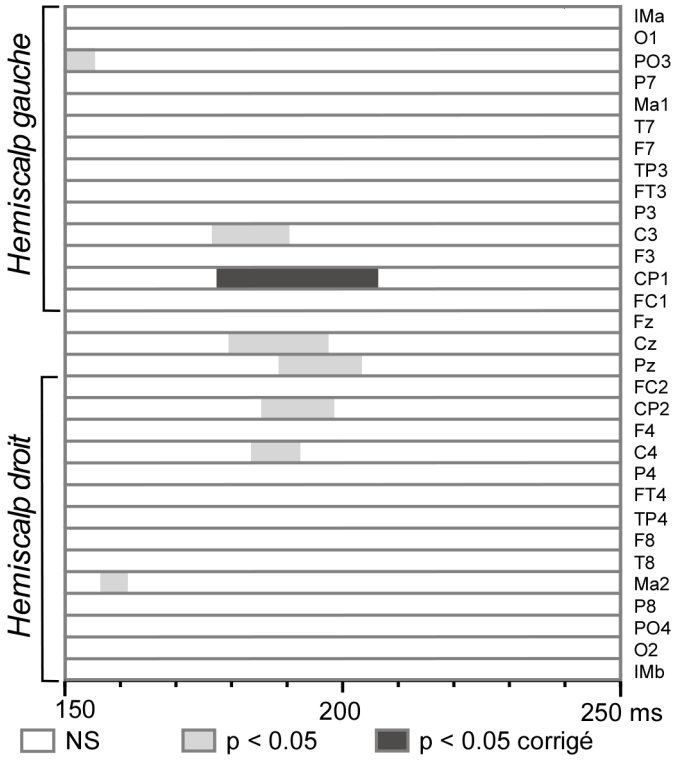

FIG. 15.4 – Résultats des tests de permutation de la différence entre les MMN$_{A'}$ et MMN$_{A'V}$, entre 150 et 250 ms. Le niveau de gris indique la significativité. $p < 0,05$ corrigé : le nombre de tests significatifs successifs dépasse le nombre d'échantillons minimal nécessaire pour limiter le risque global à 0,05.

MMN est significative également, mais seulement sur l'électrode fronto-centrale FC2. De plus, cette différence ne subsiste pas lorsque les test multiples sont pris en compte. La même

comparaison génère d'autres tests significatifs (ne résistant pas plus aux corrections) à une latence plus tardive (entre 215 et 245) ms sur plusieurs électrode pariéto-occipitales à gauche et à droite.

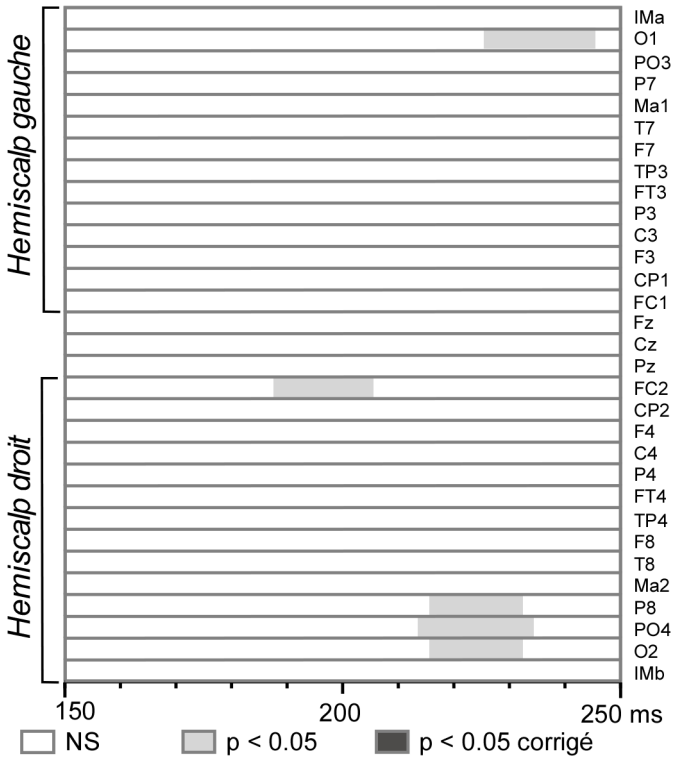

FIG. 15.5 – Résultats des tests de permutation de la différence entre les $MMN_{A'V}$ et $MMN_{A'Veq}$, entre 150 et 250 ms. Le niveau de gris indique la significativité. $p < 0,05$ corrigé : le nombre de tests significatifs successifs dépasse le nombre d'échantillons minimal nécessaire pour limiter le risque global à 0,05.

Le résultat des deux tests statistiques suggère que les MMN différaient à deux latences et en deux zones différentes du scalp. La figure 15.6 page suivante montre la topographie des deux différences testées, au cours du temps. Alors que la topographie de la différence entre les $MMN_{A'}$ et $MMN_{A'V}$ présente un pôle unique commençant sur les électrode centrales et se terminant sur les électrodes frontales, celle de la comparaison des deux conditions audio-visuelles semble être une superposition de la même différence et d'une seconde différence plus tardive et clairement occipitale.

Les densités radiales de courants correspondant à ces différences ne présentaient pas de topographie suffisamment stable (il s'agit d'une différence de différences) pour aider à cette interprétation.

15.4 Discussion

Les résultats vont dans le sens de nos hypothèses puisque la MMN générée par la déviance auditive d'un évènement audiovisuel standard diffère de la MMN unimodale générée par la même déviance auditive. La différence est faible, mais néanmoins significative, même

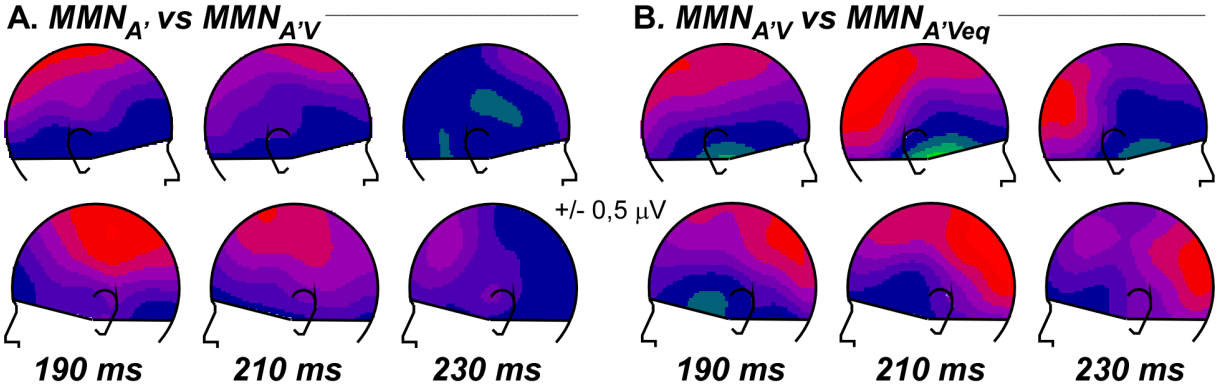

FIG. 15.6 – A. Topographies des différences entre les MMN auditives dans la condition auditive ($MMN_{A'}$) et la condition audiovisuelle ($MMN_{A'V}$), entre 190 et 230 ms. B. Topographies des différences entre les MMN auditives dans la condition audiovisuelle ($MMN_{A'V}$) et la condition audiovisuelle équiprobable ($MMN_{A'Veq}$), entre 190 et 230 ms.

en tenant compte des nombreux tests effectués. De plus, ce résultat ne peut s'expliquer simplement par un effet non spécifique de la présence de stimuli visuels, puisque, dans la condition audiovisuelle équiprobable, lorsque les stimuli auditifs étaient associés avec 4 stimuli visuels différents, il semble que les informations visuelles n'aient pas modifié la MMN auditive par rapport à la condition auditive seule.

Notre interprétation est donc que la représentation d'un évènement audiovisuel en mémoire sensorielle est différente de celle d'un évènement purement auditif, et que l'incorporation de l'élément visuel de la régularité ne peut se faire que si les stimuli auditifs et visuels sont associés de façon consistante au cours des essais. La mémoire sensorielle auditive, telle qu'indexée par la MMN auditive, semble donc stocker des régularités visuelles lorsque celles-ci sont associées à des régularités auditives.

Il semble que la MMN générée dans la condition audiovisuelle équiprobable présente des caractéristiques qui lui sont propres, puisqu'elle présentait une seconde différence par rapport à la condition audiovisuelle, vers la fin de la MMN et sur des électrodes plus occipitales. On peut supposer que la MMN auditive générée dans la condition audiovisuelle équiprobable différait de celle générée dans la condition auditive seule. Il serait hasardeux de s'aventurer à en donner une interprétation, étant donné, d'une part, que nous n'avions aucune hypothèse a priori sur cette différence, et d'autre part, que la significativité de cette différence peut s'expliquer par le nombre de tests effectués.

Il semble que la présence d'informations visuelles ait eu pour effet de ralentir le temps de réaction des sujets dans leur tâche distractive puisque celui-ci était plus rapide d'environ 10 millisecondes dans la condition auditive seule par rapport aux deux conditions audiovisuelles. Les sujets ont donc pu être légèrement distraits par le mouvement du cercle dans leur traitement de la disparition du point de fixation. Mais ils l'étaient tout autant dans la condition audiovisuelle et dans la condition audiovisuelle équiprobable. Cette distraction ne peut donc expliquer ni les effets spécifiques des informations visuelles régulières sur la MMN auditive, qui n'existaient pas dans la condition audiovisuelle équiprobable, ni les effets spécifiques à la condition audiovisuelle équiprobable.

Même si la déviance, c'est-à-dire la différence entre déviants et standards, était la même dans chacune des trois conditions, les déviants utilisés n'étaient pas, à proprement parler, identiques, puisqu'ils étaient purement auditifs dans une condition, audiovisuels dans les deux autres. Il reste donc toujours la possibilité que la différence entre les MMN proviennent simplement de la différence de traitement des déviants lorsqu'ils sont auditifs ou audiovisuels. Cette explication ne dispense pas de l'existence d'interactions entre les traitements auditifs et visuels (en effet, s'il n'y avait aucune interaction, les traitements visuels devraient purement et simplement s'éliminer dans le calcul de la MMN auditive et il n'y aurait aucune différence entre les MMN), mais compromettrait notre interprétation en termes de mémoire sensorielle auditive. Il parait difficilement soutenable, cependant, que la partie visuelle du déviant modifie son traitement dans le cortex auditif, sans qu'il en soit de même pour les standards et que donc la représentation de l'évènement standard en mémoire sensorielle auditive soit affectée par la présence d'informations visuelles.

Il semble donc que l'association régulière d'un stimulus auditif donné, avec un stimulus visuel donné, finisse par générer la perception d'un objet audiovisuel à part entière. La représentation sensorielle de cet objet pourrait être stockée en mémoire sensorielle auditive et en mémoire sensorielle visuelle (si l'on en croit les résultats de la comparaison des $\mathrm{MMN}_{\mathrm{AV}'}$ et $\mathrm{MMN}_{\mathrm{V}'}$ de l'expérience précédente). Le stockage de cette représentation audiovisuelle intégrée dans ces deux mémoires sensorielles pourrait être à l'origine de la facilitation pour la détection d'un déviant audiovisuel, mis en évidence dans la première expérience.

Chapitre 16

MMN à la conjonction audiovisuelle

Cette étude ayant été acceptée pour publication (Besle et coll., sous presse), elle ne sera que brièvement présentée ici. Les détails en sont décrits dans la publication, intégrée au manuscrit , intégrée au manuscrit en annexe (page 267). Cette expérience a été réalisée au centre MEG du CERMEP, à Lyon. Les données ont été acquises par Romaine Mayet, en DEA sous la direction de Dominique Morlet et analysées en collaboration avec Anne Caclin et Dominique Morlet.

16.1 Introduction

Nos expériences précédentes suggèrent qu'une régularité audiovisuelle est représentée en mémoire sensorielle auditive et peut-être en mémoire sensorielle visuelle. Cependant, cela n'a été montré qu'assez indirectement, en étudiant l'influence d'une régularité visuelle sur la représentation d'une régularité auditive et vice-et-versa. Dans l'expérience suivante, nous avons tenté de savoir si la représentation de cette régularité audiovisuelle peut être à l'origine d'une activité de type MMN lorsque la régularité est violée, autrement dit s'il existe une représentation mnésique sensorielle à part entière d'une association particulière et régulière entre un trait auditif et un trait visuel. Pour cela, nous avons présenté des stimuli audiovisuels déviants, ne différant de la régularité audiovisuelle que sur la façon dont les traits auditifs et visuels sont combinés (conjonction de traits), chaque trait auditif ou visuel pris isolément ne constituant pas la violation d'une régularité auditive ou visuelle.

Ces déviants à la conjonction de deux traits ont déjà été utilisés dans des études sur la MMN auditive pour montrer que la mémoire sensorielle auditive ne stocke pas uniquement des représentations indépendantes des traits acoustiques élémentaires, mais également des représentations de leurs combinaisons particulières (Gomes, Bernstein, Ritter, Vaughan & Miller, 1997 ; Sussman, Gomes, Nousak, Ritter & Vaughan, 1998 ; Takegata, Paavilainen, Näätänen & Winkler, 1999 ; Takegata, Huotilainen, Rinne, Näätänen & Winkler, 2001 ; Winkler, Czigler, Sussman, Horvath & Balazs, 2005). Dans ces expériences, plusieurs sons standards différant sur deux traits acoustiques (par exemple un son fort et aigu et un son faible et grave) sont présentés avec une probabilité équivalente. Les sons déviants occasionnels ont un trait identique à l'un des standards sur une dimension et un trait identique à un autre standard sur l'autre dimension (par exemple un son fort et grave).

Ainsi, les deux traits acoustiques du déviant pris séparément appartiennent à une régularité acoustique et sont donc représentés en mémoire sensorielle auditive. Un tel déviant génère une MMN auditive qui ne peut être attribuée à aucun des deux traits élémentaires de déviance, et on peut en conclure que la conjonction des deux traits elle-même est représentée en mémoire sensorielle auditive. Un tel résultat a également été rapporté récemment dans la modalité visuelle (Winkler et coll., 2005).

Avec un tel protocole appliqué au cas audiovisuel, on peut donc tester si la conjonction audiovisuelle en tant que telle est représentée en mémoire sensorielle, et si oui dans quelle modalité : auditive, visuelle ou les deux. Nous avons donc présenté des évènements audiovisuels standards équiprobables (A_1V_1 et A_2V_2) et des déviants audiovisuels (A_1V_2 et A_2V_1), dont les composantes auditives et visuelles sont présentes dans les standards mais dont la conjonction est inédite par rapport aux standards. Si une représentation mnésique sensorielle de la régularité audiovisuelle existe en tant que telle, on devrait observer une différence dans le traitement des standards et des déviants, bien que les traits auditifs et visuels appartiennent chacun à une régularité unisensorielle. Nos expériences précédentes suggèrent que la régularité audiovisuelle est codée à la fois en mémoire sensorielle auditive et en mémoire sensorielle visuelle. Nous prédisons donc que cette MMN à la conjonction audiovisuelle devrait présenter à la fois des générateurs auditifs et visuels.

Pour cette expérience, deux contrôles importants doivent être réalisés pour éviter de confondre la MMN à la conjonction de traits audiovisuels avec d'autres processus. D'une part, la MMN ne doit pas être due à une différence de caractéristiques physiques entre déviants, ce qui n'a pas toujours été contrôlé dans les études de MMN à la conjonction auditive (voir par exemple : Gomes et coll., 1997 ; Sussman et coll., 1998). Dans notre expérience, 2 standards et 2 déviants étaient utilisés, les 2 déviants présentant les même traits auditifs et visuels que le 2 standards, si bien que les traitements propres aux différents traits auditifs et visuels disparaissaient dans le calcul de la MMN à la conjonction.

D'autre part, une MMN ne doit pas être générée par la détection d'une déviance locale dans une seule modalité. Il a en effet été montré, dans la modalité auditive, que des représentations des traits élémentaires et de la conjonction de traits pouvaient coexister en mémoire sensorielle auditive (Takegata et coll., 2001, 1999). De la même façon, une représentation de la conjonction audiovisuelle coexiste sans doute avec les représentations des parties unimodales de la régularité.

Or, pour éviter que l'alternance des deux standards ne soit elle-même une régularité et que la MMN soit provoquée par la violation de cette régularité, on doit présenter aléatoirement les 2 standards. Dans ce cas, un déviant à la conjonction peut être précédé du même standard présenté plusieurs fois. Comme, d'une part, il suffit de trois standards pour qu'une trace se constitue (Cowan et coll., 1993), et même moins dans le cas où le stimulus a déjà été présenté précédemment (Nousak, Deacon, Ritter & Vaughan, 1996 ; Winkler, Cowan, Csepe, Czigler & Näätänen, 1996), et comme, d'autre part, un déviant à la conjonction (par exemple A_2V_1) diffère d'un standard (par exemple A_2V_2) sur un des deux traits, il est possible que la MMN générée contienne une composante unisensorielle (visuelle dans notre exemple), générée par la probabilité locale dans la série présentée.

De la même façon, un standard donné diffère de l'autre standard ou d'un déviant sur

au moins un trait. Une MMN unisensorielle pouvait donc être générée également par un standard si les stimuli précédents présentaient plusieurs fois de suite le même trait. Pour contrôler de tels effets indésirables, nous nous sommes assuré que les pourcentages de stimuli présentant un trait auditif ou un trait visuel donné, précédés par un nombre donné de traits identiques dans l'autre modalité, étaient équivalents pour les standards et les déviants à la conjonction sur l'ensemble de l'expérience. Ainsi, toute MMN unisensorielle, qu'elle soit auditive ou visuelle, devrait disparaitre dans le calcul de la MMN à la conjonction.

Cette expérience a été réalisée en MEG. Puisque nous n'avions pas, au laboratoire, l'expérience de ce qu'est une MMN auditive en MEG (souvent appelée MMF pour *Mismatch Field*[1]), les sujets ont de plus participé à une expérience purement auditive dans laquelle des sons standards et déviants étaient présentés.

16.2 Méthodes

16.2.1 Sujets

Dix sujets droitiers (dont 5 de sexe féminin) âgés en moyenne de 29 ans (écart-type : 7 ans) ont participé à cette expérience. Aucun sujet ne souffrait de troubles neurologiques. Ils avaient tous une audition normale et une vision normale ou corrigée.

16.2.2 Stimuli

Les stimuli étaient identiques à ceux utilisés dans les expériences précédentes, excepté quelques détails : les mouvements des stimuli visuels étaient constitués de 5 trames d'une durée de 33 ms chacune. La durée des sons était de 167 ms (dont 10 ms de montée/descente)

16.2.3 Procédure

L'intervalle interstimulus était de 583 ms. 2600 stimuli audiovisuels (dont 312 déviants) ont été présentés, répartis dans 10 blocs d'une durée de 2 minutes 30 environ chacun.
Dans tous les blocs, les stimuli A_1V_1 et A_2V_2 étaient utilisés comme standards, avec une probabilité d'occurrence de 44% chacun. Les déviants étaient les stimuli A_1V_2 et A_2V_1 et avaient une probabilité d'occurrence de 6% chacun.
La tâche distractive était identique à celle des expériences précédentes, excepté que la disparition du point de fixation avait une probabilité d'occurrence de 10%.

Dans le paradigme auditif unimodal, les stimuli A_1 et A_2 jouaient, tour à tour, les rôles de standards et déviants avec des probabilités d'occurrence respectives de 88 et 12%. 1700 stimuli (dont 204 déviants) ont été présentés, répartis en 4 blocs de 4 minutes 10. Le sujet devait lire une livre de son choix et ignorer les sons.
L'expérience auditive unimodale était toujours réalisée à la suite de l'expérience audiovisuelle. Dans tous les cas, chaque bloc de stimuli commençait par au moins trois standards, et un déviant était toujours précédé d'au moins 3 standards.

[1]Pour plus de clarté, nous garderons la dénomination MMN bien que la négativité n'ait pas le même sens en MEG qu'en EEG.

16.2.4 Analyses

Les champs magnétiques évoqués (CME) de chaque sujet ont été calculés en excluant les 3 essais standards de début de chaque bloc, les essais standards suivant un déviant ainsi que ceux suivant une cible. Contrairement aux analyses EEG, le seuil de rejet des artéfacts était choisi pour chaque sujet de manière à ne pas rejeter plus de 85% des essais. La ligne de base était prise entre 100 ms et 0 ms avant la stimulation.

La MMN à la conjonction audiovisuelle était calculée comme la différence entre les CME aux stimuli déviants A_1V_2 et A_2V_1 et les CME aux stimuli standards A_1V_1 et A_2V_2.

Nous avons testé la différence entre les CME aux déviants et aux standards dans une fenêtre 140-300 ms par des tests bilatéraux de permutation des conditions appariées (Efron & Tibshirani, 1993, p212 ; $2^{10} = 1024$ permutations). Les tests multiples ont été pris en compte au niveau de chaque électrode par la méthode du minimum d'échantillons significatifs successifs, avec un risque local $\alpha = 0,05$ et un risque global $\alpha_{global} = 0,05$ (voir la partie 8.1 page 112).

Contrairement à l'enregistrement EEG, dans lequel les électrodes sont disposées par rapport à des repères anatomiques propres à chaque sujet, les capteurs MEG sont disposés de façon rigide les uns par rapport aux autres, et sans rapport précis avec l'anatomie des sujets. Selon la taille de la tête et sa position dans le casque MEG, les capteurs peuvent donc enregistrer des signaux de provenances légèrement différentes selon les sujets. Cela introduit une variabilité non négligeable lors du moyennage des données de plusieurs sujet et limite la puissance des tests statistiques de groupes. Notre étude a donc été complétée par des analyses statistiques sur les données individuelles de chaque sujet (voir la partie 8.2.1 page 113). Pour chaque essai, les champs magnétiques des essais élémentaires standards et déviants ont été comparés par des tests de randomisation pour groupes indépendants. Les tests multiples ont été pris en compte de la même façon que pour les tests de groupe.

Pour avoir une idée de la variabilité de la position de la tête des sujets dans le casque, nous avons mesuré l'écart moyen de position de la tête entre les sujets pris deux à deux, et pour un même sujet entre les deux parties de l'expérience, grâce à trois bobines électromagnétiques placées sur la tête du sujet dans le casque MEG.

16.3 Résultats

Le temps de détection des sujets dans la tâche distractive était de 418 ms (écart-type : 50 ms) et le taux de cibles manquées inférieur à 1%.
Les mesures de la position relative de la tête dans le casque MEG montrent que celle-ci variait en moyenne de 4,6 mm ± 0,6 mm au cours de l'expérience. La différence inter-sujet pouvait atteindre 40 mm.

La figure 16.1 page ci-contre montre les MMN auditive (16.1.A) et à la conjonction audiovisuelle (16.1.B). Les CME standards et déviants auditifs diffèrent significativement sur un grand nombre de capteurs entre 155 et 250 ms. La topographie de la différence

consiste en une inversion de polarité au niveau des capteurs temporaux correspondant à une activité vraisemblablement générée dans le plan supratemporal[2].

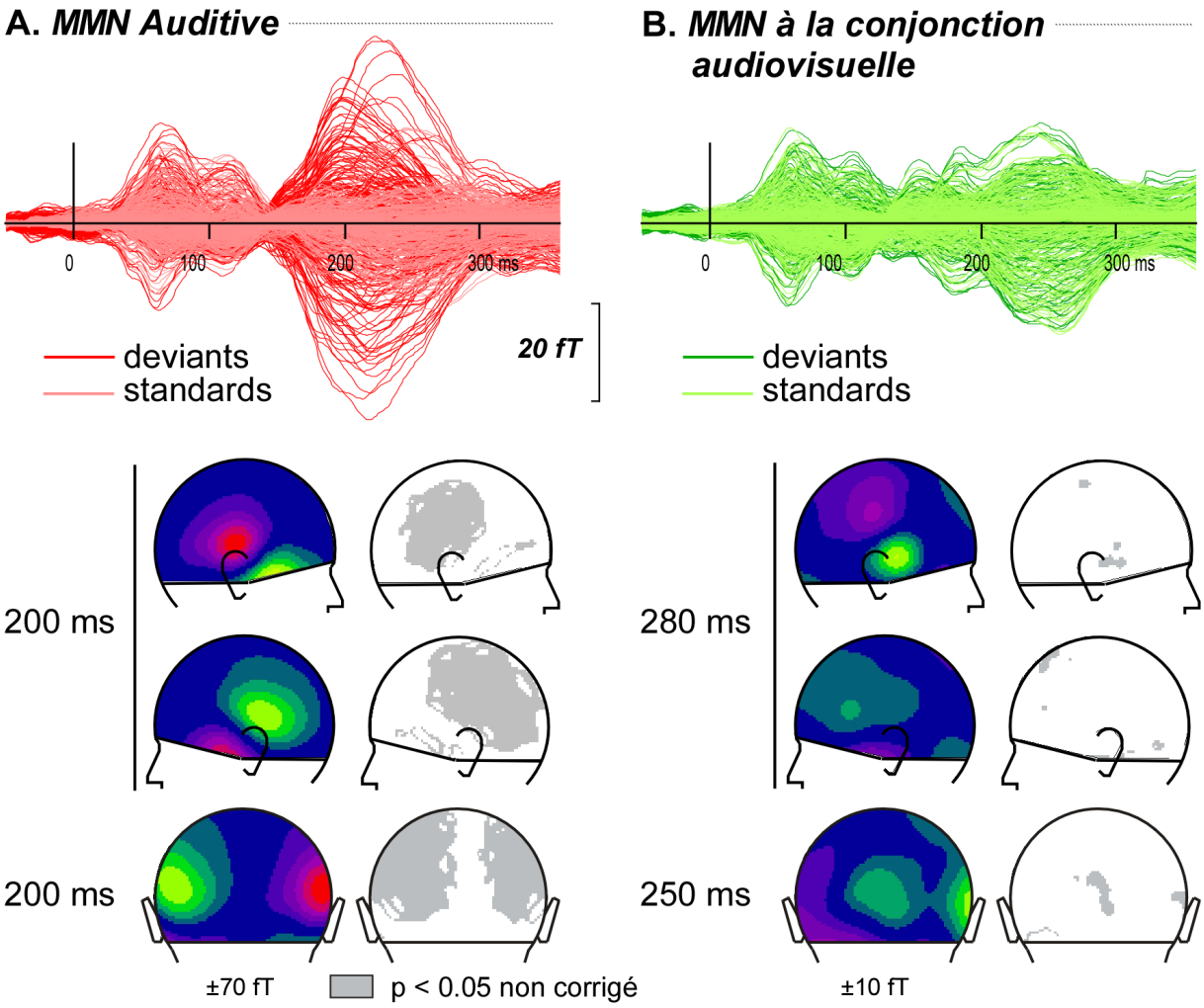

FIG. 16.1 – MMN auditive (A) et à la conjonction audiovisuelle (B). Les courbes sont les CME standards et déviants, enregistrés sur l'ensembles des capteurs et superposés. Les cartes en niveaux de gris indiquent les résultats au test de permutation de la différence entre CME standard et déviant à la latence de la topographie. Noter la différence d'échelle entre les cartes des deux MMN.

Concernant la MMN à la violation d'une conjonction audiovisuelle, les courbes évoquées par les déviants et les standards différaient à peine. Leurs amplitudes étaient cependant significativement différentes sur quelques capteurs occipitaux entre 235 et 265 ms et sur quelques capteurs temporaux gauches vers 280 ms, mais la différence ne subsistait pas à la prise en compte des tests multiples. La topographie de la différence est illustrées dans la figure 16.1.B à ces deux latences. Autour de 280 ms, elle présente une inversion de polarité

[2]Une inversion de polarité des signaux MEG signant une activité générée dans le cortex auditif doit, en principe, apparaitre perpendiculaire au plan supra-temporal. Rappelons toutefois que la topographie MEG représente les activités enregistrées au niveau du casque, et non du scalp du sujet comme en EEG, et dépend donc de l'orientation de la tête des sujets par rapport au casque.

Sujets	MMN auditive		MMN conjonction audiovisuelle		
	Temporal Gauche	Temporal Droit	Temporal Gauche	Temporal Droit	Occipital
S1	150-240	150-250	190-265	195-265	215-235
S2	190-260	200-275	-	285-330 ?	-
S3	160-230	160-230	-	-	-
S4	170-230	160-230	245-260	245-295	245-265
S5	170-250	170-240	265-275 ?	270-300 ?	280-305 ?
S6	180-260	160-270	-	-	-
S7	180-250	180-270	220-270 ?	175-205 ?	245-255 ?
S8	170-240	180-230	265-295 ?	-	-
S9	205-215	200-250	-	-	230-270
S10	150-230	170-230	275-295	275-295	245-255

TAB. 16.1 – Latences (en ms) de début et de fin des réponses significatives pour la MMN auditive et pour la MMN à la conjonction audiovisuelle, chez chacun des sujets. Le point d'interrogation désigne les réponses dont la topographie est instable.

similaire à celle de la MMN auditive et qui peut donc refléter des activités dans le cortex auditif. Par ailleurs, vers 250 ms, la topographie de la MMN à la conjonction audiovisuelle présente une composante postérieure sur les aires occipitales, qui n'est pas présente dans la MMN auditive unimodale

La table 16.1 donne, pour chaque sujet, les fenêtre de latence dans lesquelles les CME aux standards et aux déviants différaient significativement sur des capteurs temporaux ou occipitaux (seuil non corrigé). Alors que tous les sujets montraient une MMN auditive, seuls 3 sujets sur 10 montraient clairement une MMN à la violation de la conjonction audiovisuelle en regard des aires occipitales et des aires temporales de façon bilatérale, 5 sujets sur 10 ne montraient qu'une différence temporale unilatérale ou occipitale, seulement marginalement significative ou instable, et 2 sujets ne montraient aucune différence significative. La latence de ces différences était assez variable d'un sujet à l'autre, contrairement à la latence de la MMN auditive.

La figure 16.2 page ci-contre illustre la topographie de la MMN auditive et de la MMN à la conjonction audiovisuelle pour un sujet particulier (10)

16.4 Expérience comportementale complémentaire

Étant donnée la faiblesse de la MMN audiovisuelle, nous avons voulu savoir si les sujets étaient capables de détecter comportementalement une déviance à la conjonction audiovisuelle et comparer leurs performances à la détection d'une déviance auditive. Six sujets âgés de 30 ans (écart-type : 7 ans) ont participé à cette expérience complémentaire. Trois de ces sujets avaient participé à l'expérience MEG. Les stimuli auditifs et visuels étaient identiques à ceux de l'expérience MEG. La tâche des sujets consistait à cliquer le plus rapidement possible lors de la présentation d'un stimulus déviant à la conjonction audiovisuelle

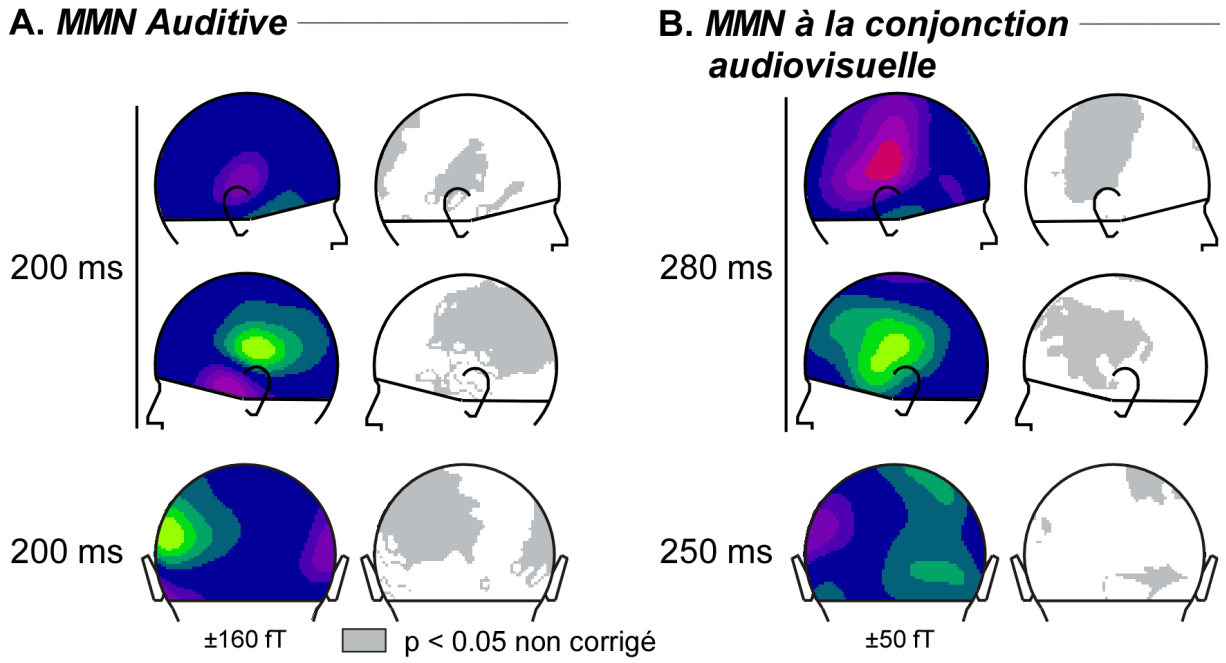

FIG. 16.2 – Topographie des MMN auditive (A) et à la conjonction audiovisuelle (B) chez le sujet S10

dans un bloc audiovisuel, et d'un stimulus déviant auditif dans un bloc contrôle auditif unimodal. Pour chacune des deux conditions, 600 stimuli (72 déviants) étaient présentés, répartis sur 6 blocs d'une durée approximative d'une minute. Les sujets étaient placés dans une situation la plus proche possible de celle des enregistrements MEG.

Les résultats de l'expérience comportementale sont rapportés dans le tableau 16.2 page suivante. Comparées à la détection auditive, les performances dans la détection d'une déviance à la conjonction audiovisuelle étaient assez médiocres puisqu'en moyenne les sujets ne détectaient que 67 % des cibles, avec un temps de détection relativement long. On peut également remarquer que, parmi les 3 sujets ayant participé aux deux expériences (S1, S9 et S10), les sujets S1 et S10, qui montraient une MMN à la conjonction, étaient également ceux dont les performances comportementales, dans la tâche de détection de la conjonction déviante, étaient les meilleures, comparées aux performances du sujet S9, qui ne montrait pas de MMN dans cette condition.

16.5 Discussion

Dans cette expérience, nous avons tenté de mettre en évidence l'existence d'une représentation intégrée de l'association particulière d'un trait auditif et d'un trait visuel en mémoire sensorielle. Les résultats sont moins clairs que ceux des expériences précédentes, dans lesquelles nous cherchions à mettre en évidence l'influence d'une régularité audiovisuelle sur la représentation d'une régularité auditive en mémoire sensorielle.

	Cibles manquées (%)		TR (ms)	
Sujets	Auditif	Audiovisuel	Auditif	Audiovisuel
S'1	2	47	406	750
S'2 (=S1)	0	2	407	673
S'3 (=S10)	2	16	425	628
S'4	2	50	378	778
S'5	5	23	390	771
S'6 (=S9)	0	56	461	739
Moyenne (± écart-type)	2 ± 2	33 ± 22	411 ± 29	723 ± 60

TAB. 16.2 – Performances et TR de détection dans l'expérience comportementale complémentaire.

Plusieurs facteurs peuvent expliquer la faiblesse des effets observés. Une première raison pourrait tenir à la technique d'enregistrement utilisée : les activités magnétiques générées par des dipôles radiaux sont relativement invisibles à la MEG. Or, lorsque l'on considère la topographie des différences de réponse obtenues dans les deux expériences précédentes, elles évoquent une origine plutôt radiale que tangentielle au scalp. La faible MMN à la conjonction observée dans la présente expérience, pourrait n'être constituée que de la composante tangentielle d'une activité principalement radiale.

Par ailleurs, il été montré que l'amplitude de la MMN dépend de la force de trace mnésique, c'est-à-dire de la régularité et de la fréquence avec lesquelles le ou les standards sont présentés : ainsi l'amplitude de la MMN auditive est plus faible lorsque deux standards plutôt qu'un seul sont présentés (Winkler, Paavilainen & Näätänen, 1992) ou lorsque le standard varie légèrement sur un trait (Winkler et coll., 1990). Comme la mise en évidence d'une MMN à la conjonction de traits nécessitait l'utilisation de deux standards différents, on ne pouvait espérer obtenir une MMN d'amplitude importante.

Enfin, dans le domaine auditif, il a été montré que l'amplitude et la latence de la MMN sont corrélées à la capacité du sujet à détecter explicitement la déviance (Tiitinen et coll., 1994). Or notre expérience comportementale complémentaire montre que la plupart des sujets avaient beaucoup de difficulté à détecter la déviance à la conjonction audiovisuelle, tant en termes de performances qu'en termes de temps de traitement. De plus il semble exister un lien entre la force de la trace et les performances puisque les sujets les plus performants dans la tâche de détection étaient également ceux qui montraient la MMN à la conjonction la plus robuste. Dans ce cas, la détection d'une violation de la conjonction audiovisuelle pourrait être un processus automatique basé sur l'existence d'une représentation de la régularité audiovisuelle, indexée par la MMN à la conjonction d'amplitude assez faible que nous observons. Le nombre de sujets est cependant insuffisant pour conclure sur cette corrélation.

Toutefois, il se pourrait également que la tâche à réaliser pour détecter la conjonction de deux traits auditif et visuel repose sur des processus différents de la détection d'une trait simple dans une modalité et non sur l'existence d'une trace mnésique de la régularité audiovisuelle. En cela, la détection de la déviance à la conjonction audiovisuelle se distinguerait du cas purement auditif puisque que les déviations de conjonctions de traits

acoustiques donnent lieu à des MMN relativement robustes (Gomes et coll., 1997 ; Sussman et coll., 1998 ; Takegata et coll., 2001, 1999 ; Winkler et coll., 2005), ce qui suggère qu'il existe des processus automatiques de détection de la violation d'une conjonction analogues à ceux de la détection d'un trait dans le cas purement auditif. Cependant, à ma connaissance, le détection d'une déviation à une conjonction auditive n'a pas été testé comportementalement.

Cinquième partie

Discussion générale

Chapitre 17

Discussion générale

17.1 Interactions audiovisuelles précoces dans la perception de la parole

Notre première expérience sur la perception audiovisuelle de la parole (chapitre 9 page 119) a montré, d'une part, que le temps de traitement de la parole auditive était diminué par la vision des mouvements articulatoires associés, sans que cette diminution ne puisse s'expliquer dans un modèle de traitement séparé des information auditives et visuelles. D'autre part, ce gain comportemental semblait associé à une diminution de l'activité auditive entre 120 ms et 200 ms de traitement.

Notre seconde expérience chez le patient épileptique (chapitre 10 page 131) a montré que cet effet n'est pas le seul à prendre place dans les 200 premières millisecondes de traitement de la syllabe auditive. Les effets les plus reproductibles correspondent, d'une part, à une activation du cortex auditif par les indices visuels de parole et, d'autre part, à une modulation (essentiellement une diminution) du traitement des indices auditifs dans le cortex auditif, dont une partie pourrait correspondre aux effets observés en EEG de scalp. L'activation du cortex auditif par les mouvements articulatoires de la parole semble avoir lieu directement après le traitement de ces stimuli dans les aires visuelles (bien que la couverture spatiale éparse des électrodes intracérébrales ne permette pas de conclure définitivement sur ce point) et correspondrait donc à une activation *feedforward* (en termes temporels). Rappelons que l'information visuelle est disponible avant l'information auditive dans les stimuli de paroles. Cette activation visuelle du cortex auditif pourrait ensuite permettre la modulation du traitement phonétique des syllabes auditives dans le cortex auditif, et ce à des étapes de traitement relativement précoces (à partir de 50 ms). Le cortex auditif primaire semble relativement épargné par ces phénomènes.

Dans notre discussion de l'expérience en EEG de scalp (page 127), nous avions proposé plusieurs possibilités d'interprétation des violations de l'additivité :

- activation du cortex auditif par les indices visuels.
- effet d'indiçage temporel intersensoriel.
- intégration des informations phonétiques auditives et visuelles à un stade pré-phonologique.
- amorçage phonologique intersensoriel.

Notre expérience en sEEG a bien montré que l'on pouvait dissocier les activations du cortex auditif par les stimuli visuels, des modulations de l'activité auditive sous influence visuelle, et donc que l'activation visuelle du cortex auditif ne pouvait expliquer la violation de l'additivité dans l'étude en EEG de scalp, ni a fortiori des autres violations du modèle additif en sEEG.

Nos expériences comportementales (chapitre 11 page 153) avaient pour but de tester l'existence d'un effet d'indiçage temporel intersensoriel sur le TR, c'est-à-dire des interactions audiovisuelles dans la perception de la parole ne reposant pas sur une intégration phonétique ou phonologique. Les résultats montrent que, si un tel effet existe et peut être mis en évidence, il s'observe uniquement lorsque la performance est diminuée par la présence de bruit. Il est donc peu probable que l'indiçage temporel intersensoriel explique à lui seul le gain de TR de la première expérience et l'ensemble des interactions audiovisuelles mises en évidence par nos mesures électrophysiologiques.

Il reste donc les deux dernières possibilités. Toutes deux impliquent l'existence d'une intégration des informations auditives et visuelles phonétiques. La première correspondrait plutôt à une intégration pré-phonologique, alors que la seconde (amorçage) préserve la possibilité d'une intégration post-phonologique. Nos données sEEG pointent plutôt vers l'hypothèse d'un amorçage, étant donné l'activation massive du cortex auditif par les indices visuels avant la présentation de la syllabe auditive. Bien entendu, le protocole utilisé dans les expériences électrophysiologiques ne permet pas de statuer sur le caractère précatégoriel ou catégoriel des représentations impliquées dans l'intégration. Pour exclure l'hypothèse d'amorçage, il faudrait utiliser des stimuli de parole dans lesquels les informations phonétiques visuelles et auditives sont synchrones. L'utilisation du modèle additif dans ce cas permettrait de caractériser la dynamique spatio-temporelle de véritables interactions phonétiques audiovisuelles.

17.2 Représentation d'un évènement audiovisuel en mémoire sensorielle auditive

Notre première expérience comportementale (chapitre 13 page 179) a montré que la détection d'une déviance audiovisuelle était plus rapide que prédit par un modèle d'indépendance des systèmes de détection auditifs et visuels. Notre seconde expérience, en EEG de scalp (chapitre 14 page 185), suggère qu'une partie des interactions audiovisuelles pouvant expliquer cette facilitation concerne l'accès à des représentations mnésiques sensorielles des évènements standards auditifs et/ou visuels (indexés par les MMN auditive et visuelle). Ces deux expériences montrent donc, a minima, l'existence d'interactions entre les systèmes auditif et visuel de détection de la déviance, eux-même basés sur l'existence de registres mnésiques spécifiques à chacune des deux modalités sensorielles.

D'un autre côté, la différence entre les MMN visuelles générées par des évènements purement visuels et audiovisuels (première expérience EEG, chapitre 14 page 185) et la différence entre les MMN auditives générées par des évènements purement auditifs et audiovisuels (seconde expérience EEG, chapitre 15 page 195), suggèrent que l'existence d'une association régulière entre stimuli auditifs et visuels peut modifier la construction de cha-

cune de ces deux traces mnésiques. La représentation d'une régularité audiovisuelle pourrait donc passer par l'inclusion réciproque d'informations auditives et visuelles dans les mémoires sensorielles sensori-spécifiques, comme le montre la topographie de la MMN à la double déviance auditive et visuelle dans la première expérience EEG. Le fait que cette modification de la MMN auditive n'ait pas lieu lorsque l'association entre le stimulus auditif et le stimulus visuel ne constitue pas elle-même un régularité (condition audiovisuelle équiprobable de la seconde expérience EEG, chapitre 15 page 195), suggère que cette inclusion réciproque correspond bien à la représentation de la régularité de l'association des deux traits auditif et visuel.

En revanche, nous ne sommes pas parvenu à montrer de façon convaincante que la violation d'une telle régularité audiovisuelle, sous la forme d'une violation de la conjonction des traits auditifs et visuels, suffit à générer activité de type MMN (chapitre 16 page 205). Il est donc possible que la représentation de la régularité audiovisuelle mise en évidence dans les trois premières expériences n'aboutisse pas à une véritable trace intégrée de la régularité audiovisuelle, permettant une détection automatique et rapide de sa violation. Une explication alternative serait que cette représentation existe mais que la force de la trace mnésique est trop faible dans notre protocole pour permettre une détection rapide de la déviance à la conjonction.

De manière générale, les interactions audiovisuelles mises en évidence dans cette série d'expériences étaient de faible amplitude et toujours à la limite de la significativité statistique, et donc probablement à la limite de la sensibilité de la technique d'enregistrement utilisée. En cela, elles s'opposent aux effets audiovisuels massifs provoqués par les syllabes McGurk, l'illusion de ventriloquie ou des stimuli audiovisuels entretenant des liens plus étroits ou plus écologiques. En utilisant de tels stimuli peut-être pourrait-on mettre en évidence de manière plus convaincante une MMN à la conjonction audiovisuelle.

17.3 Interactions audiovisuelles dans le cortex auditif

Nos deux séries d'expériences ont mis en évidence, entre autres, une influence des informations visuelles sur les traitements réalisés dans le cortex auditif. Cette influence pouvait se manifester de deux façons : soit par une activation des structures auditives en réponse à un stimulus visuel (cas de la parole), soit par une modulation de l'activité enregistrée en réponse à un stimulus auditif (cas de la parole et des représentations en mémoire sensorielle auditive).

Comment peut-on expliquer ces activations intersensorielles au regard de l'architecture connue des systèmes sensoriels auditif et visuel et en particulier de leurs interrelations ? Depuis la mise en évidence d'effets d'interaction multisensorielle dans les cortex sensori-spécifiques par les méthodes d'imagerie fonctionnelle chez l'homme (par exemple : Calvert et coll., 1997 ; Giard & Peronnet, 1999 ; Calvert et coll., 1999), principalement deux hypothèses anatomiques, pouvant expliquer ces effets, ont été proposées, l'une dans le cadre du modèle classique de convergence tardive, l'autre en opposition à ce modèle.

On trouve dans une revue de Mesulam (1998), une description détaillée de l'architecture du modèle classique de convergence tardive chez l'homme. Dans ce modèle, les aires

corticales auditives et visuelles sont totalement ségrégées dans le sens ascendant (*feedforward*) : il n'existe ni connections latérales entre cortex de différentes modalités sensorielles et encore moins de projections sous-cortico-corticales intersensorielles (les projections sous-cortico-sous-corticales ne sont pas discutées). Les informations de différentes modalités sensorielles ne convergent qu'au niveau d'aires associatives hétéromodales, qui sont au nombre de quatre : le cortex pré-frontal, le cortex pariétal postérieur, le cortex temporal latéral (dont le STS) et le gyrus para-hippocampique, qui sont analogues à celles mises en évidence chez l'animal (voir la partie 1.5 page 17). Cependant, un aspect important de ce modèle, est que toutes les connections sont bidirectionnelles. Ainsi, même en l'absence de connections ascendantes ou latérales entre systèmes sensoriels, une influence intersensorielle dans les cortex modalité-spécifiques est possible par le biais de projections descendantes (*feedback*), depuis les aires associatives hétéromodales. Plusieurs auteurs ont proposé que ces voies descendantes soient à l'origine des effets intersensoriels dans les cortex auditifs ou visuels (par exemple : Calvert, 2001 ; Driver & Spence, 2000).

Pour d'autres auteurs, une partie de ces effets doit nécessairement s'expliquer en sortant de ce modèle (Schroeder et coll., 2003 ; Schroeder, Molholm, Lakatos, Ritter & Foxe, 2004 ; Bulkin & Groh, 2006 ; Ghazanfar & Schroeder, 2006). Ces auteurs se basent principalement sur deux arguments. D'une part, selon Foxe et Schroeder (2005), certains effets audiovisuels ont une latence trop courte pour être explicables par des projections descendantes (par exemple : Giard & Peronnet, 1999 ; Molholm, Ritter, Javitt & Foxe, 2004 ; Fort et coll., 2002a). D'autre part, des projections intersensorielles latérales et ascendantes entre cortex auditif et visuel existent : les premières ont déjà été mentionnées dans la partie 1.5 page 18 et ont été principalement observées du cortex auditif vers le cortex visuel ; selon Schroeder et coll. (2003), les secondes correspondraient au système de projection thalamo-cortical koniocellulaire (aussi appelé non spécifique), un système de projection diffus ne respectant pas la ségrégation des aires corticales (revue dans E. G. Jones, 2001). Bulkin et Groh (2006) mentionnent également l'existence de connections sous-cortico-sous-corticales audiovisuelles, par exemple entre le colliculus supérieur et le colliculus inférieur (une structure sous-corticale auditive ; Doubell, Baron, Skaliora & King, 2000)

Les effets que nous avons mis en évidence dans le cortex auditif nécessitent une analyse détaillée des stimuli, afin, soit de discriminer les informations phonétiques, soit de distinguer les déformations de cercles dans différentes directions. Cela semble exclure les projections koniocellulaires qui ne possèdent pas une spécificité spatiale suffisante pour porter de telles informations (Schroeder et coll., 2003). De même le colliculus supérieur semble être impliqué dans des fonctions liées plus à la détection et la localisation du stimulus visuel qu'à son identification (voir la partie 1.4.1 page 13).

Les résultats concernant la mémoire sensorielle auditive donnent peu d'informations temporelles permettant de trancher entre projections latérales ou descendantes. En effet, ils ne donnent qu'une borne temporelle supérieure des interactions audiovisuelles nécessaires à l'inclusion d'informations visuelles dans la trace sensorielle auditive. Or ces effets étaient observés à la latence de la MMN, c'est-à-dire autour de 200 ms.

Dans le cas le parole, nous avions proposé (Besle, Fort, Delpuech & Giard, 2004, annexe page 245) que la modulation de l'activation du cortex auditif pouvait être due à des

connections *feedback* entre le STS et le cortex auditif, le STS pouvant être activé par les informations visuelles présentées en avance. Cette proposition étaient basée sur le manque de données anatomiques montrant des projections d'aires sensorielles visuelles vers les aires sensorielles auditives, sur l'existence de projections d'aires polysensorielles homologues du STS vers le cortex auditif (Pandya, Hallett & Kmukherjee, 1969 ; Seltzer & Pandya, 1978), ainsi que des données électrophysiologiques chez le macaque, montrant que l'influence d'un stimulus visuel sur le traitement d'un son dans le cortex auditif possède un profil laminaire (profil spatial le long de différentes couches du cortex) de type descendant (Schroeder & Foxe, 2002). Depuis, des projections directes du cortex visuel vers le cortex auditif ont été mises en évidence (Hishida et coll., 2003 ; Cappe & Barone, 2005), mais il semble que, même dans ce cas, les aires visuelles d'origine possèdent déjà un caractère multisensoriel audiovisuel (voire la partie 1.5 page 18).

Les données en sEEG suggèrent que des composantes auditives relativement précoces (50 ms) sont modulées par les informations visuelles. Le problème est qu'il existe une asynchronie fondamentale entre les composantes visuelles et auditives de la parole, et que, dans nos stimuli, les informations visuelles étaient disponibles avant les informations auditives. De fait, les interactions audiovisuelles dans le cortex auditif étaient précédées d'une activation visuelle de ce même cortex, dès 120 millisecondes avant la présentation du son, soit 170 ms avant les premières modulations de l'activité auditive. Ce laps de temps est largement suffisant pour permettre à des effets descendants de se mettre en place.

En revanche, l'activation du cortex auditif par les stimuli visuels semble relativement précoce, puisqu'elle suivait immédiatement celle d'aires occipito-temporales et temporales postérieures, vraisemblablement visuelles. Les activations provenant du STS semblent, elles, se produire plus tardivement. Mais le STS est une vaste structure qui ne présente sans doute pas une unité fonctionnelle très prononcée et certaines parties du STS n'ont pas du tout été explorées chez nos patients. De même, les zones visuelles occipitales n'ont pas été explorées dans cette étude, si bien qu'on ignore à quelle latence avaient lieu les premières activations visuelles. Il est donc difficile de dire exactement à quel point les activations du cortex occipito-temporal et du cortex auditif étaient "précoces" et donc de se prononcer sur la nature *feedback* ou *feedforward* de ces activations.

Annexe A

Données individuelles des patients

Patient	Région explorée	type de réponse	Latence de début (ms)	Latence de fin (ms)	Côté	Nom des contacts	Coordonnées de Talairach X	Y	Z
8	STS postérieur	1	-80	400	G	E'6-8	-42	-53	1
9	GTM postérieur	1	-100	160	G	V'12-14	-46	-63	12
6	GTI/STI postérieur/Gyrus fusiforme	1	-80	140	D	L11	55	-55	0
10	Jonction occipito-temporale	1	-40	600+	G	V'10-12	-35	-65	1
10	Jonction occipito-temporale	1	-40	600+	D	V9-12	34	-64	1
10	Gyrus occipito-temporal ventral/fissure calcarine	1	-40	600+	D	V7-9	26	-63	2
10	Gyrus occipito-temporal supérieur	1	-20	350	G	W'9-10	-34	-63	11
6	Gyrus occipito-temporal ventral	1	60	350	D	L3-4	32	-55	0
7	Planum temporale/gyrus supra-marginal	2	-20	450	G	G'11-14	-56	-40	21
5	Planum temporale/STS	2	0	600+	G	G'3-5	-57	-39	17
8	STS/GTS	2	-60	500	G	T'8-9	-62	-10	-3
8	Planum temporale	2	0	550	G	H'10-15	-47	-24	6
8	Insula/ planum polare	2	0	600+	G	T'3	-40	-10	-2
3	Fond du STS	2	-40	600+	D	C8	46	-16	-11
3	GTM latéral	2	-40	600+	D	C13	57	-15	-12
3	STS antérieur	2	-120	600+	D	T7-9	57	-7	1
3	GTS/ gyrus transverse antérieur latéral	2	100	600+	G	T'6-9	-57	-4	4
1	Gyrus transverse postérieur médial/ Planum temporale	2	-20	600+	D	H7-10	42	-29	7
1	Gyrus transverse antérieur latéral	2	40	600+	D	T7-10	54	-9	-4
8	Insula/Gyrus transverse médial	2	160	550	D	T2	34	-12	6
2	Gyrus transverse antérieur médial	3	-120	450	D	H6-7	33	-25	11
2	Planum temporale	3	-120	450	D	H8-12	44	-27	10
10	Planum temporale/Gyrus transverse	3	-20	600+	D	H11-14	57	-19	6
10	Gyrus transverse antérieur médial	3	-20	600+	D	H6-9	33	-19	7
3	Planum temporale /STS	3	60	600+	G	H'6-13	-46	-30	5
7	Planum polare	3	140	300	G	T'2-7	43	-10	7
7	STS antérieur	4	60	600+	G	A'10-11	-49	-8	-16
3	GTM latéral	4	60	600+	G	B'11	-62	-20	-14
3	GTS/STS	4	160	600+	D	A11	57	-3	-9
7	STS *	4	160	600+	D	B7-9	49	-23	-9
8	MTG/STI (A'9-10)	4	220	600+	G	A'9-10	-56	0	-19
9	STS	5	40	600+	G	H'11-13	-60	-24	6
6	Bord inférieur STS	5	60	600+	D	B10	49	-16	0
5	GTS	5	220	600+	G	H'8-10	-61	-24	7
4	Fond du STS/ STI *	5	140	300	G	B'7-9	-46	-24	-9
4	GTS/STS antérieur	5	240	600+	G	T'6-9	-54	-15	-2
4	Bord supérieur STS	5	260	600+	G	H'11-15	-58	-28	10
4	GTM/STS	5	300	600+	G	B'10-12	-56	-24	-9

Patient	Région explorée	type de réponse	Latence de début (ms)	Latence de fin (ms)	Côté	Nom des contacts	Coordonnées de Talairach X	Y	Z
3	GTM	6	140	600+	G	L'13-14	-64	-43	-3
1	STS postérieur	6	160	450	D	R2	43	-39	3
10	Gyrus supramarginal/gyrus post-central	7	-20	600+	D	E13	52	-26	33
8	Gyrus supramarginal G	7	80	450	G	G'11-15	50	-33	33
8	Gyrus supramarginal D	7	160	600+	D	G15	57	-31	32
4	STS postérieur (lésion...)	8	100	400	D	X11	47	-53	16
5	Fond du STS postérieur/ gyrus supra marginal	8	100	500	G	Y'8-10	-35	-50	23
10	Gyrus supramarginal/ gyrus post-central/sillon post-central *	8	100	200	D	E8-10	31	-26	33
1	Gyrus angulaire/gyrus supramarginal *	8	220	350	D	V11-12	47	-43	34
10	Gyrus supramarginal/gyrus post-central/sillon post-central	8	240	400	D	E11	46	-25	33
10	Gyrus angulaire/supramarginal *	8	300	600+	D	G11-12	44	-43	26
5	Gyrus supramarginal	8	300	600+	G	Y'13-15	-53	-49	24
7	Gyrus post-central	9	100	260	G	R'1	-47	-18	27
8	Sillon post central D *	9	160	600+	G	G'8	-31	-36	30
5	Opercule post-central	9	220	600+	G	N'8-10	-58	-14	20
2	Opercule post-central	9	160	600+	D	N2	33	-14	12
7	Insula antérieure G/gyrus post-central	9	160	450	G	P'2-3	-37	-3	3
7	Insula/opercule post-centrale	9	240	600+	G	N'2-3	-38	-10	20
5	Insula/opercule post-central*	9	400	600+	G	N'3-4	-38	-14	20
4	Insula*	10	140	400	D	T4-5	38	-15	2
8	Insula/Planum temporale	10	160	550	D	T2	34	-12	6
4	Insula/Planum polare/gyrus transverse antérieur latéral *	10	220	400	G	T'4-5	-42	-15	-2
8	Gyrus cingulaire postérieur/ précunéus*	11	160	600+	G	G'3-4	-14	-37	30
10	Cingulaire postérieur*	11	200	600+	D	G3-4	12	-45	26
10	Gyrus cingulaire postérieur	11	220	600+	G	G'3	-12	-40	28
6	Opercule précentral	12	-60	550	D	N4-7	43	-7	10
1	Opercule précentral	12	140	600+	D	N7-9	52	12	21
7	Opercule précentral	12	240	600+	G	P'8	-58	-1	18
7	Gyrus frontal inférieur postérieur	12	240	600+	G	E'5-8	-51	11	7
9	Opercule précentral	12	240	400	G	P'4-10	-48	6	5

Patient	Région explorée	type de réponse	Latence de début (ms)	Latence de fin (ms)	Côté	Nom des contacts	Coordonnées de Talairach X	Y	Z
7	Hippocampe	13	80	400	G	B'2-5	-35	-23	-5
8	Hippocampe	13	180	600+	D	B'1-2	-30	-20	-10
8	Hippocampe	13	220	600+	D	B2-3	30	-22	-7
10	Gyrus parahippocampique/amygdale	13	220	600+	D	D3	22	-3	-33
10	Hippocampe/insula/gyrus transverse médial	13	260	600+	G	B'2-4	-34	-15	-10
9	Gyrus parahippocampique/gyrus lingual	13	300	600+	G	L'2	-18	-48	5
10	Gyrus lingual *	14	0	120	D	L5	32	-31	-15
10	Fond fissure calcarine *	14	280	400	D	W7	25	-63	9
10	GTM postérieur *	14	450	600+	D	W12	41	-63	7
10	Sillon frontal inférieur	14	160	300	D	K9-11	33	37	24
6	Gyrus frontal médian *	14	400	600+	D	K10-11	37	26	21
2	Opercule post-central inférieur/gyrus transverse	14	80	550	D	N4-5	43	-15	12
5	Sillon intrapariétal *	14	160	600+	G	Q'6	-23	-47	50
6	Gyrus supramarginal/sillon intrapariétal	14	-60	600+	D	R7-11	30	-48	35
8	Sillon colatéral	14	80	600+	G	D'5-7	-44	-5	-30
8	Gyrus temporal ventral antérieur*	14	120	350	G	A'4-7	-39	-1	-19
5	Cunéus/fissure pariéto-occipitale *	14	0	600+	G	W'3-5	-14	-68	36
5	Lobule pariétal inférieur/ fissure pariéto-occipitale *	14	0	600+	G	W'6-7	-26	-66	36
10	Gyrus lingual	14	180	600+	D	L2-5	28	-31	-14
10	Gyrus lingual	14	200	600+	G	L'4-5	-30	-30	-13
9	Sillon colatéral	14	300	600+	D	E'5-6	-42	-30	-9

TAB. A.1 – Coordonnées, localisation et latence des réponses aux syllabes visuelles. Type de réponse : 1. Activité spécifique à la condition visuelle enregistrée autour du GTM postérieur. 2. Activité enregistrée dans le lobe temporal supérieur et dont les sources ressemblent à celles de la réponse auditive entre 50 et 100 ms. 3. Activité enregistrée dans le lobe temporal supérieur et dont les sources ressemblent à celles de la réponse auditive après 100 ms. 4. Réponse autour du STS antérieur, commune aux conditions auditives et visuelles. 5. Activité enregistrée dans le lobe temporal supérieur, spécifique à la condition visuelle. 6. Réponse autour du STS postérieur commune aux conditions auditive et visuelle. 7. Réponse autour du gyrus supramarginal spécifique à la condition visuelle. 8. Réponse autour du gyrus supra marginal commune aux conditions auditives et visuelles. 9. Activité enregistrée autour de l'opercule post-central ou de l'insula, commune aux conditions A et V. 10. Réponse visuelle dans l'insula, sans réponse équivalente en condition auditive (mais peut-être cachée par l'activité provenant du cortex auditif). 11. Activité enregistrée autour du gyrus cingulaire. 12. Réponse enregistrée autour de l'opercule pré-central et du gyrus frontal inférieur. 13. Activité enregistrée autour de l'hippocampe et du gyrus parahippocampique. 14. Activités diverses. Les régions suivies d'une étoile sont celles dans lesquelles la réponse n'était significative qu'en montage bipolaire. 600+ : la réponse continue au-delà de 600 ms post-stimulus. GTM : Gyrus temporal moyen. GTS : Gyrus temporal supérieur. STI : Sillon temporal inférieur. STS : Sillon temporal supérieur.

Patient	Région explorée	type de réponse	Latence de début (ms)	Latence de fin (ms)	Côté	Nom des contacts	Coordonnées de Talairach X	Y	Z
6	Gyrus transverse antérieur médial	1	30	600+	D	H3-5	39	-20	5
2	Planum temporale antérieur	1	40	110	D	T8-10	61	-13	-2
7	GTS / gyrus supramarginal	1	70	600+	G	G'9-14	-54	-39	21
8	Planum temporale	1	70	500	G	H'10-15	-58	-23	6
3	GTS supérieur	1	80	450	G	T'8	-59	-4	4
10	Gyrus transverse médial	1	90	600+	D	H7-10	38	-19	7
3	Matière blanche du GTS	1	100	250	G	H'12-15	-61	-28	5
1	Planum temporal/ gyrus transverse	1	110	250	D	H8-10	44	-28	7
3	Bord supérieur du STS	1	120	600+	D	T7-8	55	-7	2
8	Planum polare/gyrus transverse latéral	1	120	250	D	T6-7	51	-11	8
8	GTS	1	120	180	D	T9-10	62	-11	9
9	GTM/bord inférieur du STS *	1	120	250	G	L'12-14	-56	-47	5
10	Planum Temporale/Gyrus transverse latéral	1	130	600+	D	H11-13	57	-19	6
7	Gyrus précentral	1	130	500	G	N'6-7	-52	-10	20
5	Planum temporale/Gyrus supramarginal	1	130	250	G	G'4	-57	-39	17
8	Bord supérieur du STS/GTS	1	130	200	G	T'8-9	-61	-9	-3
8	Gyrus supramarginal	1	130	300	G	G'13	-50	-35	29
7	Planum polare/bord supérieur du STS	1	140	400	G	T'4-9	-51	-4	2
8	Insula/Planum polare	1	160	450	D	T2-3	37	-11	7
6	Gyrus transverse postérieur latéral **	2	40	120	D	H7-9	53	-20	5
8	Bord supérieur du STS **p<0,005	2	50	90	G	T'7-8	-58	-9	-3
3	GTS ** p<0,01	2	50	100	G	H'12-15	-61	-28	5
8	Planum temporale	2	60	120	G	H'11	-50	-23	7
3	Bord supérieur du STS** p<0,005	2	60	120	D	T7-8	55	-7	2
7	Bord supérieur du STS* p<0,05	2	60	100	G	T'7-8	-57	-4	2
10	Gyrus transverse médial	2	80	160	D	H7-10	38	-19	7
8	Planum temporale	2	80	160	G	H'13-15	-62	-23	7
7	Bord supérieur du STS/planum polare	2	120	200	G	T'5-7	-57	-4	2
1	Gyrus cingulaire postérieur *	3	20	60	D	W4	13	-50	16
10	Fissure calcarine	3	30	80	D	V2	8	-68	5
2	Bord supérieur du STS/GTS	3	40	110	D	C9-13	61	-25	-4
2	GTS/planum temporale	3	50	100	D	H15	63	-26	10
6	Cunéus	3	60	130	D	G5-6	19	-57	16
2	Planum temporale*	3	70	120	D	H12	52	-26	10
9	Bord inférieur STS	3	100	120	G	B'11	-59	-17	-13
7	Gyrus transverse postérieur latéral	3	120	250	G	H'8-9	-57	-22	9
8	Gyrus cingulaire postérieur/précunéus	3	120	300	G	G'4	-15	-36	30
9	STS/GTS *	3	120	250	G	H'13	-62	-22	6
6	Insula	3	120	200	D	T3	38	-10	0
6	GTI ventral postérieur	3	130	200	D	L6-7	39	-55	-8
3	MTG	3	140	550	G	B'12	-66	-19	-14
6	Gyrus précentral	3	140	200	D	N7	50	-7	10
6	STI/GTI	3	140	250	D	L10	51	-54	-8
9	GTM/bord inférieur du STS *	3	160	250	G	V'14-15	-51	-61	12

TAB. A.2 – Coordonnées, localisations et latences des violations du modèles additif commençant entre 0 et 200 ms. Type de violation de l'additivité : 1. le profil spatiotemporel de la violation est identique à celui de la réponse visuelle et de polarité opposée. 2. Le profil spatiotemporel est identique à celui de la réponse auditive et de polarité opposée. 3. autre type de violation. * la violation n'était significative qu'en montage bipolaire. ** la violation n'était significative qu'en montage monopolaire. 600+ : la violation continue au-delà de 600 ms post-stimulus. GTS : Gyrus temporal supérieur. STS : Sillon temporal supérieur.

FIG A.1 (page suivante) - Les représentations tridimensionnelle et bidimensionnelle du ruban cortical sont propres au patient. La représentation tridimensionnelle est celle du lobe temporal et les représentations bidimensionnelle sont faites dans le plan coronal colinéaire à l'axe de de pénétration des électrodes. Sur les cartes de profil spatio-temporel, les zones entourées en jaune sont les échantillons significatifs au seuil corrigé. L'amplitude indiquée sous chaque couple de cartes monopolaire/bipolaire correspond aux couleurs les plus vives aux extrémités de l'échelle (jaune pour une différence de potentiel positive et rouge pour une différence de potentiel négative).

Pour l'électrode T, la réponse visuelle était soutenue entre 40 et 600 ms et le profil spatial ressemble à la première composante auditive entre 50 et 100 ms (foyers négatif sur T6-10 en monopolaire et foyers négatifs sur T6 et T8-10 en bipolaire). Pour l'électrode H, la réponse visuelle est constituée de plusieurs foyers positifs entre -20 et 600 ms que l'on retrouve en condition auditive, en particulier la réponse auditive transitoire vers 100 ms sur H9-10. En montage bipolaire, on retrouve une inversion de polarité autour de H10 dans les deux modalités. De manière générale, la réponse visuelle est plus soutenue que la réponse auditive, comme on peut le voir facilement sur les courbes. Le profil spatio-temporel de la violation présente une ressemblance évidente avec la réponse visuelle sur ces deux électrodes.

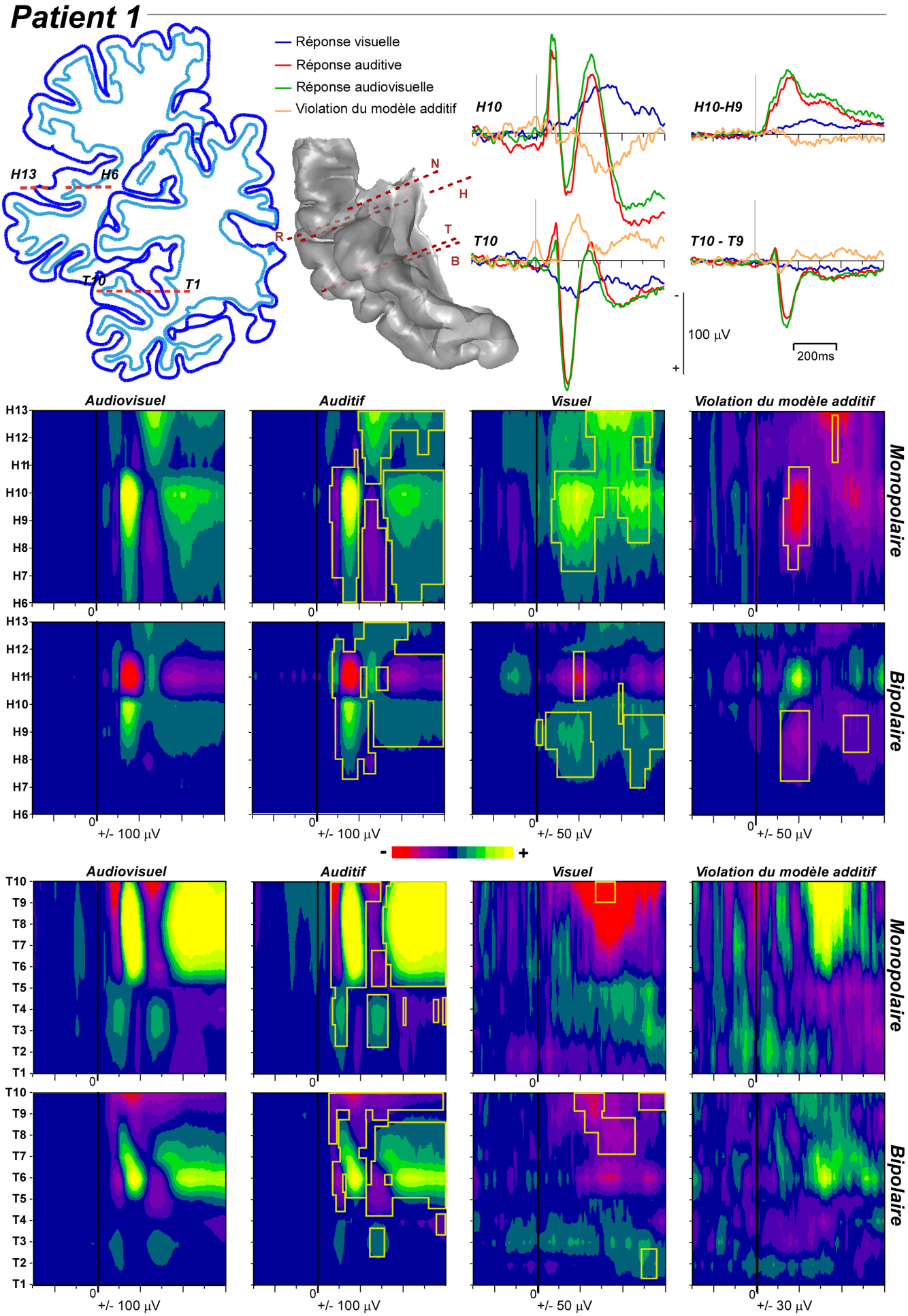

FIG. A.1 – Localisation et activités enregistrées aux électrodes H et T (hémisphère droit) pour le patient 1.

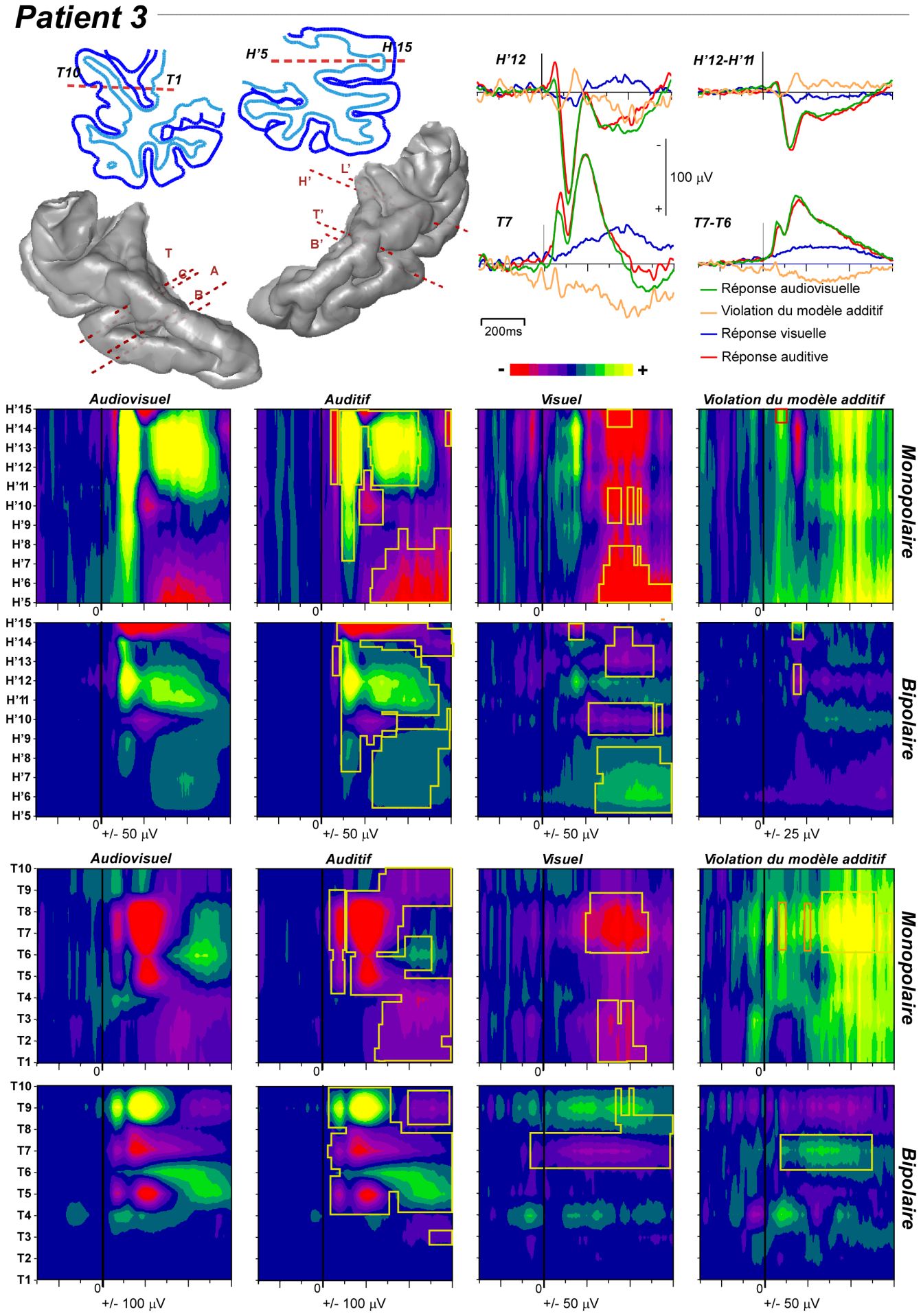

FIG. A.2 – Localisation et activités enregistrées aux électrodes H' (hémisphère gauche) et T (hémisphère droit) pour le patient 3.

FIG A.2 (page ci-contre) - Sur l'électrodes H', la première réponse visuelle significative apparait sur le contact H'15 vers 100 ms comme une composante positive en montage monopolaire ; cette première réponse ressemble à la réponse auditive entre 90 et 160 ms. Cette première réponse visuelle est suivie d'une réponse plus soutenue à partir de 160 ms qui semble ne correspondre à aucune composante auditive. Sur l'électrode T, la réponse visuelle soutenue commençant à -120 ms (elle est significative à partir de -70 sur T7 en bipolaire) sur T7-9 a le même profil spatial que les deux réponses auditives transitoires enregistrées entre 40 et 100 ms puis entre 100 et 300 ms, à la fois en montage monopolaire (foyer négatif sur T7-8) et en montage bipolaire (inversion de polarité entre T7 et T9). Sur les deux électrodes la ressemblance entre le profil spatiotemporel de la violation du modèle additif et la réponse visuelle est évidente (seulement sur les contacts les plus latéraux pour l'électrode T). On observe de plus quelques foyers qui ne peuvent s'expliquer par l'activation visuelle : sur les contacts H'12-15, entre 50 et 100 ms la violation a le même profil spatial que la réponse auditive transitoire à la même latence, mais uniquement en montage monopolaire. Cette modulation est visible sur les courbes du contact H'12. De même le foyer positif sur T7-8 entre 60 et 120 ms (montage monopolaire) correspond à la fois à la réponse auditive transitoire et à la réponse visuelle, mais son amplitude ne peut s'expliquer uniquement par l'activation visuelle. Les zones entourées en orange et rouge correspondent respectivement aux seuils $p < 0,005$ et $p < 0,01$.

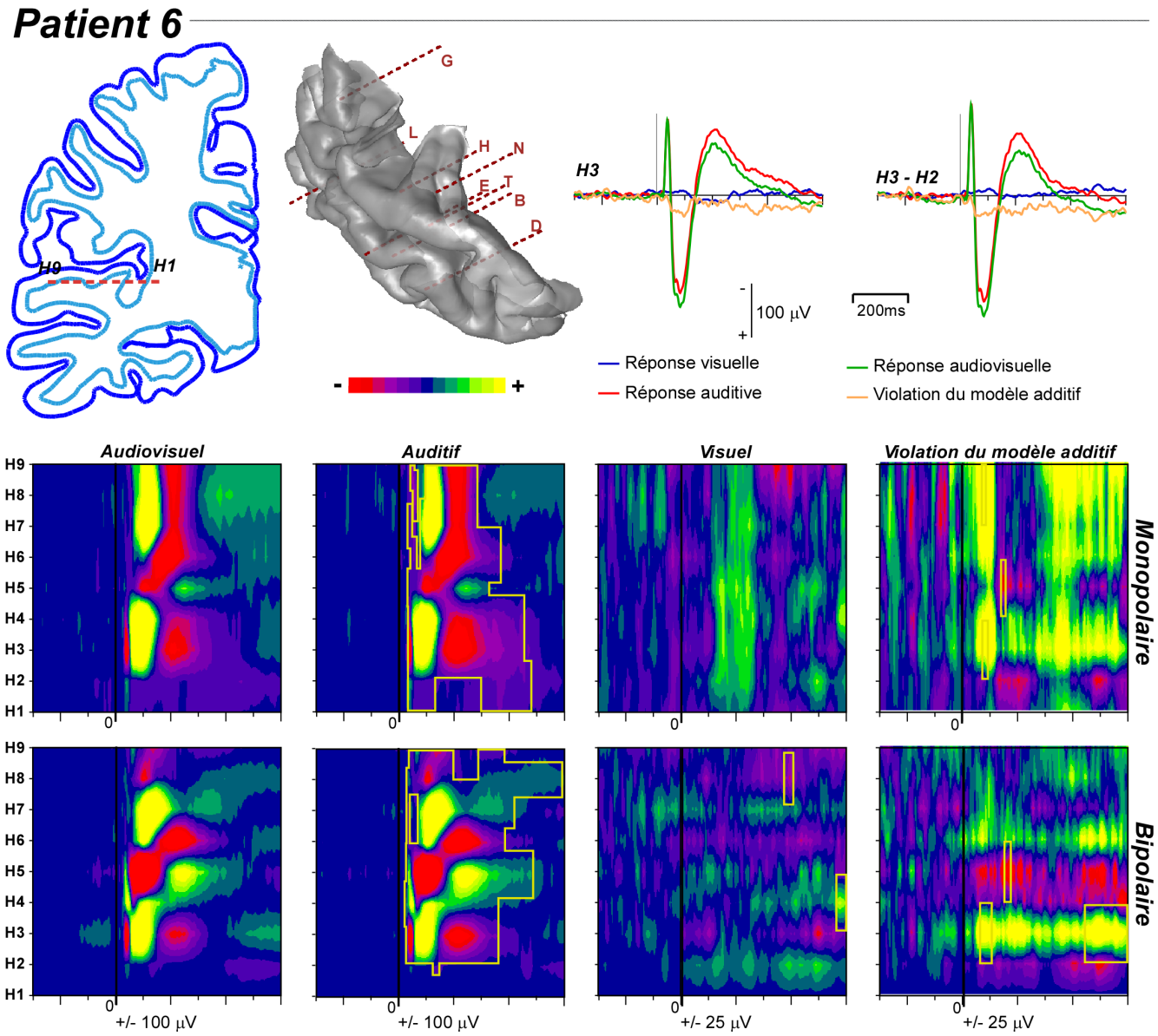

FIG. A.3 – Localisation et activités enregistrées à l'électrode H (hémisphère droit) pour le patient 6. Les premières réponse auditives sur les contacts H3-4 apparaissent dès 23 ms et sont enregistrées en montage monopolaire et bipolaire. La réponse visuelle émerge peu du bruit et ne devient significative que tardivement. On devine cependant l'existence de réponses soutenues dont le profil spatial évoque celui des réponses auditive transitoires, y compris aux niveau des contacts H3-4 où étaient enregistrées des réponses auditives primaires. De même le profil spatiotemporel de la violation ressemble à celui de la réponse visuelle, avec une amplitude plus importante. Le début de la violation de l'additivité sur les contacts H3-4 (entre 40 et 120 ms) pourrait également provenir de la modulation de la réponse transitoire auditive. Mais contrairement aux autres patients, il s'agit ici d'une augmentation de la réponse auditive en condition audiovisuelle.

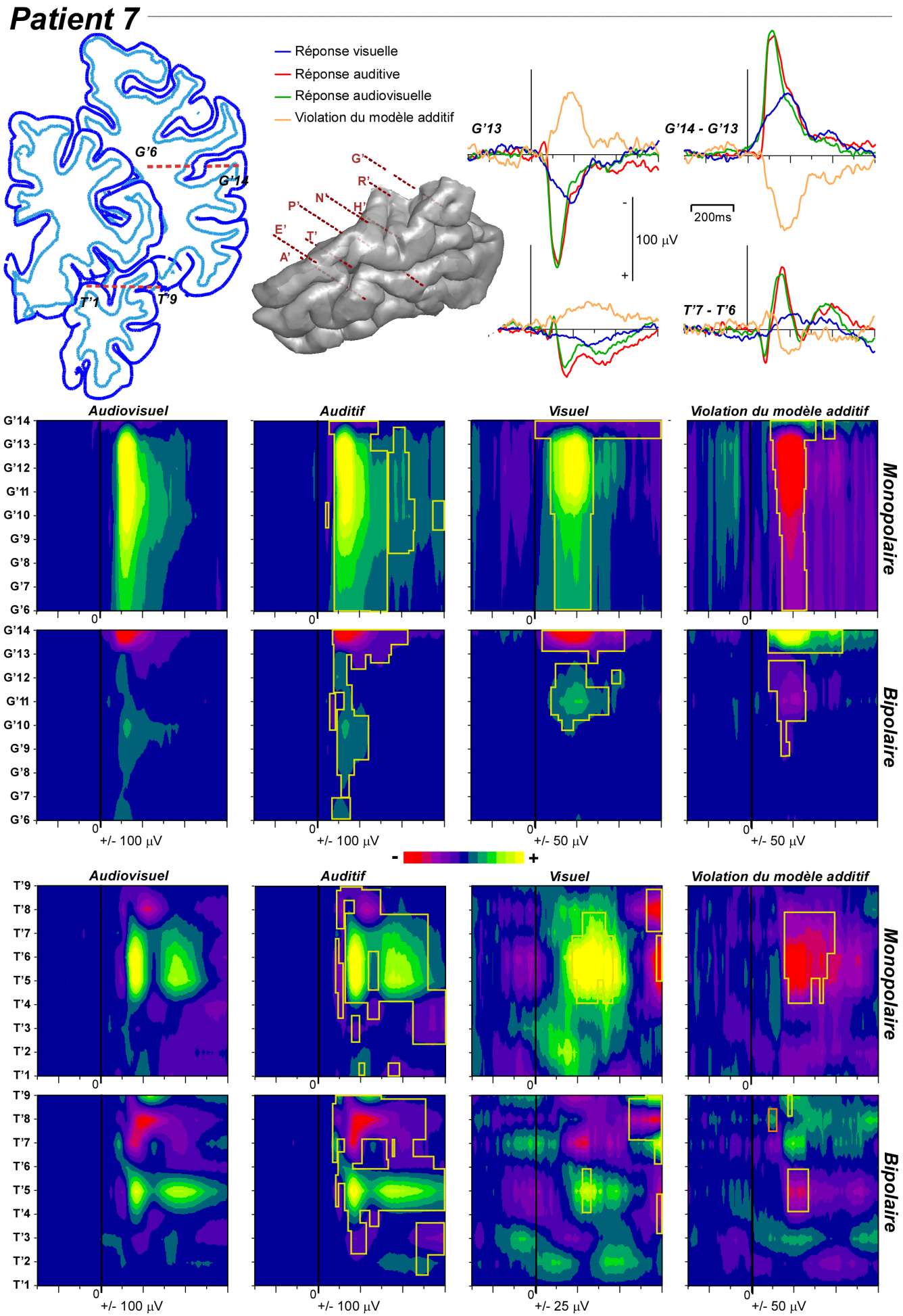

FIG. A.4 – Localisation et activités enregistrées aux électrodes G' et T' (hémisphère gauche) pour le patient 7.

FIG A.4 (page précédente) - Sur l'électrode T, la réponse soutenue commençant vers 100 ms et terminant vers 400 ms sur les contacts T'5-7 a le même profil spatial que la réponse auditive transitoire entre 120 et 200 ms (aussi bien en montage bipolaire que monopolaire). Sur l'électrode G', il existe également une certaine ressemblance entre les réponses visuelles et auditive, notamment en montage bipolaire au niveau du contact G'14. Sur les deux électrodes, la ressemblance entre le profil spatio-temporel de la violation de l'additivité et celui de la réponse visuelle est évidente. De plus sur les contacts T'7-8 entre 60 et 100 ms et T'5-7 entre 120 et 200 ms, la violation a le même profil spatio-temporel que les deux réponses transitoires auditives aux même latences, comme on peut le voir sur la courbe du montage bipolaire T'7-T'6. À ces latences l'amplitude de la réponse visuelle ne suffit pas à expliquer la violation, ce qui suggère l'existence d'une diminution de ces deux réponses auditives en condition audiovisuelle. Les zones entourées en orange correspondent au seuil $p<0{,}05$.

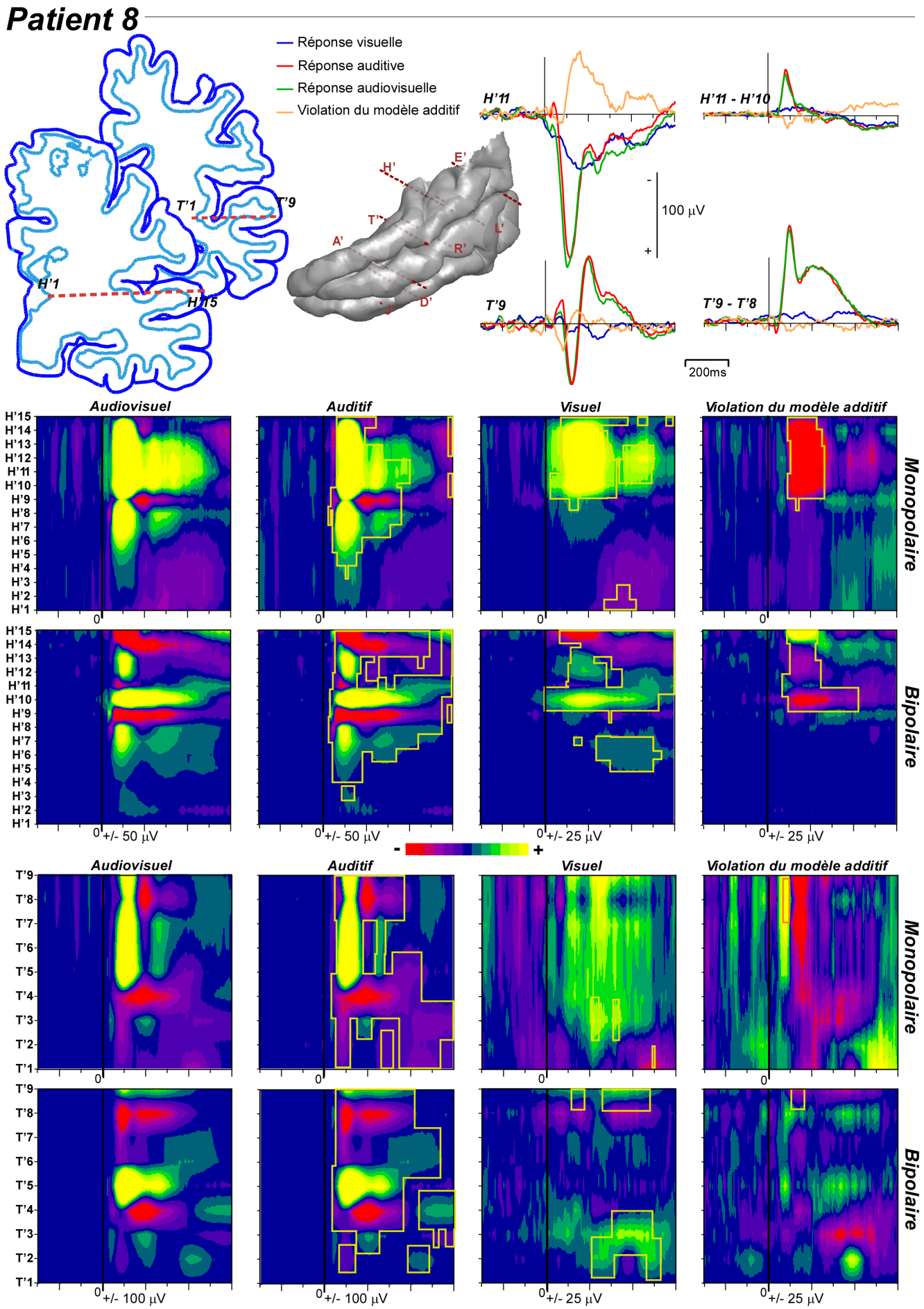

FIG. A.5 – Localisation et activités enregistrées pour les électrodes H' et T' (hémisphère gauche) pour le patient 8.

FIG A.5 (page précédente) - Les premières réponse auditive significatives apparaissent sur les contacts H'8-9 à partir de 25 ms en montage monopolaire et bipolaire. Sur l'électrode H', la réponse visuelle soutenue commençant à 0 et se terminant à 550 ms sur les contacts H'10-15 présente le même profil spatial que la réponse transitoire entre 50 et 150 ms, qui évolue elle-même en réponse soutenu ressemblant beaucoup à la réponse visuelle. Sur l'électrode T', la réponse visuelle soutenue enregistrée en montage bipolaire sur les contacts T'8-9 a également le même profil spatiale que la réponse transitoire/soutenue observée en condition auditive sur le mêmes contacts entre 50 et 400 ms. La réponse visuelle générée dans le cortex auditif est donc pour ce patient enregistrée sur des contacts différents de la réponse auditive primaire. Sur les deux électrodes H' et T', aux mêmes contacts que la réponse visuelle, le profil spatio-temporal de la violation de l'additivité ressemble de manière évidente à celle de la réponse visuelle. En montage bipolaire, un foyer positif au niveau du contact H'11 entre 60 et 120 ms n'est pas présent en condition visuelle mais correspond à la modulation de la réponse transitoire auditive, comme on peut le voir sur la courbe de l'activité bipolaire H'11-H'10. Sur le contact H'13, en montage bipolaire, on note également que la violation semble commencer à une latence inférieure à celle de la réponse visuelle. Cette violation pourrait être due à la diminution de la composante auditive transitoire enregistrée à ce contact entre 80 et 150 ms. De la même façon, la violation positive visible en monopolaire entre 50 et 100 ms sur les contacts T'5-9 suggère l'exitence d'une diminution de la composante négative transitoire auditive entre 50 et 100 ms, comme on peut le voir sur la courbe de l'activité monopolaire au contact T'9. Les zones entourées en orange correspondent au seuil $p<0{,}005$.

Patient 10

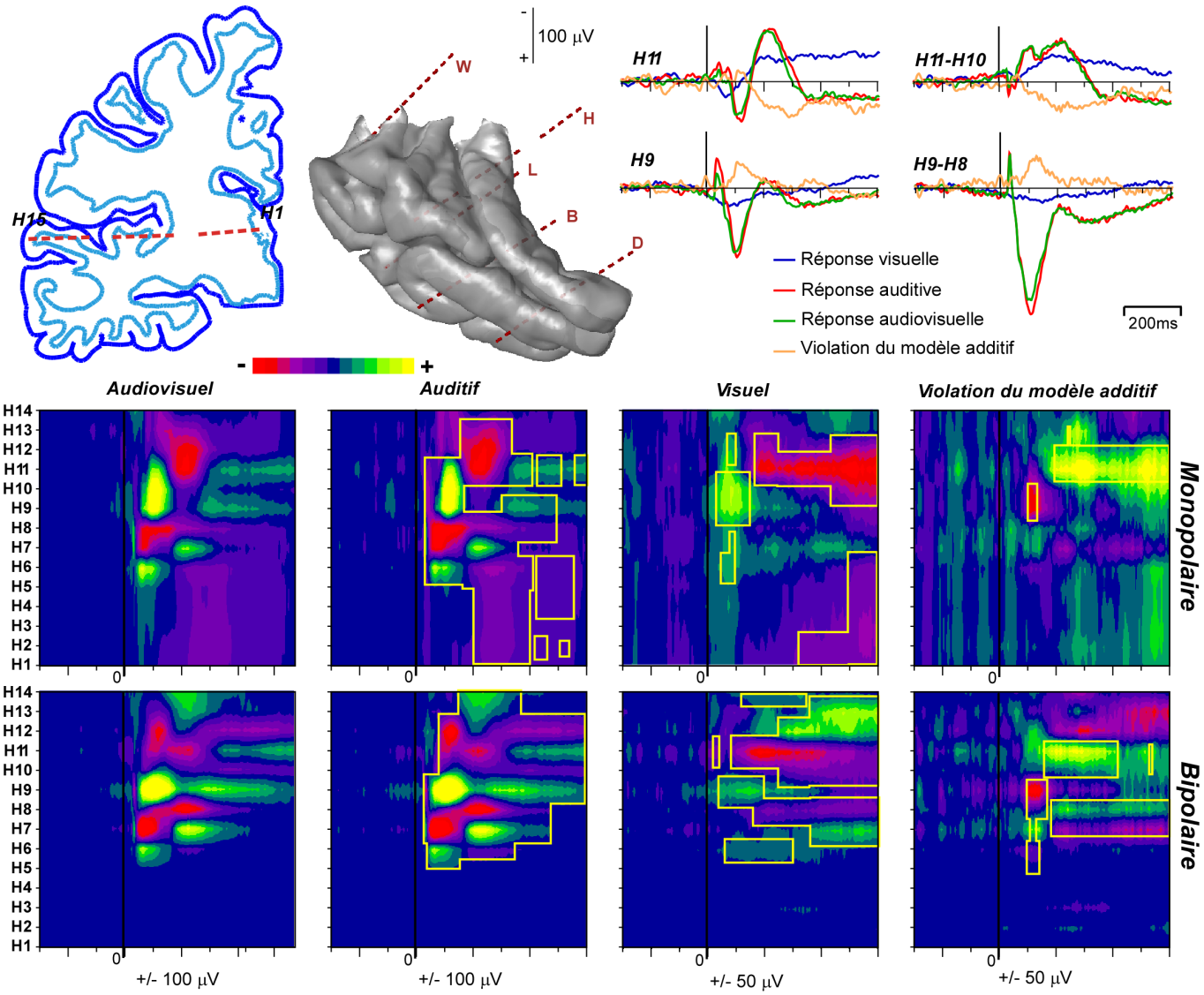

FIG. A.6 – Localisation et activités enregistrées pour l'électrode H (hémisphère droit) pour le patient 10. Les premières réponses auditives apparaissent à partir de 17 ms sur H6, H8 et H10. Les réponses visuelles sont constituées en montage monopolaire d'un réponse transitoire centrée sur les contacts H9-10 dont le profil spatiotemporel correspond à la réponse auditive transitoire entre 80 et 180 ms. En montage bipolaire apparaissent surtout des réponses soutenues dont le profil spatial ressemble à celui de la réponse auditive transitoire entre 50 et 150 ms sur H7-9, mais pas sur les électrodes plus latérales. Le profil spatiotemporel de la violation du modèle additif montrait une ressemblance certaine avec celui de la réponse visuelle aux mouvements articulatoires, excepté sur les contacts H6-8 entre 80 et 160 ms où le profil spatiotemporel était identique à celui de la réponse auditive transitoire et semble refléter une diminution de cette composante en condition audiovisuelle. Notons que pour ce patient la modulation de la réponse auditive transitoire et la réponse visuelle du cortex auditif semblent avoir lieu au niveau du cortex primaire.

Annexe B

Articles

Cogn Process (2004) 5: 189 192
DOI 10.1007/s10339-004-0026-y

LETTER TO THE EDITOR

Julien Besle · Alexandra Fort · Marie-Helene Giard

Interest and validity of the additive model in electrophysiological studies of multisensory interactions

Received: 10 June 2004 / Revised: 21 June 2004 / Accepted: 23 June 2004 / Published online: 22 July 2004
© Marta Olivetti Belardinelli and Springer-Verlag 2004

Over the past decades, the Stein group has provided a fundamental neural model of multisensory integration at the single-neuron level in animals. They have shown in cat and monkey that when inputs from different modalities are presented in close temporal and spatial proximity, multisensory neurons in the superior colliculus (SC) can increase their firing rate to a level exceeding that predicted by summing the responses to each unimodal cue (review in Stein and Meredith 1993).

Although this supra-additive effect applies to the single neuron, it has inspired a wider model that has been used at the integrated level of cortical populations (brain sites) in various functional brain imaging (ERP, MEG, fMRI) studies of multisensory integration. The rationale is that, under certain conditions that will be described below, neural activities induced by a bimodal stimulus (e.g., audiovisual, AV) should be equal to the sum of the responses generated separately by the two unisensory stimuli (e.g., auditory, A, and visual, V), if the two dimensions of the stimulus were to be independently processed. Hence, any neural activity departing from the mere summation of unimodal activities should be attributed to the bimodal nature of the stimulation, that is to interactions between the inputs from the two modalities. Using this model, it is therefore possible to estimate the crossmodal interactions in the differences between the brain responses to bimodal stimuli and the algebraic sum of the unimodal responses.

$$\text{AV Interactions} = \text{Response to (AV)} - [\text{Response to (A)} + \text{Response to (V)}]$$

Edited by: Marie-Hélène Giard and Mark Wallace

J. Besle · A. Fort · M.-H. Giard (✉)
Mental Process and Brain Activation Lab,
U280 INSERM, 151 Cours Albert Thomas,
69003 Lyon, France
E-mail: giard@lyon.inserm.fr

Note that these interactions may include modulations of unimodal responses as well as new activities in sensory or polysensory areas.

This procedure, first used by Berman (1961) in event-related corticograms of cat, was later more formally expounded by Barth et al. (1995) in a study in which they identified the brain regions that responded (evoked potentials) uniquely to bimodal AV stimuli in rat cortex: "The model assumes that if subpopulations of cells that respond separately to auditory and visual stimulation do not respond uniquely to multisensory stimuli, their contribution to the [AV-ERP] will be the linear sum of their contributions to the [A-ERP] and [V-ERP] respectively. This assumption is valid for extracellular volume conducted potentials in a [sic] purely resistive extracellular media, and is based on the law of superposition of electrical fields. The sum [A-ERP] + [V-ERP] was then subtracted from the actual [AV-ERP] to obtain a difference waveform complex [AV – (A + V)]. The [AV – (A + V)] complex was used to determine cortical regions that were uniquely activated by polysensory stimulation." (Barth et al. 1995, p 179)

Although this model theoretically can be applied to any measure of human brain activity, it has been used mainly in electrophysiological data (scalp ERP and magneto-encephalography, MEG: Miniussi et al. 1998; Giard and Peronnet 1999; Foxe et al. 2000; Raij et al. 2000; Fort et al. 2002a, b; Molholm et al. 2002; Klucharev et al. 2003; Möttönen et al. 2004). On the other hand, its use has been recurrently criticized (Teder-Sälejärvi et al. 2002; Calvert and Thesen 2004) because of the multiple biases it can generate in the estimation of the crossmodal interactions if several important conditions are not fulfilled. We discuss in this note what these biases are and how to avoid or minimize them, with particular emphasis on electromagnetic (EEG/MEG) recordings. Finally, we explain why, in spite of its strict conditions of application, the supra-additive model is particularly interesting in ERP/MEG studies of multisensory integration.

Potential biases and artifacts generated by the additive model

1. The additive model is valid only when the brain responses that are analyzed do not include activity common to all conditions. Indeed these activities would be added only once but subtracted twice in the [AV − (A + V)] model, which would confound the derivation of the multisensory interaction. "Common activity" may be of several types. One type is neural responses related to late semantic processes, target processing (e.g., N2b/P3 waves in ERP/MEG recordings), response selection, or motor processes. ERP literature has shown that these activities usually arise about 200 ms post-stimulus, whereas earlier latencies are characterized by sensory-specific responses (review in Hillyard et al. 1998). One way to avoid this problem is to restrict the analysis period to the early time frame (< 200 ms) of stimulus processing. While this procedure is very simple in ERP/MEG recordings since their time resolution is of the order of the millisecond, sorting the response components according to their latency is still virtually impossible in hemodynamic imaging techniques. Second, in paradigms requiring speeded responses with rapidly presented stimuli, "anticipatory" slow responses may arise before each (unimodal and bimodal) stimulus and continue for a time after stimulus onset. These anticipatory responses appear, when present in ERP/MEG recordings, as slow ramp-like deflections in the prestimulus and early poststimulus periods. These deflections can thus give rise to spurious residual effects in the [AV − (A + V)] signals that may be confused with early cross-modal interactions (Teder-Sälejärvi et al. 2002). Note that such anticipatory processes are independent of the technique used and can also be included in fMRI/PET responses. At the level of the experimental design, a procedure that may be applied to avoid or strongly reduce anticipatory processes, whatever the neuroimaging technique, is to present the stimuli at random interstimulus time intervals during data acquisition. In ERP/MEG signal analysis, two further methods have been proposed to control for these effects: modify the latency of the prestimulus period that will be used as the reference baseline, and/or high-pass the data (e.g., 2 Hz cut-off frequency) to remove the slow wave effects.

2. Several functional imaging studies using block-designed paradigms have shown a decrease in activation in sensory-specific cortices (e.g., the auditory cortex) when subjects were presented with continuous stimulation in another (e.g., visual) modality (Haxby et al. 1994; Kawashima et al. 1995; Laurienti et al. 2002). There are two possibilities to explain this. First, these effects may reflect cross-sensory driving and/or inhibition of lower-order sensory areas via direct projections from one sensory cortex to another (Falchier et al. 2002; Rockland and Ojima 2003; review in Schroeder et al. 2004). There is, however, no experimental evidence that such "cross-modal effects" in unimodal conditions may be seen at the integrated level of scalp ERP/MEG or fMRI signals irrespective of the task or stimulus delivery context. In addition, even in this case, the additive model should still apply since any difference in these processes between a unimodal and a bimodal condition should appear—if strong enough—as low-level cross-modal interactions in the model, and further represent one possible neural mechanism for multisensory integration. A second, more likely explanation is that when a particular sensory cortex is continuously and exclusively activated during a whole block, while the other non-matching cortices are not activated, the attentional resources are dedicated to the relevant modality (even in passive tasks or tasks that demand little attention), while the other modalities are more or less voluntarily ignored (deactivated) to optimize the processing in the relevant sensory cortex (see also Ghatan et al. 1998; Kawashima et al. 1999, for similar attentional effects). In studies of multisensory integration, the [AV − (A + V)] model should therefore not be used in experiments based on block-designed paradigms, since these unimodal deactivations would be subtracted from the bimodal activations, resulting in artificial increases of the "crossmodal" effects. One way to eliminate or considerably reduce such attention-related deactivations in unisensory cortices is to consider paradigms in which the stimuli are randomly and equiprobably delivered across all modality conditions (e.g., Giard and Peronnet 1999; Calvert et al. 2000, 2001; Foxe et al. 2000; Raij et al. 2000; Fort et al. 2002a, b; Molholm et al. 2002; Wright et al. 2003).

3. Random mixing of conditions, however, may not be sufficient for a correct control of attention. Although a classical design to avoid attentional biases in the additive model is to require the same task in the three modalities, in some paradigms, the task may be easier and require less effort in one unimodal condition than in the other. This problem can be overcome by equating the levels of difficulty across unimodal conditions (by equating the behavioral performance in both unimodal conditions, e.g., Giard and Peronnet 1999). However, in some particular cases, this may not be possible and using the same task across all the conditions can lead to noticeable spurious effects in the computation of interactions. Consider, for example, speech stimuli (lip movements associated with syllable sounds) randomly presented in the three A, V and AV conditions: if a discrimination task (e.g., respond to target syllables) is required under the three modality conditions, the processing of syllables in the lip-reading condition alone will include an important visual attention effect that will not be eliminated in the AV − (A + V) derivation, since in speech perception (unlike what is likely to occur for bimodal non-speech objects), normal subjects will *naturally* engage much less visual attention to process AV than V stimuli. Alternatively, if the subjects are required to respond only whenever they hear (A and AV conditions) a target syllable, their (selective) auditory attention effect

will be expressed rather similarly for A and AV stimuli and eliminated in the [AV – (A + V)] model; in the same way, a lesser (if any) effect of visual attention (rather similar for V and AV stimuli) should be mostly eliminated in the model. A general principle, therefore, in dealing with attentional problems is, in addition to systematically mixing conditions, to equate the attentional load between each unimodal condition and the bimodal condition (but not necessarily between the two unimodal conditions).

Advantages of the additive model in ERP/MEG studies of cross-modal interactions

All the examples above show that non-biased estimation of multisensory interactions in the human cortex using the additive model requires taking important precautions both in the experimental design and in data analysis. While the constraints relative to the control of attention may be easily respected whatever the neuroimaging technique used, caveats concerning the temporal selection of the response components to be analyzed can be overcome only in EEG/MEG approaches, because of the excellent time information provided by these techniques.

In addition, the additive model has a further fundamental interest in ERP/MEG analysis of crossmodal interactions. Indeed, unlike what is observed at the voxel level in fMRI or PET signals, a significant value at a particular electrode (sensor) in ERP/MEG recordings does not mean that the structure beneath the electrode/sensor is active. Rather what is recorded at the scalp surface results from the diffusion of electrical currents inside the brain originating from *distant* "generators," and the interpretation of the surface signals needs to take into account these volume conduction factors (using topographic analysis, generator modeling, etc.). Interestingly, the additive [AV – (A + V)] model in ERP/MEG has the fundamental property of avoiding the problem of overlaps of volume conduction effects in the different subcomponents of the bimodal response by removing the conduction effects of the corresponding unimodal responses. In this respect, the additive model is not a mere application of the single-cell model used by Stein's group and other authors: it applies not only at the *local* structure level (single cell, voxel), but also at the *distant* electrode/sensor level (volume conduction effects) because it is based on the superposition principle of electrical fields, in which the potentials from separate current sources in a conductive medium sum linearly. If its conditions of application are fulfilled, the additive model will therefore isolate the (volume conduction) effects specifically related to the interactions (which will have to be analyzed in turn in terms of topography and generators).

We therefore believe that the additive model is particularly well suited to ERP/MEG study of multisensory interactions in humans, and that its multiple advantages make it worthwhile dealing with the several constraints it imposes. Provided that its conditions of application are respected, the model can reveal the existence of genuine cross-modal interactions without making a priori assumptions about the congruent/incongruent character of the bimodal inputs, or introducing supra-additive/sub-additive criteria for integration (e.g., Calvert 2001; Calvert et al. 2001). Rather the additive model allows one to access the dynamics of the multisensory interactions and observe both supra-additive and sub-additive modulations of unimodal activities in sensory-specific cortices—which appear to form a highly flexible network of cross-modal operations—as well as to observe new processes specifically activated by the bimodal nature of the stimulus.

References

Barth DS, Goldberg N, Brett B, Di S (1995) The spatiotemporal organization of auditory, visual and auditory visual evoked potentials in rat cortex. Brain Res 678:177 190

Berman AL (1961) Interaction of cortical responses to somatic and auditory stimuli in anterior ectosylvian gyrus of cat. J Neurophysiol 24:608 620

Calvert GA (2001) Crossmodal processing in the human brain: insights from functional neuroimaging studies. Cereb Cortex 11:1110 1123

Calvert GA, Thesen T (2004) Multisensory integration: methodological approaches and emerging principles in the human brain. J Physiol Par (in press)

Calvert GA, Campbell R, Brammer MJ (2000) Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. Curr Biol 10:649 657

Calvert GA, Hansen PC, Iversen SD, Brammer MJ (2001) Detection of audio visual integration sites in humans by application of electrophysiological criteria to the bold effect. Neuroimage 14:427 438

Falchier A, Clavagnier S, Barone P, Kennedy H (2002) Anatomical evidence of multimodal integration in primate striate cortex. J Neurosci 22(13):5749 5759

Fort A, Delpuech C, Pernier J, Giard MH (2002a) Dynamics of cortico-subcortical crossmodal operations involved in audio visual object detection in humans. Cereb Cortex 12(10):1031 1039

Fort A, Delpuech C, Pernier J, Giard MH (2002b) Early auditory visual interactions in human cortex during nonredundant target identification. Cogn Brain Res14:20 30

Foxe JJ, Morocz IA, Murray MM, Higgins BA, Javitt DC, Schroeder CE (2000) Multisensory auditory somatosensory interactions in early cortical processing revealed by high-density electrical mapping. Cogn Brain Res 10:77 83

Ghatan PH, Hsieh JC, Petersson KM, Stone-Elander S, Ingvar M (1998) Coexistence of attention-based facilitation and inhibition in the human cortex. Neuroimage 7:23 29

Giard MH, Peronnet F (1999) Auditory visual integration during multimodal object recognition in humans: a behavioral and electrophysiological study. J Cogn Neurosci 11(5):473 490

Haxby JV, Horwitz B, Ungerleider LG, Maisog JM, Pietrini P, Grady CL (1994) The functional organization of human extrastriate cortex: a pet-rcbf study of selective attention to faces and locations. J Neurosci 14(11):6336 6353

Hillyard SA, Teder-Sälejärvi WA, Munte TF (1998) Temporal dynamics of early perceptual processing. Curr Opin Neurobiol 8:202 210

Kawashima R, O'Sullivan BT, Roland PE (1995) Positron-emission tomography studies of cross-modality inhibition in selective attentional tasks: closing the "mind's eye". Proc Natl Acad Sci USA 92:5969 5972

Kawashima R, Imaizumi S, Mori K, Okada K, Goto R, Kiritani S et al (1999) Selective visual and auditory attention toward utterances-a pet study. Neuroimage 10:209 215

Klucharev V, Mottonen R, Sams M (2003) Electrophysiological indicators of phonetic and non-phonetic multisensory interactions during audiovisual speech perception. Cogn Brain Res 18(1):65 75

Laurienti PJ, Burdette JH, Wallace MT, Yen YF, Field AS, Stein BE (2002) Deactivation of sensory-specific cortex by crossmodal stimuli. J Cogn Neurosci 14(3):420 429

Miniussi C, Girelli M, Marzi CA (1998) Neural site of the redundant target effect: electrophysiological evidence. J Cogn Neurosci 10:216 230

Molholm S, Ritter W, Murray MM, Javitt DC, Schroeder CE, Foxe JJ (2002) Multisensory auditory visual interactions during early sensory processing in humans: a high-density electrical mapping study. Cogn Brain Res 14(1):115 128

Möttönen R, Schurmann M, Sams M (2004) Time course of multisensory interactions during audiovisual speech perception in humans: a magnetoencephalographic study. Neurosci Lett 363(2):112 115

Raij T, Uutela K, Hari R (2000) Audiovisual integration of letters in the human brain. Neuron 28(2):617 625

Rockland KS, Ojima H (2003) Multisensory convergence in calcarine visual areas in macaque monkey. Int J Psychophysiol 50(1 2):19 26

Schroeder C, Molholm S, Lakatos P, Ritter W, Foxe JJ (2004) Human-simian correspondence in the early cortical processing of multisensory cues. Cogn Process DOI 10.1007/s10339-004-0020-4

Stein BE, Meredith MA (1993) The merging of the senses. MIT Press, Cambridge

Teder-Sälejärvi WA, McDonald JJ, Di Russo F, Hillyard SA (2002) An analysis of audio visual crossmodal integration by means of event-related potential (erp) recordings. Cogn Brain Res 14(1):106 114

Wright TM, Pelphrey KA, Allison T, McKeown MJ, McCarthy G (2003) Polysensory interactions along lateral temporal regions evoked by audiovisual speech. Cereb Cortex 13(10):1034 1043

European Journal of Neuroscience, Vol. 20, pp. 2225 2234, 2004 © Federation of European Neuroscience Societies

Bimodal speech: early suppressive visual effects in human auditory cortex

Julien Besle, Alexandra Fort, Claude Delpuech and Marie-Hélène Giard
INSERM U280, Mental Processes and Brain Activation, 151 Cours Albert Thomas, 69424 Lyon Cedex 03, France

Keywords: audiovisual, electrophysiology, multisensory integration, speech perception,

Abstract

While everyone has experienced that seeing lip movements may improve speech perception, little is known about the neural mechanisms by which audiovisual speech information is combined. Event-related potentials (ERPs) were recorded while subjects performed an auditory recognition task among four different natural syllables randomly presented in the auditory (A), visual (V) or congruent bimodal (AV) condition. We found that: (i) bimodal syllables were identified more rapidly than auditory alone stimuli; (ii) this behavioural facilitation was associated with cross-modal [AV − (A + V)] ERP effects around 120–190 ms latency, expressed mainly as a decrease of unimodal N1 generator activities in the auditory cortex. This finding provides evidence for suppressive, speech-specific audiovisual integration mechanisms, which are likely to be related to the dominance of the auditory modality for speech perception. Furthermore, the latency of the effect indicates that integration operates at pre-representational stages of stimulus analysis, probably via feedback projections from visual and/or polymodal areas.

Introduction

It is commonly known and agreed that vision may improve the comprehension of a talker in a face-to-face conversation or on the television. In behavioural studies, the influence of visual information on auditory speech perception has been particularly explored in the 'McGurk effect' (McGurk & McDonald, 1976), an auditory illusion produced for particular syllables when the lip movements do not match the auditory signal (for example, auditory /ba/ combined with visual /ga/ is perceived as /da/).

Yet, the neural mechanisms by which auditory and visual speech information is combined in normal communication are still poorly understood. Several functional neuroimaging studies have identified possible sites of multisensory convergence and integration for linguistic material with various results. Haemodynamic responses to semantically congruent audiovisual speech stimuli were found to be enhanced in sensory specific auditory and visual cortices, compared to the responses to unimodal or incongruent bimodal inputs (Calvert *et al.*, 1999). However, when only the brain areas presenting supra-additive response enhancement to congruent bimodal inputs and subadditive response to incongruent cues were considered as integration sites, only the left superior temporal sulcus (STS) exhibited significant integration effects. In another functional magnetic resonance imaging experiment, a supra-additive enhancement was found only in the left claustrum/insula whereas activation of the STS occurred for lip-reading alone (Olson *et al.*, 2002; see also Calvert & Campbell, 2003). Whatever the precise sites of multisensory integration, Calvert (2001) hypothesized that increased activity in sensory-specific cortices would be a result of backward projections from polymodal areas such as the STS.

However, this assumption is beyond the reach of haemodynamic imaging techniques because of their poor temporal resolution. In contrast, neuromagnetic (MEG) and event-related potential (ERP) recordings can provide significant insights into the timing of bimodal speech integration.

In three studies using audiovisual oddball paradigms (Sams *et al.*, 1991; Colin *et al.*, 2002; Möttönen *et al.*, 2002), deviant 'McGurk syllables' differing from standard syllables only on the visual dimension were found to elicit a mismatch negativity (MMN) around 150 180 ms post-stimulus, an ERP/MEG component generated for the main part in the auditory cortex. As MMN probably reflects a neuronal mismatch between deviant auditory inputs and a neural representation of the past stimuli in auditory sensory memory (review in Näätänen & Winkler, 1999), it can be concluded from these previous studies that visual speech information has been integrated to the auditory input before the MMN process was triggered, that is before about 150 ms. This McGurk paradigm, however, only put an indirect upper bound on the timing of multisensory integration and the question remains open as to when and where in the sensory processing chain, and by which neural mechanisms auditory-visual speech is combined.

One way to investigate these questions is to compare the electrophysiological responses to bimodal sensory inputs with the sum of the responses to unimodal cues presented separately. This approach was used in humans to analyse the mechanisms of audiovisual integration in bimodal object recognition (Giard & Peronnet, 1999) and revealed the existence of multiple interactions within the first 200 ms post-stimulation, expressed both as modulations (increase and decrease) and as new activations in sensory-specific and polymodal brain areas. Subsequent experiments using this additive model have provided evidence for different integrative operations according to the stimulus type, the modalities involved, or the task required (Foxe *et al.*, 2000; Raij *et al.*, 2000; Fort *et al.*, 2002a,b; Molholm *et al.*, 2002). In the present study we therefore used the same approach to investigate the time-course and neural mechanisms of audiovisual integration in the particular case of speech perception.

Correspondence: Dr M.-H. Giard, as above.
Email: giard@lyon.inserm.fr

Received 14 November 2003, revised 27 July 2004, accepted 3 August 2004

doi:10.1111/j.1460-9568.2004.03670.x

Materials and methods

Subjects

Sixteen right-handed native French speakers (mean age 23.0; eight females) were paid to participate in the study, for which they gave a written informed consent in accordance with the Code of Ethics of the World Medical Association (Declaration of Helsinki). All subjects were free from neurological disease, had normal hearing and normal or corrected-to-normal vision.

Thirteen other subjects (mean age 24.3; nine females) participated in an additional behavioural-only experiment.

Stimuli

ERP study of multisensory integration requires to strictly control the timing of the unimodal input signals, a particularly heavy constraint in the case of natural speech. We therefore proceeded in the following way:

1. A hundred utterances of four different audiovisual syllables (/pa/, /pi/, /po/ and /py/) were produced by a female French speaker and recorded with a DV camera at a video sampling rate of 25 fps and an audio sampling rate of 44.1 kHz.
2. Visual inspection of the video stream showed that for most utterances, six frames (240 ms) separated the first detectable lip movements from the opening of the mouth (corresponding roughly with the beginning of the speech sound). To have stimuli with similar auditory-visual structures, we selected a subset of these syllables. The sound onset was then strictly postsynchronized with the onset of the 7th frame. This point (240 ms after the beginning of lip movements) was taken as time zero for ERP averaging and latency measurements (see Fig. 1B). The voice onset times (the intervals between the consonant burst and the voicing corresponding to the vowel), originally ranging from 15 to 26 ms, were artificially shortened to 15 ms for all the stimuli.
3. Using a unique exemplar of each syllable (/pa/, /pi/, /po/ or /py/) could have led subjects to learn and recognize the stimuli on the basis of low-level sensory features specific to each stimulus but irrelevant for phonetic processing. We therefore selected three exemplars of each syllable, that is 12 different utterances.
4. Eventually, lip movements preceding the sound emission anticipate the shape that will produce the vowel (coarticulation) and can therefore slightly differ between the different syllables. Although the prevowel lip movements were very faint during the first six frames of the video stream, we ensured that they could not allow the subjects to deduce the identity of the syllable before the sound onset (7th frame): we asked seven subjects (who did not participate in the main experiment) to visually identify the syllables on the basis of the first 6, 8 or 13 frames. Results showed that subjects did respond at chance level in the 6-frame condition.

All the images of the video stream were cropped in order to keep only the mouth, the cheeks and the bottom of the nose (see Fig. 1B). In the final frames, the mouth was about 5 cm wide and was presented on a video monitor placed 130 cm in front of the subjects' eyes, subtending a visual angle of 2.2°. The duration of the 12 sounds corresponding to the 12 syllables ranged from 141 to 210 ms; their amplitudes were adjusted to have the same perceived intensity (kept constant for all subjects).

Procedure

Subjects were seated in a dark, sound-attenuating room and were given instructions describing the task along with a practice block of 70 trials (a trial is described in Fig. 1A). Then subjects were presented with 31 repetitions of the 12 syllables in each of the three following conditions: auditory-only (A), visual-only (V) and audiovisual (AV). These 1116 trials were divided into 16 blocks (block duration, about 2 min 35 s; mean ISI, 2210 ms). In all blocks, trials were delivered pseudorandomly with the constraint that two stimuli of the same condition could not occur in a row.

At the beginning of each block, one of the four syllables (/pa/, /pi/, /po/ or /py/) was designated as the target (so that each syllable could be target or nontarget depending on the block). The subjects' task was to press a mouse-button with the right forefinger whenever they heard (A and AV conditions) the target-syllable in the block sequence.

An auditory task alone was chosen because estimation of the crossmodal interactions using the additive [AV − (A + V)] model (see data analysis) requires that the attention level in each modality is similar between the unimodal and the bimodal conditions (but not necessarily between the two unimodal conditions). Indeed, as the subjects are instructed to make an auditory discrimination task, the auditory attention effect will be expressed rather similarly in A and AV

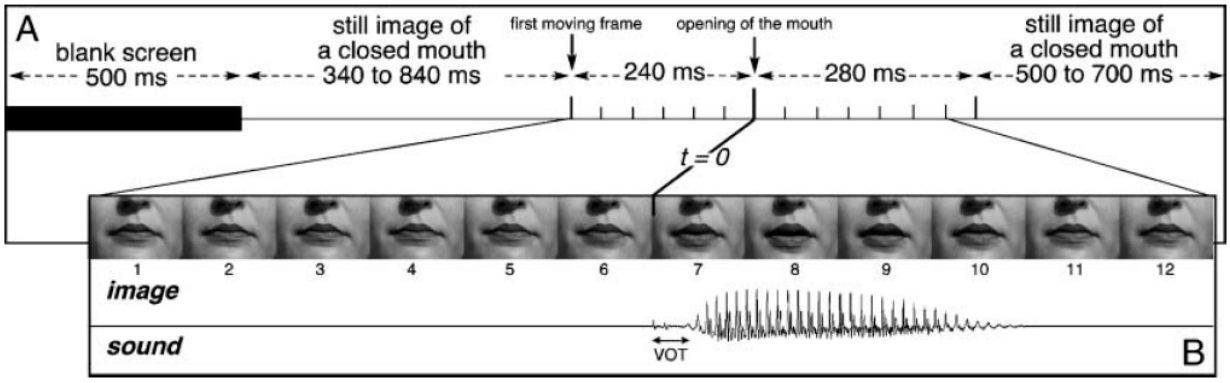

FIG. 1. Time-course of an auditory-visual non-target trial. Each trial began with the presentation of a blank screen for 500 ms; then a still image of a closed mouth was displayed during a random period of 340 840 ms. The mouth began to move for 240 ms (six frames) before opening (time zero). Then, the corresponding sound was played. The lip movement ended 280 ms after time zero with an image of the closed mouth that remained for a random time of 500 700 ms for nontarget trials, and until the key press for target trials (or for 1500 ms if the subject did not respond). In the visual-only condition, the time course was similar except that the sound was not played. In the auditory-only condition, the mouth remained closed all along the trial. VOT, voice onset time.

brain responses and mostly eliminated in [AV − (A + V)]. By contrast, because lip-reading is unnatural and difficult for untrained, normal-hearing subjects, a task in the three A, V and AV conditions would have led the subjects to naturally engage much more visual attention to process visual than bimodal stimuli. As a consequence, a larger visual attention effect in the V than in AV responses would not have been eliminated in the model. On the contrary, the task used here required a rather similar (if any) visual attention effort to process the visual and audio-visual stimuli, then minimizing any attentional bias in the additive model.

Electroencephalogram recording

Electroencephalograms (EEG) were was recorded continuously via a Neuroscan Compumedics system through Synamps DC-coupled amplifiers (0.1 200 Hz analogue band width; sampling rate, 1 kHz) from 36 Ag AgCl scalp electrodes referenced to the nose and placed according to the International 10 20 System: Fz, Cz, Pz, POz, Iz; Fp1, F3, F7, FT3, FC1, T3, C3, TP3, CP1, T7, P3, P7, PO3, O1, and their counterparts on the right hemiscalp; Ma1 and Ma2 (left and right mastoids, respectively); IMa and IMb (midway between Iz-Ma1 and Iz-Ma2, respectively). Electrode impedances were kept below 5 kΩ. Horizontal eye movements were recorded from the outer canthus of the right eye; eye blinks and vertical eye movements were measured in channels Fp1 and Fp2.

Data analysis

EEG analysis was undertaken with the ELAN Pack software developed at the INSERM U280 laboratory (Lyon, France). Trials with signal amplitudes exceeding 100 μV at any electrode from 2000 ms before time zero to 500 ms after were automatically rejected to discard the responses contaminated by eye movements or muscular activities. One subject was excluded from analysis for general noise in EEG at most sites. For seven other subjects, the excessively noisy signals at one or two electrodes were replaced by their values interpolated from the remaining electrodes.

ERPs to nontarget stimuli were averaged offline across the 12 different syllables separately for each modality (A, V, AV), over a time period of 1000 ms including 500 ms prestimulus (the zero time corresponding to the onset of the sound, or the onset of the 7th video frame for visual-only trials). Trials including false alarms were not taken into account when averaging. The mean numbers of averaged trials (by subject) were 155, 157 and 170 in the A, V and AV conditions, respectively (about 40% of the trials were discarded because of important eye movements).

ERPs were finally digitally filtered (bandwidth, 1 30 Hz; slope, 24 dB/octave). The mean amplitude over the [−300 to −200 ms] prestimulus period was taken as the baseline for all amplitude measurements.

Estimation of audiovisual interactions

We assumed that at an early stage of stimulus processing, if auditory (A) and visual (V) dimensions of the stimulus were to be independently processed, the neural activities induced by the audiovisual (AV) stimulus should be equal to the algebraic sum of the responses generated separately by the two unisensory stimuli. Hence, any neural activity departing from the mere summation of unimodal activities should be attributed to the bimodal nature of the stimulation, that is to interactions between the inputs from the two modalities (Barth *et al.*, 1995; Miniussi *et al.*, 1998; Giard & Peronnet, 1999; see Discussion in Besle *et al.*, 2004). This assumption is valid only if the period of analysis does not include nonspecific activities that would be common to all three types of stimuli, and particularly late activities related to semantic processing, response selection or motor processes. ERP litterature shows that these 'nonspecific' components generally arise after about 200 ms, whereas the earlier latencies are characterized by sensory-specific responses (e.g. Hillyard *et al.*, 1998 for a review). We have therefore restricted the analysis period to [0 200] ms and used the following summative model to estimate the AV interactions:

$$\text{ERP (AV)} = \text{ERP (A)} + \text{ERP (V)} + \text{ERP (A} \times \text{V interactions)}$$

This expression is valid whatever the nature, configuration or asynchrony of the underlying neural generators and is based on the law of superposition of electric fields. However, estimation of AV interactions using this procedure further requires that: (i) the levels of modality-specific attention are similar between each unimodal condition and the bimodal condition (see Procedure); and (ii), the effects potentially found in a particular structure cannot be attributed to deactivation processes in that structure under unimodal stimulation of a concurrent modality (see Discussion).

Significant interaction effects were assessed by Student's *t*-tests comparing the amplitudes of the [AV − (A + V)] difference waves to zero for each time sample at each electrode. Student's *t*-maps could then be displayed at each latency. Correction for multiple comparisons was performed using the procedure of Guthrie & Buchwald (1991), which tabulates the minimum number of consecutive time samples that should be significant in ERP differences, in order to have a significant effect over a given time series. As these tables are given for a horizon of 150 time samples, we under-sampled our data at 500 Hz over the [0 200 ms] analysis period, as proposed by the authors. We therefore considered as significant interactions the spatiotemporal patterns having a stable topography with significant amplitude ($P < 0.05$) during at least 12 consecutive time samples (24 ms), which is an upper bound for 15 subjects over 100 time samples (200 ms).

Topographic analysis and dipole modelling

To facilitate the interpretation of the voltage values recorded at multiple electrodes over the scalp surface, we analysed the topographic distributions of the potentials and the associated scalp current densities (SCDs). Scalp potential maps were generated using two-dimensional spherical spline interpolation and radial projection from T3 or T4 (left and right lateral views, respectively), which respects the length of the meridian arcs. SCDs were obtained by computing the second spatial derivative of the spline functions used in interpolation (Perrin *et al.*, 1987, 1989). SCDs do not depend on any assumption about the brain generators or the properties of deeper media, and they are reference free. In addition, SCDs reduce the spatial smearing of the potential fields because of the volume conduction of the different anatomical structures, and thus enhance the contribution of local intracranial sources (Pernier *et al.*, 1988).

Topographic analysis was complemented by spatiotemporal source modelling (Scherg & Von Cramon, 1985, 1986; Giard *et al.*, 1994) based on a three-concentric sphere head model for conductive volumes (brain, skull and scalp) and equivalent current dipoles (ECDs) for generators (local activity of brain regions). Data were modelled using two stationary dipoles with symmetrical positions (one in each hemisphere). The dipole parameters were determined by a

nonlinear iterative procedure (Marquardt minimization method) for the spatial parameters (location and orientation) and with a linear least-mean square algorithm for the time-varying magnitude (Scherg, 1990). The model adequacy was assessed by a goodness-of-fit criterion based on the percentage of experimental variance explained by the model. Note that the modelling procedure was not used here to localize the brain generators involved in the auditory response and/or the cross-modal interactions, but rather to test whether the dipole configuration best explaining the [AV − (A + V)] interactions could also explain most of the auditory response.

Results

Behavioural results

Subjects ($n = 16$) identified the target syllables more rapidly when presented in the audiovisual condition (mean response time 400 ms) than in the auditory-alone condition (423 ms, $F_{1,15} = 18.76$, $P < 0.001$). The error rate was less than 1% in each of the two conditions.

According to the race models (Raab, 1962), a shorter reaction time (RT) in bimodal condition (known as the redundant-stimulus effect) does not necessarily imply the existence of crossmodal interactions before the response, as the first of the two unimodal processes completed could have determined the reaction time. Miller (1982) has shown that under this last hypothesis, particular assumptions can be made on the distribution of RTs:

$$P(\mathrm{RT_{AV}} < T) = P(\mathrm{RT_A} < T) + P(\mathrm{RT_V} < T), \qquad (1)$$

for any reaction time T, where $P(\mathrm{RT} < T)$ is the cumulative probability density function (CDF) of RT.

To test this hypothesis on the speech material used in the ERP study, we performed an additional behavioural-only experiment using the same stimuli and paradigm, except that the subjects ($n = 13$) had to respond to the target syllables in the three (A, V, AV) modalities. The mean RTs to identify the auditory, visual and audiovisual stimuli were 418 ms, 496 ms and 356 ms, respectively (Fig. 2A). Following the procedure proposed by Ratcliff (1979) (see also Miller, 1982), the CDFs of RTs for each subject in each of the three conditions were divided into 19 fractiles (0.05, 0.10, …, 0.90, 0.95) and RTs were group averaged at each fractile, yielding a group distribution (Fig. 2B). Comparison of the AV-CDF with the sum of the A- and V-CDFs using Student t-tests reached statistical significance ($P < 0.001$) for the nine first fractiles (0.05 0.45) showing that inequality (1) was violated for shorter RTs.

Electrophysiological results

Figure 3 presents the ERPs elicited by nontarget unimodal and bimodal stimuli from 150 ms before time zero (onset of the auditory signal) up to 300 ms after, at a subset of electrodes (and corresponding SCDs at Cz). The unimodal A and V waveforms display morphologies typical of activities in sensory-specific areas: the auditory N1 wave was maximum at 136 ms at fronto-central sites (−8.34 μV at Cz) with a small polarity reversal at mastoid electrodes (Ma1, 0.35 μV at 111 ms; Ma2, 0.01 μV at 108 ms). This spatiotemporal configuration is known to reflect neural activity in the supra-temporal auditory cortex (Vaughan & Ritter, 1970; Scherg & Von Cramon, 1986). Auditory N1 was followed by the P2 wave peaking at 221 ms (3.97 μV at Cz) with polarity reversals at mastoids (Ma1, −2.86 μV at 205 ms and Ma2, −2.25 μV at 207 ms).

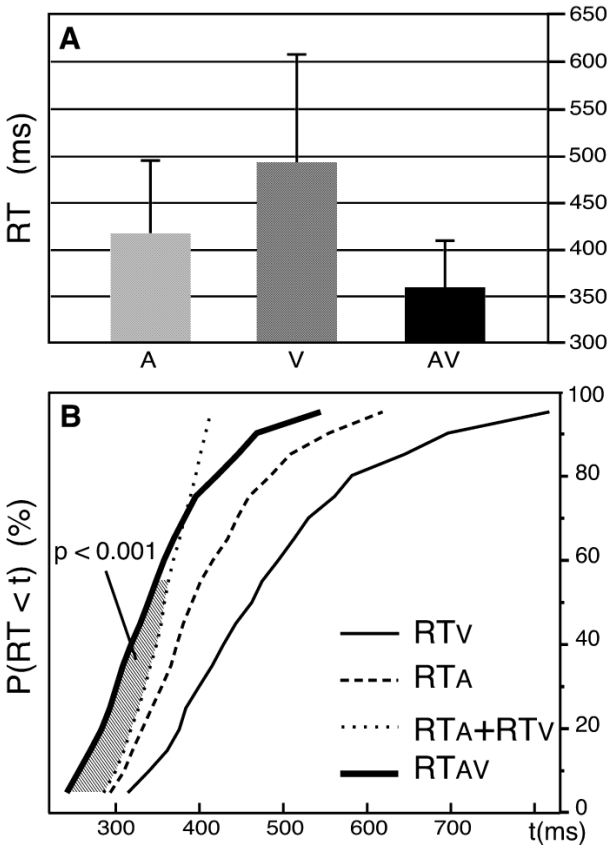

FIG. 2. Violation of the race model inequality in the behavioural-only experiment. (A) Mean reaction times for the auditory, visual and audiovisual trials. (B) Cumulative probability density functions (CDFs) of the reaction times in the three (A, V and AV) conditions of presentation, pooled across subjects. The stimuli and procedure were similar to those used in the main experiment, except that subjects responded to the targets in the three conditions. For shorter reaction times, the CDF for AV responses (thick line) is above the sum of the A and V CDFs (thin dotted line). The hatched area between these two curves illustrates the fractiles for which the violation of the race model inequality [$P(\mathrm{RT_{AV}} < t) \leq P(\mathrm{RT_A} < t) + P(\mathrm{RT_V} < t)$] is statistically significant ($P < 0.001$).

The first deflection in visual ERPs peaked around 40 ms at occipito-parietal electrodes (−3.04 μV at PO3 and −3.60 μV at PO4). Although the onset of the visual stimulus began 240 ms before time zero, this wave is likely to correspond to the visual 'N1' wave, usually peaking around 180 ms post-stimulus. Indeed typical visual N1 responses are usually obtained with stimuli characterized by steep visual energy changes. In our paradigm, the first lip movements were very faint with small progressive changes every 40 ms (see Fig. 1). Therefore the global ERP signal must have also developed progressively, by successive overlaps of small visual responses to each frame delayed by 40, 80, 120 ms, until reaching a 'ceiling' level that appeared about 280 ms after the onset of the first frame. In addition, because we used 12 different visual stimuli (three exemplars of four syllables), the variability of the responses averaged across these stimuli may have reinforced the apparent smoothness of the visual ERP.

A second negative visual component was elicited by lip movements with maximum amplitudes at parieto-central electrodes at about 160 ms (C3, −2.46 μV at 143 ms; C4, −2.16 μV at 179 ms).

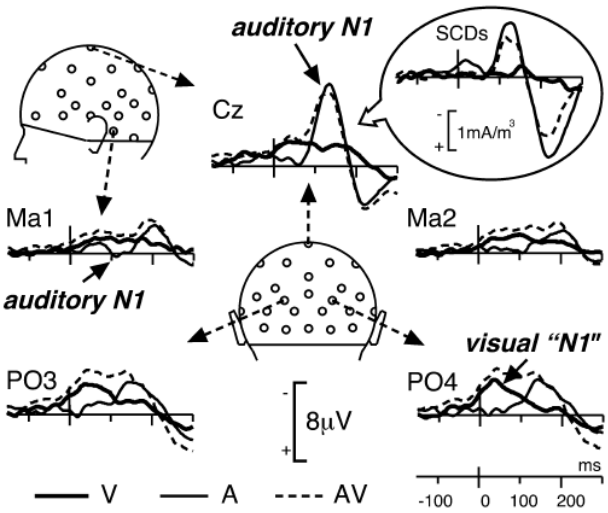

FIG. 3. Unimodal and bimodal responses grand-average ERPs at five illustrative electrodes in each of the three conditions of presentation (A, V and AV) from 150 ms before time zero to 300 ms after. The unimodal auditory N1 wave peaks at 136 ms post-stimulus around Cz with small polarity reversals at mastoid sites (Ma1 and Ma2). The visual 'N1' wave is maximum around occipito-parietal electrodes (PO3 and PO4) at about 40 ms after time zero (this short latency is due to the fact that lip movements began before time zero). Insert: grand-average SCDs at Cz are presented to illustrate the difficulty of interpreting interaction effects locally.

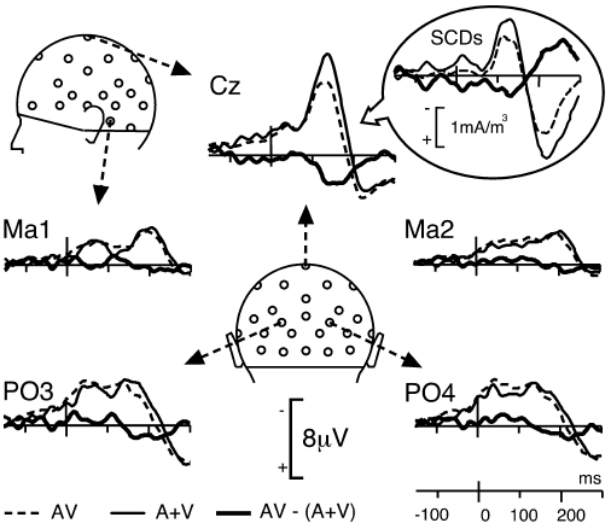

FIG. 4. Bimodal vs. sum of unimodal responses. Comparison of the response to bimodal AV stimuli (dotted lines) with the sum (A + V) of the unimodal responses (thin lines) at five illustrative electrodes, from −150 to +300 ms. The AV response closely follows the A + V trace, except at central sites (illustrated here at Cz) where the two traces significantly differ from about 120 to 190 ms after time zero (see Fig. 5). Insert: for homogeneity with Fig. 3, grand-average SCDs are also presented at this electrode.

Figure 4 displays the superimposition of the ERPs to nontarget bimodal stimuli and the algebraic sum of the responses to unimodal stimuli. Although the morphology of the bimodal response resembles the sum of unimodal ERPs, the differences between the two traces

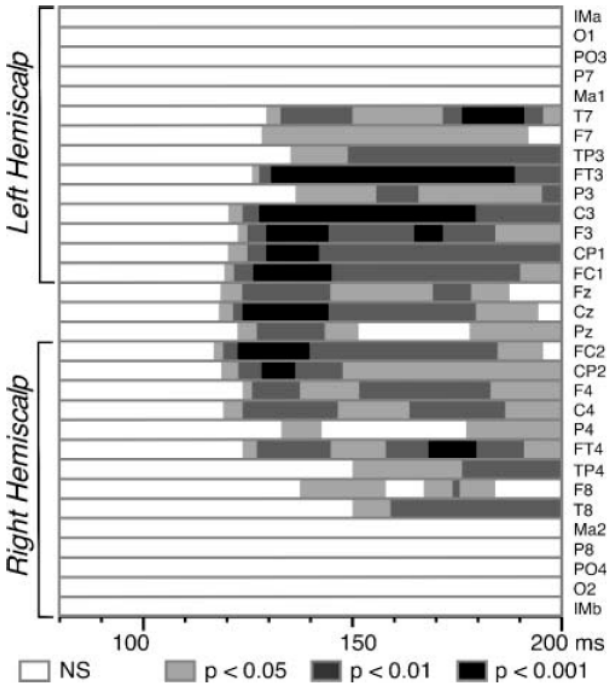

FIG. 5. Statistical significance of the auditory visual interactions. Results of the Student's t-tests (n = 15 subjects) comparing the [AV − (A + V)] amplitudes to zero at each latency from 80 to 200 ms after time zero. Electrodes at the centre of the figure correspond to frontal and central sites and those at the extrema (top and bottom) to more lateral sites. Significant interactions start around 120 ms over fronto-central areas with stronger effects ($P < 0.001$) on the left hemiscalp.

were highly significant over a wide central region within the first 200 ms after time zero. Using the additive criterion [AV − (A + V)] to estimate the interactions, significant patterns were found bilaterally from about 120 to 190 ms at most fronto-central electrodes, that is, in the spatiotemporal range corresponding to the auditory N1 response and the second visual component − {mean amplitude of [AV − (A + V)] over Fz, FC1, FC2, F3, F4, Cz, C3, C4, CP1 and CP2 between 120 and 190 ms: 2.23 (μV)}. The detailed statistical significance of the effect is depicted in Fig. 5. The topography of the cross-modal effect remains roughly stable over the whole 120–190 ms time interval and significance reached the 0.001 threshold at several fronto-central electrodes of the left hemiscalp between 125 and 145 ms latency.

As evidenced by Giard & Peronnet (1999), AV interactions can take multiple forms that are not mutually exclusive: (i) new components that are not present in the unimodal responses; (ii) modulation of the visual response; and (iii), modulation of the auditory response. To assess the nature of the interactions, we therefore compared the topography of [AV − (A + V)] with those of the unimodal responses at the corresponding latencies, with the following reasoning: if the interaction pattern has the same (or inverse) topography as either unimodal response, it is likely to express a modulation (increase or decrease) of that unimodal response. {Note. As ERPs recorded at the scalp surface result from volume conduction activities, it is fundamental to interpret the data from the global topography of the electrical fields and not from a local analysis at one particular electrode. For example, in Fig. 3, the peak amplitudes of the potentials at Cz are similar for A and AV responses, and one could argue that an effect in

[AV − (A + V)] at that electrode could stem only from the (non null) V signal. However, the corresponding SCD traces (Fig. 3) show a different response pattern at Cz in which (i) A and AV traces clearly differ around their peak latency, and (ii) the V signal is close to zero. As the differences between voltage and SCD signals are mainly due to the reduction of the volume conduction effects from distant generators in SCDs, this example illustrates the difficulty of interpreting local measures (potential or SCD) and the need to take the global topography of responses into account.}

Figure 6 displays the topography of the auditory and visual responses, the bimodal responses, the sum of the unimodal responses and the [AV − (A + V)] pattern at the latency of the auditory N1 peak (136 ms). As can be seen, the distributions of the interaction pattern over the left and right hemiscalps (top and bottom panels, respectively) strongly resemble those of the auditory response, but with opposite polarities. This similarity appears not only in the potential maps (Fig. 6, first and third rows), but also in the SCD distributions (Fig. 6, second and fourth rows). Indeed, on the left hemiscalp, the SCD map of [AV − (A + V)] displays sharp, positive current sources at C3, Cz and Pz and a negative current sink around Ma1-T5 (Fig. 6, row 2, col. 5). A similar current sink/source pattern with opposite signs can be observed in the auditory N1 map (Fig. 6, row 2, col. 1). On the right hemiscalp, the polarity reversals at temporal sites are less clear, but the configurations are again very similar in the A and [AV − (A + V)] patterns (Fig. 6, row 4, col. 1 and 5).

By contrast, the latency range of the interaction effects (Fig. 4, Cz) also overlaps that of the second component of the visual response (Fig. 3, PO3/PO4/Cz), suggesting that this component could be modulated by the bimodal inputs. This hypothesis would be supported if the topography of the interactions mirrored that of the unimodal visual response. While the SCD distribution of [AV − (A + V)] could also include part of the topography of the visual reponse around the same latency (particularly over central areas), the overall SCD distributions of the interactions are more complex and differ over occipito-temporal scalp sites.

Finally, we modelled the grand-average [AV − (A + V)] signal in the 110–150 ms period (around the peak of the auditory N1 wave), using two symmetrical (one in each hemisphere) ECDs. The best fitting ECDs were found at an eccentricity of 0.37 and explained the experimental data with a goodness-of-fit of 95.1%. When applied to the same time interval of the auditory response, these ECDs explained 92.3% of the variance of the data whereas they explained only 29.3% of the unimodal visual response within the same latency window. (As a comparison, applied to a 40-ms window in the baseline period, the goodness-of-fit was 39.3%).

Hence, although the pattern of audiovisual interactions observed here may include some contributions from central processes activated by visual stimuli alone, it appears to originate for its main part from the same sources as the auditory N1 response, and may therefore reflect a decrease of activity in the N1 generators in the auditory

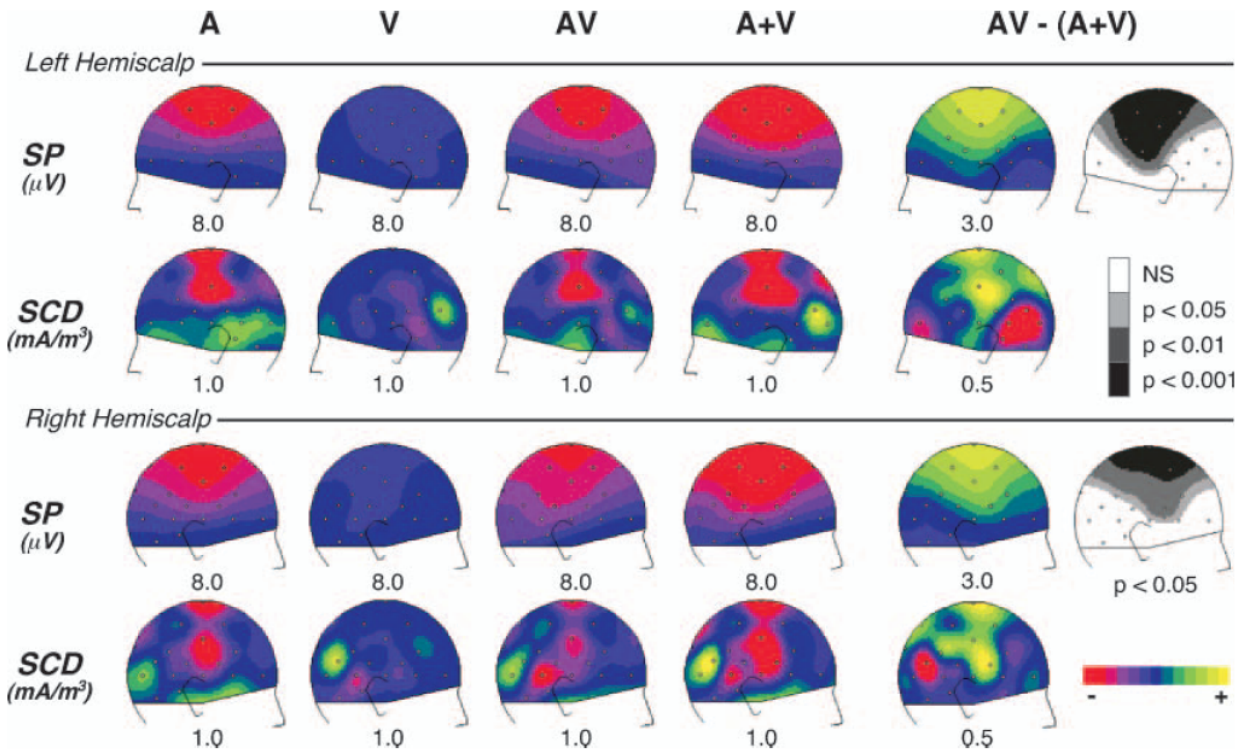

FIG. 6. Comparison of the AV interactions with the auditory N1 wave. Scalp potential (SP) and current density (SCD) topographies over the left and right hemiscalps, at the latency of the unimodal auditory N1 wave (136 ms). In each row the left part displays the distributions of the auditory (A), visual (V), bimodal (AV) responses and the sum of auditory and visual (A + V) responses; the right part shows the distributions of the [AV − (A + V)] interaction pattern with the associated Student's *t*-map estimated on potential values at the same latency (136 ms). The grey colours in *t*-maps indicate the scalp areas where [AV − (A + V)] significantly differs from zero. In potential and SCD maps, half the range of the scale (in μV or mA/m^3) is given below each map. The topography of the crossmodal interaction pattern is similar to that of the unimodal auditory N1 wave, but with opposite polarities. This interaction could therefore reflect a decrease of the unimodal N1 response in auditory cortex.

cortex for the bimodal response compared to the unimodal auditory response.

Furthermore, as we have noted, the SCD maps to auditory alone stimuli display two additional currents sinks on parietal (around Pz) and more anterior (Cz) midline, that both resolve in current sources in the [AV − (A + V)] maps. While a precise interpretation of these current patterns is difficult, the topography of the parietal currents fits with the findings of auditory responses in the posterior intraparietal sulcus (Schroeder *et al.*, 2004). By contrast, the anterior midline currents at Cz could correspond to the 'frontal component' of the auditory N1 response described by Giard's group (Alcaini *et al.*, 1994; Giard *et al.*, 1994). The fact that the same patterns are found with opposite polarities in the auditory-alone condition and in [AV − (A + V)] may indicate that more than the supratemporal component of the auditory response has been modulated by bimodal stimulation.

Discussion

Behavioural facilitation of bimodal speech perception

To date, experimental evidence of behavioural facilitation in bimodal speech perception has been provided almost exclusively on qualitative categorization of phonological continua in situations of sensory conflict (e.g. McGurk illusion: Massaro, 1993) or on detection or intelligibility threshold for degraded unimodal inputs (e.g. speech in noise: Sumby & Pollack, 1954; Grant & Seitz, 2000). Unlike these approaches, our behavioural-only data clearly show that: (i) bimodal information facilitates speech processing also in normal conditions of perception (see also Reisberg *et al.*, 1987; Arnold & Hill, 2001); and (ii), this facilitation can be expressed in chronometric measures, similar to the redundant-stimulus effects widely reported in behavioural studies on cross-modal interactions for nonspeech stimuli (e.g. Hershenson, 1962; Nickerson, 1973; Giard & Peronnet, 1999; Fort *et al.*, 2002a). RT distributions in the three modalities (A, V, AV) falsified the *race* models (Raab, 1962), thereby indicating that unimodal speech inputs interacted somehow during stimulus analysis to speed up the response (coactivation model, Miller, 1982, 1986).

Genuine crossmodal interactions in the ERP paradigm?

The differences in tasks and in the observed RTs for bimodal stimuli preclude a direct application of the previous conclusion to the ERP paradigm on the basis of the sole behavioural measures. It could be argued that the RT effects observed in that paradigm could only result from alertness processes as the visual stimulus started before the auditory stimulus. However, although alerting and spatial orienting of attention are two subcomponents of the attentional system that are probably carried out by separate internal mechanisms (Fernandez-Duque & Posner, 1997), it has been shown that the two processes have similar neural effects on the processing of a subsequent incoming stimulus namely, in the visual modality, an increased activation in extra-striate cortex (Thiel *et al.*, 2004). It is well known in ERP/MEG literature that directing attention to an auditory stimulus results in increased activities in the auditory cortex in a wide latency window including the N1 range (reviews in Näätänen, 1992; Giard *et al.*, 2000). If the auditory processing in the bimodal condition was affected by an alerting process due to the visual signal preceding the acoustic input, the effect would therefore very probably be expressed as a *larger* auditory N1 amplitude for bimodal than for auditory-alone stimuli. Yet we observed a decrease of the auditory N1 amplitude (see next section), strongly suggesting that the alerting hypothesis can be ruled out as a main explanation for the bimodal facilitation in the ERP experiment.

By contrast, several functional imaging studies have reported a decrease of activation in sensory-specific cortices in paradigms where subjects were continuously and exclusively exposed to stimuli in a concurrent modality (Kawashima *et al.*, 1995; Lewis *et al.*, 2000; Bense *et al.*, 2001; Laurienti *et al.*, 2002). Such a cortical deactivation would have led to spurious effects in the [AV − (A + V)] model. However, in the present experiment, all auditory, visual and bimodal stimuli were delivered randomly with equal probability, which should considerably reduce this possibility. Furthermore, because attention was mainly focused on the auditory modality, deactivation processes could have occurred only in the visual cortex (where in fact no [AV − (A + V)] effects were found). Therefore, our significant [AV − (A + V)] effects in temporal areas very probably reflect genuine cross-modal interactions.

Cross-modal depression in the auditory cortex

Both the potential and scalp current density distributions of the crossmodal interactions from about 120 to 150 ms after sound onset mimic those of the unimodal auditory N1 wave in the same latency range. This similarity, also evident in the results of spatio-temporal dipole modelling, strongly suggests that audiovisual integration in speech perception operates at least in part by decreasing the N1 generator activities in supratemporal auditory cortex. This interpretation (that does not preclude the involvement of other additional mechanisms) raises several comments.

First, Miki *et al.* (2004) recently reported no difference in the auditory M100 response (the MEG analogue of the N100 or N1 response) to vowel sounds when they were presented together with the stilled image of a closed mouth or the image of an open mouth pronouncing this vowel. Several reasons may explain these different results: Miki *et al.* used only one stilled image of a mouth pronouncing /a/ in a passive task while we required an auditory discrimination between 12 (three exemplars of four syllables) ecological moving lip movements. First, the use of a passive task prevents from knowing whether their stimuli induced a behavioural facilitation relative to their control condition. In addition, still images and moving speech stimuli may access partly different cortical networks (Calvert & Campbell, 2003). Given that the cross-modal integrative operations are highly sensitive to both the nature of the task and the sensory 'effectiveness' of the unimodal inputs (Fort & Giard, 2004), any of the differences in experimental parameters between the two studies might have explained the differences in results.

Second, the spatial resolution of scalp ERPs does not allow one to rule out the hypothesis that at least part of the interactions are generated in the STS, which has roughly the same orientation as the supratemporal plane. Suppressive effects have indeed been found in the STS in an MEG study comparing the responses to spoken, written and bimodal letters in a recognition task (Raij *et al.*, 2000). However, these effects took place around 380 540 ms and were related to grapheme/phoneme conversion, and do certainly not reflect the same processes as our early interactions occurring at the latency of the auditory sensory N1 response.

Thirdly, congruent bimodal inputs have generally been found to enhance activation in sensory-specific cortices (estimated either by cerebral blood flow measurements: Calvert *et al.*, 1999; Macaluso *et al.*, 2000; or by electric measurements: Giard & Peronnet, 1999; Foxe *et al.*, 2000; Fort *et al.*, 2002b). Yet Giard & Peronnet (1999) reported a decrease of the visual N185 wave (155 200 ms) to

bimodal relative to unimodal visual stimuli in an object discrimination task. This ERP component generated in extra-striate cortex (Mangun, 1995) has been specifically related to visual discrimination processes (Vogel & Luck, 2000). The reduced N185 response was therefore interpreted as reflecting a lesser energetic demand (neural facilitation) from the visual system to discriminate stimuli made more salient by the addition of an auditory cue (see also Fort & Giard, 2004). As the auditory N1 wave is known to be related to stimulus feature analysis in the auditory cortex (Näätänen & Picton, 1987; Näätänen & Winkler, 1999), our results could indicate that lip movements have facilitated feature analysis of the syllables in the auditory cortex by a depression mechanism similar to that found in the visual cortex for object processing. This interpretation makes sense if one considers the general advantage of the cognitive system for visual processing (Posner *et al.*, 1976), and the obvious dominance of the auditory modality in the speech domain: cross-modal facilitation would operate, among other neural mechanisms, as suppressive modulation in the more responsive sensory system.

Finally, a reduced response at the auditory N1 latency appears to be specific to audiovisual speech integration, because this effect differs from those found not only during object recognition (Giard & Peronnet, 1999), but also during the discrimination of verbal material presented in spoken (heard) and written forms (Raij *et al.*, 2000).

Latency of the cross-modal effects

In our paradigm, the first auditory information distinguishing between two different vowels appeared 15 ms after time zero (after the voice onset time; see Material and Methods). The onset latency of the cross-modal effects relative to relevant auditory analysis can therefore be estimated at approximately 105 ms. Several studies on multisensory integration using synchronous nonspeech stimuli have reported very early cross-modal effects (from 40 to 50 ms) in sensory-specific cortices (Giard & Peronnet, 1999; Foxe *et al.*, 2000; Fort *et al.*, 2002a; Molholm *et al.*, 2002). In the speech domain, Lebib *et al.* (2003) recently reported that the processing of congruent and incongruent bimodal inputs generated different ERP effects on the auditory P50 component. Although these and our results might seem hardly compatible with the hypothesis of backprojections from higher-level multisensory areas (Calvert, 2001), one might note that audiovisual speech is special compared with other bimodal objects in that its unimodal inputs are intrinsically asynchronous: coarticulation implies that visual information most often precedes speech sounds, so that visual processing has already started when the sound reaches the auditory system. It is therefore possible that our early effects in the auditory cortex are mediated through visual backprojections from the visual associative system (the visual component peaking bilaterally at occipital sites around 40 ms could well have fed subsequent crossmodal processes in auditory cortex) or from the STS (found to be activated by articulatory lip movements alone and by biological motion in general, review in Calvert & Campbell, 2003). This latter hypothesis fits well with two sets of findings:

At the neural level, although there is growing anatomical and electrophysiological evidence in the primate suggesting that every sensory cortex is likely to receive inputs from each other (from auditory to visual cortex, Falchier *et al.*, 2002; from somatosensory to auditory cortex, Schroeder *et al.*, 2001; Schroeder & Foxe, 2002), no direct pathway from the visual to the auditory cortex has yet been found to our knowledge. However, electrophysiological experiments in monkeys have shown that the associative auditory cortex receives visual inputs with laminar patterns typical of feedback connections, which suggests that visual information is conveyed in the auditory cortex by back-projections from associative areas (Schroeder & Foxe, 2002). The upper bank of the STS (which receives feed-forward auditory and visual information) has been proposed as a candidate for the origin of these visual feedback inputs towards the auditory cortex (see also Pandya *et al.*, 1969; Seltzer & Pandya, 1978).

At a functional level, although we ensured that the subjects could not identify the syllables on the basis of visual information preceding the sound onset, the very first lip movements could have preactivated phonetic units in the auditory cortex via the STS. Several ERP studies have shown that unimodal (e.g. Holcomb & Neville, 1990) and intersensory (e.g. Holcomb & Anderson, 1993) semantic priming effects can decrease the amplitude of the N400 wave, a component associated with late semantic processes. In the same line, the reduced auditory N1 amplitude observed in the present study might reflect an intersensory priming effect on phonetic units at an earlier stage of sensory analysis. Intersensory phonetic priming may therefore be seen as a genuine integrative mechanism by which auditory feed-forward and visual feedback information are combined.

According to Näätänen & Winkler (1999), the auditory N1 component corresponds to a prerepresentational stage of stimulus analysis, during which acoustic features are analysed individually, whereas the first neural correlate of an integrated auditory trace is the MMN, the latency onset of which closely follows that of the N1 wave. Several studies have shown that MMN for speech stimuli is sensitive to phonological (categorical) information (review in Näätänen, 2001). If the MMN is an index of the first phonological trace in the auditory processing chain, then our early cross-modal interactions may reflect online binding of audiovisual information at a prerepresentational stage of stimulus analysis, before the phonological (categorical) trace is built. This chronology of events is in agreement both with the observation of MMN to McGurk syllables (Sams *et al.*, 1991; Colin *et al.*, 2002; Möttönen *et al.*, 2002) and with psycholinguistic models of speech perception by ear and eye (Summerfield, 1987; Massaro & Cohen, 2000).

Abbreviations

A, auditory-only; AV, audiovisual; ECD, equivalent current dipoles; EEG, electroencephalogram; ERP, event-related potential; MEG, neuromagnetic; MMN, mismatch negativity; RT, reaction time; SCD, scalp current density; STS, superior temporal sulcus; V, visual only.

References

Alcaini, M., Giard, M.H., Thevenet, M. & Pernier, J. (1994) Two separate frontal components in the N1 wave of the human auditory evoked response. *Psychophysiology*, **31**, 611–615.

Arnold, P. & Hill, F. (2001) Bisensory augmentation: a speechreading advantage when speech is clearly audible and intact. *Br. J. Psychol.*, **92**, 339–355.

Barth, D.S.N., Goldberg, B., Brett, B. & Di, S. (1995) The spatiotemporal organization of auditory, visual and auditory-visual evoked potentials in rat cortex. *Brain. Res.*, **678**, 177–190.

Bense, S., Stephan, T., Yousry, T.A., Brandt, T. & Dieterich, M. (2001) Multisensory cortical signal increases and decreases during vestibular galvanic stimulation (fMRI). *J. Neurophysiol.*, **85**, 886–899.

Besle, J., Fort, A. & Giard, M.-H. (2004) Interest and validity of the additive model in electrophysiological studies of multisensory interactions. *Cognitive Processing*, **5**, 189–192.

Calvert, G.A. (2001) Crossmodal processing in the human brain: Insights from functional neuroimaging studies. *Cereb. Cortex*, **11**, 1110–1123.

Calvert, G.A., Brammer, M.J., Bullmore, E.T., Campbell, R., Iversen, S.D. & David, A.S. (1999) Response amplification in sensory-specific cortices during crossmodal binding. *Neuroreport*, **10**, 2619–2623.

Calvert, G.A. & Campbell, R. (2003) Reading speech from still and moving faces: The neural substrates of visible speech. *J. Cogn. Neurosci.*, **15**, 57–70.

Colin, C., Radeau, M., Soquet, A., Demolin, D., Colin, F. & Deltenre, P. (2002) Mismatch negativity evoked by the McGurk-MacDonald effect: a phonetic representation within short-term memory. *Clin. Neurophysiol.*, **113**, 495–506.

Falchier, A., Clavagnier, S., Barone, P. & Kennedy, H. (2002) Anatomical evidence of multimodal integration in primate striate cortex. *J. Neurosci.*, **22**, 5749–5759.

Fernandez-Duque, D & Posner, M.I. (1997) Relating the mechanisms of orienting and alerting. *Neuropsychologia*, **35**, 477–486.

Fort, A., Delpuech, C., Pernier, J. & Giard, M.H. (2002a) Dynamics of cortico-subcorical crossmodal operations involved in audio-visual object detection in humans. *Cereb. Cortex*, **12**, 1031–1039.

Fort, A., Delpuech, C., Pernier, J. & Giard, M.H. (2002b) Early auditory–visual interactions in human cortex during nonredundant target identification. *Brain Res. Cogn. Brain Res.*, **14**, 20–30.

Fort, A. & Giard, M.-H. (2004) Multiple electrophysiological mechanisms of audio-visual integration in human perception. In Calvert, G., Spence, C. & Stein, B. (eds), *The Handbook of Multisensory Processes*, MIT Press, Cambridge, pp. 503–514.

Foxe, J.J., Morocz, I.A., Murray, M.M., Higgins, B.A., Javitt, D.C. & Schroeder, C.E. (2000) Multisensory auditory–somatosensory interactions in early cortical processing revealed by high-density electrical mapping. *Brain Res. Cogn. Brain Res.*, **10**, 77–83.

Giard, M.-H., Fort, A., Mouchetant-Rostaing, Y. & Pernier, J. (2000) Neurophysiological mechanisms of auditory selective attention in humans. *Front. Biosci.*, **5**, 84–94.

Giard, M.H. & Peronnet, F. (1999) Auditory-visual integration during multimodal object recognition in humans: a behavioral and electrophysiological study. *J. Cogn. Neurosci.*, **11**, 473–490.

Giard, M.H., Perrin, F., Echallier, J.F., Thevenet, M., Froment, J.C. & Pernier, J. (1994) Dissociation of temporal and frontal components in the human auditory N1 wave: a scalp current density and dipole model analysis. *Electroencephalogr. Clin. Neuro.*, **92**, 238–252.

Grant, K.W. & Seitz., P.F. (2000) The use of visible speech cues for improving auditory detection of spoken sentences. *J. Acoust. Soc. Am.*, **108**, 1197–1208.

Guthrie, D. & Buchwald, J.S. (1991) Significance testing of difference potentials. *Psychophysiology*, **28**, 240–244.

Hershenson, M. (1962) Reaction time as a measure of intersensory facilitation. *J. Exp. Psychol.*, **63**, 289–293.

Hillyard, S.A., Teder-Salejarvi, W.A. & Munte, T.F. (1998) Temporal dynamics of early perceptual processing. *Curr. Opin. Neurobiol.*, **8**, 202–210.

Holcomb, P.J. & Anderson, J.E. (1993) Cross-modal semantic priming – a time-course analysis using event-related brain potentials. *Lang. Cogn. Proc.*, **8**, 379–411.

Holcomb, P.J. & Neville, H.J. (1990) Auditory and visual semantic priming in lexical decision: a comparison using event-related brain potentials. *Lang. Cogn. Proc.*, **5**, 281–312.

Kawashima, R., O'Sullivan, B.T. & Roland, P.E. (1995) Positron-emission tomography studies of cross-modality inhibition in selective attentional tasks: Closing the 'mind's eye'. *Proc. Natl. Acad. Sci. USA.*, **92**, 5969–5972.

Laurienti, P.J., Burdette, J.H., Wallace, M.T., Yen, Y.F., Field, A.S. & Stein, B.E. (2002) Deactivation of sensory-specific cortex by cross-modal stimuli. *J. Cogn. Neurosci.*, **14**, 420–429.

Lebib, R., Papo, D., de Bode, S. & Baudonnière, P.M. (2003) Evidence of a visual-to-auditory cross-modal sensory gating phenomenon as reflected by the human P50 event-related brain potential modulation. *Neurosci. Lett.*, **341**, 185–188.

Lewis, J.W., Beauchamp, M.S. & DeYoe, E.A. (2000) A comparison of visual and auditory motion processing in human cerebral cortex. *Cereb. Cortex*, **10**, 873–888.

Macaluso, E., Frith, C. & Driver, J. (2000) Modulation of human visual cortex by crossmodal spatial attention. *Science*, **289**, 1206–1208.

Mangun, G.R., (1995) Neural mechanisms of visual selective attention. *Psychophysiology*, **32**, 4–18.

Massaro, D.W. (1993) Perceiving asynchronous bimodal speech in consonant-vowel and vowel syllables. *Speech Comm.*, **13**, 127–134.

Massaro, D.W. & Cohen, M.M. (2000) Tests of auditory-visual integration efficiency within the framework of the fuzzy logical model of perception. *J. Acoust. Soc. Am.*, **108**, 784–789.

McGurk, H. & McDonald, J. (1976) Hearing lips and seeing voices. *Nature*, **264**, 746–748.

Miki, K., Watanabe, S. & Kakigi, R. (2004) Interaction between auditory and visual stimulus relating to the vowel sounds in the auditory cortex in humans: a magnetoencephalographic study. *Neurosci. Lett.*, **357**, 199–202.

Miller, J.O. (1982) Divided attention: Evidence for coactivation with redundant signals. *Cognit. Psychol.*, **14**, 247–279.

Miller, J.O. (1986) Time course of coactivation in bimodal divided attention. *Percept. Psychophys.*, **40**, 331–343.

Miniussi, C., Girelli, M. & Marzi, C.A. (1998) Neural site of the redundant target effect: Electrophysiological evidence. *J. Cogn. Neurosci.*, **10**, 216–230.

Molholm, S., Ritter, W., Murray, M.M., Javitt, D.C., Schroeder, C.E. & Foxe, J.J. (2002) Multisensory auditory–visual interactions during early sensory processing in humans: a high-density electrical mapping study. *Brain Res. Cogn. Brain Res.*, **14**, 115–128.

Möttönen, R., Krause, C.M., Tiippana, K. & Sams, M. (2002) Processing of changes in visual speech in the human auditory cortex. *Brain Res. Cogn. Brain Res.*, **13**, 417–425.

Näätänen, R. (2001) The perception of speech sounds by the human brain as reflected by the mismatch negativity (MMN) and its magnetic equivalent (MMNm). *Psychophysiology*, **38**, 1–21.

Näätänen, R. (1992) *Attention and Brain Function*, Lawrence Erlbaum Associates, Hillsdale, NJ.

Näätänen, R. & Picton, T.W. (1987) The N1 wave of the human electric and magnetic response to sound: a review and an analysis of the component structure. *Psychophysiology*, **24**, 375–425.

Näätänen, R. & Winkler, I. (1999) The concept of auditory stimulus representation in cognitive neuroscience. *Psychol. Bull.*, **125**, 826–859.

Nickerson, R.S. (1973) Intersensory facilitation of reaction time: Energy summation or preparation enhancement? *Psychol. Rev.*, **80**, 489–509.

Olson, I.R., Gatenby, J.C. & Gore, J.C. (2002) A comparison of bound and unbound audio-visual information processing in the human cerebral cortex. *Brain Res. Cogn. Brain Res.*, **14**, 129–138.

Pandya, D.N., Hallett, M. & Kmukherjee, S.K. (1969) Intra- and interhemispheric connections of the neocortical auditory system in the rhesus monkey. *Brain. Res.*, **14**, 49–65.

Pernier, J., Perrin, F. & Bertrand, O. (1988) Scalp current density fields: Concept and properties. *Electroencephalogr. Clin. Neuro.*, **69**, 385–389.

Perrin, F., Pernier, J., Bertrand, O. & Echallier, J.F. (1989) Spherical splines for scalp potential and current density mapping. *Electroencephalogr. Clin. Neuro.*, **72**, 184–187.

Perrin, F., Pernier, J., Bertrand, O. & Giard, M.H. (1987) Mapping of scalp potentials by surface spline interpolation. *Electroencephalogr. Clin. Neuro.*, **66**, 75–81.

Posner, M.I., Nissen, M.J. & Klein, R.M. (1976) Visual dominance: An information-processing account of its origins and significance. *Psychol. Rev.*, **83**, 157–171.

Raab, D.H. (1962) Statistical facilitation of simple reaction times. *Trans. NY Acad. Sci.*, **24**, 574–590.

Raij, T., Uutela, K. & Hari, R. (2000) Audiovisual integration of letters in the human brain. *Neuron*, **28**, 617–625.

Ratcliff, R. (1979) Group reaction time distributions and an analysis of distribution statistics. *Psychol. Bull.*, **86**, 446–461.

Reisberg, D.J., McLean, J & Goldfield, A. (1987) Easy to hear but hard to understand: A lipreading advantage with intact auditory stimuli. In Dodd, B. & Campbell, R. (eds), *Hearing by Eye: the Psychology of Lipreading*. Lawrence Erlbaum Associates, London, pp. 93–113.

Sams, M., Aulanko, R., Hamalainen, H., Hari, R., Lounasmaa, O.V., Lu, S.T. & Simola, J. (1991) Seeing speech: Visual information from lip movements modifies activity in the human auditory cortex. *Neurosci. Lett.*, **127**, 141–145.

Scherg, M. (1990) Fundamentals of dipole source potential analysis. In Grandori, F., Hoke, M. & Romani, G.L. (eds), *Auditory Evoked Magnetic Fields and Electric Potentials. Advances in Audiology*, Vol. 5. Karger, Basel, pp. 40–69.

Scherg, M. & Von Cramon, D. (1985) A new interpretation of the generators of BAEP waves I–V: Results of a spatio-temporal dipole model. *Electroencephalogr. Clin. Neuro.*, **62**, 290–299.

Scherg, M. & Von Cramon, D. (1986) Evoked dipole source potentials of the human auditory cortex. *Electroencephalogr. Clin. Neuro.*, **65**, 344–360.

Schroeder, C.E., Lindsley, R.W., Specht, C., Marcovici, A., Smiley, J.F. & Javitt, D.C. (2001) Somatosensory input to auditory association cortex in the macaque monkey. *J. Neurophysiol.*, **85**, 1322–1327.

Schroeder, C., Molholm, S., Lakatos, P., Ritter, W. & Foxe, J.J. (2004) Human-simian correspondence in the early cortical processing of multisensory cues. *Cognitive Processing*, **5**, 140–151.

Schroeder, CE & Foxe, J.J. (2002) The timing and laminar profile of converging inputs to multisensory areas of the macaque neocortex. *Brain Res. Cogn. Brain Res.*, **14**, 187–198.

Seltzer, B. & Pandya, D.N. (1978) Afferent cortical connections and architectonics of the superior temporal sulcus and surrounding cortex in the rhesus monkey. *Brain. Res.*, **149**, 1–24.

Sumby, W.H. & Pollack, I. (1954) Visual contribution to speech intelligibility in noise. *J. Acoust. Soc. Am.*, **26**, 212–215.

Summerfield, Q. (1987) Some preliminaries to a comprehensive account of audio-visual speech perception. In Dodd, B. & Campbell, R., eds. *Hearing by Eye: the Psychology of Lipreading*. Lawrence Erlbaum Associates, Location, pp. 3–52.

Thiel, C.M., Zilles, K. & Fink, G.R. (2004) Cerebral correlates of alerting, orienting and reorienting of visuospatial attention: An event-related fMRI study. *Neuroimage*, **21**, 318–328.

Vaughan, H.G. & Ritter, W. (1970) The sources of auditory evoked responses recorded from the human scalp. *Electroencephalogr. Clin. Neuro.*, **28**, 360–367.

Vogel, E.K. & Luck, S.J. (2000) The visual N1 component as an index of a discrimination process. *Psychophysiology*, **37**, 190–203.

Exp Brain Res (2005) 166: 337 344
DOI 10.1007/s00221-005-2375-x

RESEARCH ARTICLE

Julien Besle · Alexandra Fort · Marie-Hélène Giard

Is the auditory sensory memory sensitive to visual information?

Received: 4 August 2004 / Accepted: 9 November 2004 / Published online: 23 July 2005
© Springer-Verlag 2005

Abstract The mismatch negativity (MMN) component of auditory event-related brain potentials can be used as a probe to study the representation of sounds in auditory sensory memory (ASM). Yet it has been shown that an auditory MMN can also be elicited by an illusory auditory deviance induced by visual changes. This suggests that some visual information may be encoded in ASM and is accessible to the auditory MMN process. It is not known, however, whether visual information affects ASM representation for any audiovisual event or whether this phenomenon is limited to specific domains in which strong audiovisual illusions occur. To highlight this issue, we have compared the topographies of MMNs elicited by non-speech audiovisual stimuli deviating from audiovisual standards on the visual, the auditory, or both dimensions. Contrary to what occurs with audiovisual illusions, each unimodal deviant elicited sensory-specific MMNs, and the MMN to audiovisual deviants included both sensory components. The visual MMN was, however, different from a genuine visual MMN obtained in a visual-only control oddball paradigm, suggesting that auditory and visual information interacts before the MMN process occurs. Furthermore, the MMN to audiovisual deviants was significantly different from the sum of the two sensory-specific MMNs, showing that the processes of visual and auditory change detection are not completely independent.

J. Besle
Univ. Lyon 2, Lyon, France

J. Besle · A. Fort · M.-H. Giard
Univ. Lyon 1, Lyon, France

J. Besle · A. Fort · M.-H. Giard
IFNL IFR19, Lyon, France

J. Besle (✉) · A. Fort · M.-H. Giard
INSERM U280, Mental Processes and Brain Activation,
69675 Bron Cedex, France
E-mail: besle@lyon.inserm.fr
Tel.: +33-472-138907
Fax: +33-472-138901

Keywords Electrophysiology · Audiovisual · MMN · Multisensory integration · Memory

Introduction

The most counter-intuitive effect of audiovisual interactions in the brain is, perhaps, the fact that sensory-specific cortices (e.g. the auditory cortex) seem to be sensitive to information from other modalities, even in primary cortices (Bental et al. 1968) and at very early stages of sensory processing (Fort and Giard 2004).

The mismatch negativity (MMN) component of event-related potentials (ERPs) is elicited in the auditory cortex when incoming sounds are detected as deviating from a neural representation of acoustic regularities and is computed by subtracting the responses to frequent standard sounds from those to infrequent deviant sounds. MMN implies the existence of an auditory sensory memory (ASM) that stores a neural representation of the standard against which any incoming auditory input is compared (Ritter et al. 1995). It is mainly generated in the auditory cortex (Kropotov et al. 1995; Alain et al. 1998) and has long been regarded as specific to the auditory modality (Nyman et al. 1990; Näätänen 1992).

It has, however, recently been discovered that the MMN is not completely impervious to crossmodal influences. For example, in bimodal speech processing, an MMN has been shown to be elicited by deviant syllables differing from the standards only on their visual dimension. In this so-called McGurk illusion (McGurk and McDonald 1976), the same physical sound is therefore differently perceived and processed in ASM, depending on the lip movements that are simultaneously seen (Sams et al. 1991; Möttönen et al. 2002; Colin et al. 2002b, 2004). To keep in line with the auditory-specificity assumption, several non-exclusive explanations have been proposed, that are related to the special status of speech. Either there would exist a phonetic MMN

process that is sensitive to the phonetic nature of articulatory movements (Colin et al. 2002b) or visual speech cues could have specific access to the MMN generators in auditory cortex because, like auditory speech, they carry time-varying information (Möttönen et al. 2002).

Nonetheless, generation of an MMN by visual-only deviants is not restricted to the speech domain, because it can also be observed with the ventriloquist illusion, in which the perceived location of a sound is shifted by a spatially disparate visual stimulus (Stekelenburg et al. 2004; see also Colin et al. 2002a). Rather, what these two phenomena have in common is that they give rise to irrepressible audiovisual illusions that seem to occur at a sensory level of representation (McGurk effect: Soto-Faraco et al. 2004; ventriloquist effect: Bertelson and Aschersleben 1998; Vroomen et al. 2001).

The question therefore arises whether any visual change of an audiovisual event, even in the absence of perceived audiovisual illusion, is likely to access the ASM indexed by the MMN. In other words, does the ASM encode more than the auditory part of an audiovisual event?

When replacing articulatory lip-movements by non-speech visual stimuli in a McGurk MMN paradigm, Sams et al. (1991) found no evidence of an auditory MMN elicited by visual variations alone of the audiovisual event. However, it is very possible that in the absence of strong illusion, the effect is of much less amplitude. Moreover, the effect of visual deviance on the MMN process could occur only in a suitable auditory-deviance context: thus the MMN elicited by both auditory and visual deviances of an audiovisual event should be different from the MMN elicited by auditory deviances alone, while a visual deviance alone would not be detected by the auditory system.

We therefore conducted an audiovisual oddball paradigm in which audiovisual deviants differed from audiovisual standards (AV) either on the visual dimension (AV′), on the auditory dimension (A′V) or on both dimensions (A′V′), with the following hypothesis: If visual information is represented in ASM, AV′ deviants should elicit an auditory MMN, or A′V′ deviants should at least elicit an MMN different from those elicited by A′V deviants.

This would be the whole story if there were not a spoilsport: visual mismatch negativity (vMMN). Several studies have recently shown that visual stimuli deviating from repetitive visual standards can elicit a visual analogue of the MMN in the same latency range (review in Pazo-Alvarez et al. 2003). This vMMN seems to be mainly generated in occipital areas (Berti and Schroger 2004) with possibly a more anterior component (Heslenfeld 2003; Czigler et al. 2002), to be independent of attention (Heslenfeld 2003), and to rely also on memory processes (Stagg et al. 2004; Czigler et al. 2002; see however Kenemans et al. 2003). However, it seems that a greater amount of deviance is necessary to evoke a vMMN than an auditory MMN (Pazo-Alvarez et al. 2003).

If visual-specific components are evoked by visual deviances, then it is necessary in our audiovisual paradigm to separate them from the effect of visual information on the auditory-specific MMN process. To disentangle the contributions of each unisensory process (vMMN and auditory MMN) and isolate the effect of visual information on the auditory MMN, we have therefore:

1. conducted an additional visual oddball paradigm[1]. using the same visual inputs as in our main experiment, so as to elicit a genuine visual MMN (V′MMN); and
2. analyzed the voltage and scalp current density (SCD) distributions of that V′MMN relative to the AV′ MMN elicited in the audiovisual paradigm.

If, on the other hand, the two unisensory MMN processes do not somehow interact, then the two unisensory MMNs should be strictly additive.

Methods

Participants

Fifteen right-handed adults (eight female, ages 20 25 years, mean age 23.1 years) were paid to participate in the study, for which they gave a written informed consent in accordance with the Code of Ethics of the World Medical Association (Declaration of Helsinki). All subjects were free of neurological disease, had normal hearing, and normal or corrected-to-normal vision.

Stimuli

The stimuli were inspired from those previously used by our group in various experiments which revealed a variety of crossmodal interactions in the first 200 ms of processing (Giard and Peronnet 1999; Fort et al. 2002a, b; Fort and Giard 2004).

Visual stimuli consisted in the deformation of a circle into an ellipse either in the horizontal or in the vertical direction (Giard and Peronnet 1999). The basic circle had a diameter of 4.55 cm and was displayed on a video screen placed 130 cm in front of the subjects' eyes, subtending a visual angle of 2°. The amount of deformation in either direction relative to the diameter of the circle was 33% and lasted 140 ms. Between each deformation, the circle remained present on the screen; a cross at its centre served as the fixation point.

Auditory stimuli were rich tones (the fundamental and the second and the fourth harmonics) shifting

[1]As, on the one hand, the topography of the auditory MMN is well known and on the other hand, it would have needlessly lengthened the recording session, we chose not to conduct an auditory oddball paradigm

linearly in frequency (fundamental) either from 500 to 540 Hz or from 500 to 600 Hz. Their duration was also 140 ms, including 14 ms rise/fall time.

All stimuli consisted in the synchronous presentation of a visual and an auditory feature. One association (e.g. an elongation in the horizontal direction and a frequency shift from 500 to 540 Hz) was delivered in 76% of the trials (AV standard). Each remaining association was presented in 8% of the trials: the A′V deviant had the same visual feature as the standard but a different auditory feature, the AV′ deviant had the same auditory but a different visual feature, and the A′V′ deviant differed from the AV standard on both dimensions.

To ensure that the MMNs obtained could not be attributed to physical differences between the standard and deviants, the features of the standard and the A′V′ deviant were exchanged in half of the experimental blocks (and so were the features of A′V and AV′ deviants)

Distractive task

An important characteristic of the auditory (and visual) MMN is that it is automatic and pre-attentive. A "pure" MMN (that is not contaminated by attentional processes) is elicited by stimuli that are irrelevant to the subject. This is a more difficult constraint for visual or audiovisual oddball paradigms, because visual stimuli have to be presented in the visual field of the subjects, but outside their attentional focus. We therefore required a task on the fixation cross. From time to time (13% of the trials) the fixation cross disappeared for 120 ms. This disappearance occurred unpredictably within a standard trial but it was desynchronised relative to the trial's onset and could not occur in a standard preceding a deviant trial. Subjects had to stare at the fixation cross and click a button as quickly and accurately as possible when the cross disappeared.

Visual control

To control for the existence of a vMMN and to study its topography, a visual oddball paradigm was conducted. The experimental parameters were those used in the audiovisual paradigm with the sound off: standard visual stimuli occurred in 84% of the trials (V standards) and the other visual feature occurred in 16% of the trials (V′ deviants). Standards and deviants were exchanged in half of the experimental blocks.

Procedure

After setting the ERP recording apparatus, subjects were seated in a dark, sound-attenuating room and were given instructions describing the distractive task along with an audiovisual practice block of 267 trials. They were told to stare at the fixation cross at the centre of the screen, to respond as accurately and as quickly as possible to the cross disappearance and not to pay attention to the circle or the tones. In the audiovisual paradigm, 256 deviant trials of each type (AV′, A′V and A′V′) were randomly delivered among 2432 AV-standard trials, over 12 blocks including 267 trials each (except the last block that included 263 trials) at a fixed ISI of 560 ms, with the constraint that two deviants could not occur in a row. In the visual oddball paradigm, 256 deviant trials were randomly delivered among 1344 V-standard trials, over 6 blocks including 267 trials each, except the last block that included 265 trials (same ISI). The 12 audiovisual and the six visual blocks were randomly presented to the subjects.

EEG recording

EEG was continuously recorded via a Neuroscan Compumedics system through Synamps AC coupled amplifiers (0.1 200 Hz analogue bandwidth; sampling rate: 1 kHz) from 36 Ag to AgCl scalp electrodes referred to the nose and placed according to the International 10 20 System: Fz, Cz, Pz, POz, Iz; Fp1, F3, F7, FT3, FC1, T3, C3, TP3, CP1, T7, P3, P7, PO3, O1, and their counterparts on the right hemi scalp; Ma1 and Ma2 (left and right mastoids, respectively); IMa and IMb (midway between Iz-Ma1 and Iz-Ma2, respectively). Electrode impedances were kept below 5 kΩ. Horizontal eye movements were recorded from the outer canthus of the right eye; eye blinks and vertical eye movements were measured in channels Fp1 and Fp2.

Data analysis

The EEG analysis was undertaken with the Elan Pack software developed at the INSERM U280 laboratory (Lyon, France). Trials with signal amplitudes exceeding 100 μV at any electrode from 300 ms before time 0 to 500 ms after were automatically rejected to discard the responses contaminated by eye movements or muscular activities.

The ERPs to audiovisual stimuli were averaged offline separately for the six different stimulus types (AV and V standards, A′V, AV′, A′V′ and V′ deviants), over a time period of 800 ms including 300 ms pre-stimulus. Trials including disappearance of the fixation cross were not taken into account when averaging. The mean numbers of averaged trials (by subject) were 1299, 649, and 204 for AV-standard, V-standard, and deviants of each type, respectively (about 20.4% of the trials were discarded because of eye movements).

The ERPs were finally digitally filtered (bandwidth: 1.5 30 Hz, slope: 24 dB/octave). The mean amplitude over the [−100 to 0 ms] pre-stimulus period was taken as the baseline for all amplitude measurements.

Topographic analysis

To facilitate interpretation of the voltage values recorded at multiple electrodes over the scalp surface, we analyzed the topographic distributions of the potentials and the associated scalp current densities. Scalp potential maps were generated using two-dimensional spherical spline interpolation and radial projection from T3, T4 or Oz (left, right and back views, respectively), which respects the length of the meridian arcs. The SCDs were obtained by computing the second spatial derivative of the spline functions used in interpolation (Perrin et al. 1987, 1989). The SCDs do not depend on any assumption about the brain generators or the properties of deeper media, and they are reference-free. In addition, SCDs reduce the spatial smearing of the potential fields due to the volume conduction of the different anatomical structures, and thus enhance the contribution of local intracranial sources (Pernier et al. 1988).

Statistical analysis

The MMNs were statistically assessed by *t*-tests comparing the averaged amplitude of the deviant minus standard difference waveform to zero in the 40 ms time-window around the latency of the peak in the grand-average responses. Results are displayed as statistical probability maps associated with the *t*-tests at each electrode.

Results

Behavioral measures

Mean reaction time to respond to the disappearance of the fixation cross was 404 ms (SD = 51 ms) in the audiovisual oddball paradigm and 409 ms (SD = 52 ms) in the visual paradigm. The mean error ratios were respectively 3.51% (SD = 3.13%) and 3.24 % (SD = 3.11%). Neither the reaction times nor the error rates significantly differed between the two paradigms.

A′V MMN

The response to A′V deviants in the audiovisual paradigm began to differ from AV standards at about 120 ms of processing, being more negative at Fz, and more positive at mastoid sites (Ma1 and Ma2) until about 250 ms (Fig. 1). As can be seen from the difference waveforms (Fig. 2), the MMN elicited by A′V deviants was maximum at 199 ms at Fz (−3.75 μV) and at 190 ms around the mastoids sites (2.77 μV at Ma1 and 2.14 μV at Imb).

Scalp potential and current density topographies (Fig. 3. upper left panel) display a clear-cut polarity reversal around the supra-temporal plane, as expected from an auditory MMN generated in the auditory cortex (Giard et al. 1990). Student *t*-tests on the MMN amplitude around its peak latency (199 ms) are significant at most electrodes around the reversal plane.

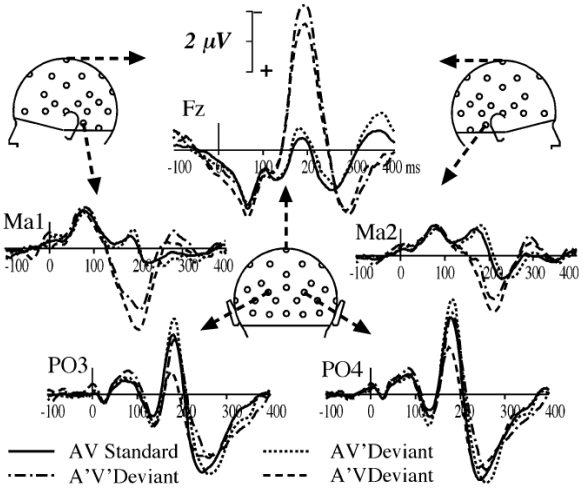

Fig. 1 Potential waveforms elicited by AV standards, AV′, A′V, and A′V′ deviants at a subset of five electrodes (Fz, Ma1, Ma2, PO3, and PO4) from 100 ms pre-stimulus to 400 ms post-stimulus. Negative values are plotted upwards

AV′ MMN

Responses to AV′ deviants and to AV standards are hardly different (Fig. 1). However, the deviant minus standard difference curves (Fig. 2) reveal an occipital deflection that peaks bilaterally at a latency of 192 ms (−0.87 μV at PO4 and −0.63 μV at PO3), with a second peak around 215 ms (−0.72 μV at PO3 and PO4). Figure 3 (upper right panel) illustrates the bilateral occipital

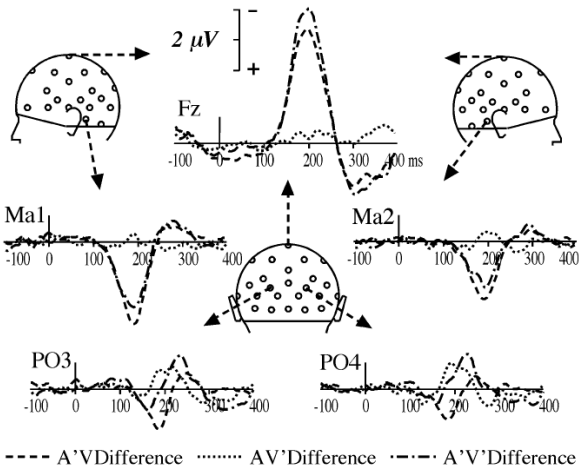

Fig. 2 Deviant minus AV standard difference waveforms for each AV′, A′V, and A′V′ deviant type in the audiovisual oddball paradigm at the same subset of electrodes as in Fig. 1

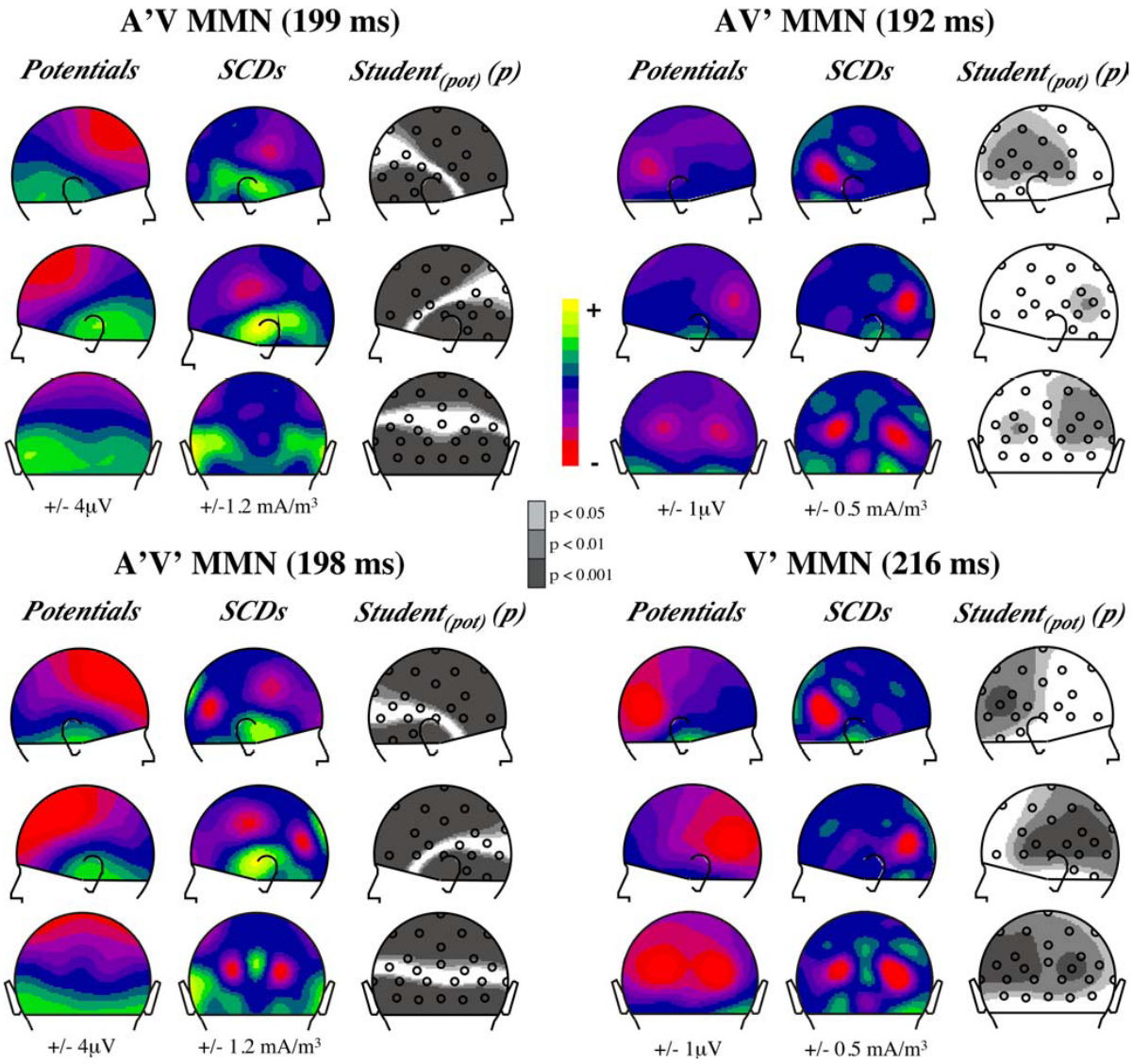

Fig. 3 Topographies of the MMNs elicited by each deviant type in the audiovisual oddball paradigm and by the visual deviant in the visual oddball paradigm (*lower right panel*). Scalp potentials (1st column of each panel), scalp current densities (2nd column) and probability maps associated to Student *t*-tests (3rd column) are presented in right, left, and back views at the latency of the MMN peak indicated above each panel. In potential and SCD maps half the range of the colour scale is given below each column. In Student *t*-maps *grey* areas include electrodes where the averaged potential in a 40 ms time-window around the indicated latency significantly differs from zero

topography of the AV′ MMN at the latency of its largest peak and the statistical significance of its amplitude on the scalp around its peak latency.

A′V′ MMN

Although the responses elicited by A′V′ deviants most resemble those elicited by A′V deviants at fronto-central and mastoid sites (Figs. 1, 2), they tend to come near the curves elicited by AV′ deviants at occipital sites (Fig. 2). Figure 3 (lower left panel) displays the SCD distribution of A′V′ MMN at the latency of its peak in ERPs (198 ms). It clearly shows that it consists of the SCD patterns observed in both the auditory MMN component and the component elicited by AV′ deviants at occipital sites.

Additivity of the MMNs

The additivity of the MMNs elicited by each deviant type was tested by Student *t*-tests comparing the

A′V′ MMN to the sum of the two AV′ MMN and A′V MMN, averaged in the 178 to 218 ms latency window (around the peak latency of both the A′V and the A′V′ MMNs at Fz). Figure 4A shows that additivity is violated at several left parieto-temporal electrodes.

Visual oddball paradigm

Figure 5 displays the ERPs elicited by V standards and V′ deviants in the visual oddball paradigm, and the deviant minus standard difference curve. As in the audiovisual paradigm, the V′ deviants elicited a bilateral occipital component that peaked at a latency of 215 ms (that is at the latency of the second peak of the AV′ MMN), with a larger amplitude (−1.19 μV at PO3 and −1.21 μV at PO4). Figure 3 (lower right panel) displays the topography of the vMMN and the brain areas where its amplitude is statistically significant. There was no hint of an anterior component as was found in other studies investigating the visual MMN.

We further compared the vMMN elicited in the visual paradigm (V′ deviants) and the audiovisual paradigm (AV′ deviants) with Student *t*-tests in the 40 ms time-window around the peak latency that was common to both vMMNs (216 ms). As shown in Fig. 4B, tests were significant at several electrodes over the left hemi scalp.

Discussion

Auditory deviance of an audiovisual event elicited a classical MMN with topography typical of activities in the auditory cortex. Visual deviance of an audiovisual object elicited a bilateral occipital component in the same latency range as the auditory MMN. This component was very similar to that found in the visual oddball paradigm. The spatio-temporal characteristics of these two components are consistent with previous reports of an analogue of the MMN in the visual modality (vMMN) and, especially, with the study by Berti and Schroger (2004) who reported a vMMN with a bilateral occipital topography at a latency of 240 ms.

Thus in our data, the visual variation of an audiovisual event does not seem to elicit an auditory MMN, unlike what has been observed in McGurk and ventriloquist illusions (e.g. Möttönen et al. 2002; Stekelenburg et al. 2004), suggesting that real or illusory perception of an auditory change is necessary to elicit a clear MMN response in the auditory cortex.

In addition, we found that the MMN to deviance on both the auditory and visual dimensions of a bimodal event includes both supratemporal and occipital components, suggesting that the deviance detection processes operate separately in each modality.

However, the vMMNs elicited in the visual (V′ MMN) and the audiovisual (AV′ MMN) oddball paradigms were found to significantly differ, while the only difference between these paradigms was the presence or absence of the same sound that was constantly associated with the visual standards and deviants. Two mutually non-exclusive explanations could account for this finding:

1. either an auditory MMN of small amplitude induced by visual change of the audiovisual event by the same phenomenon as in audiovisual illusions is superimposed to the vMMN and alters its topography;

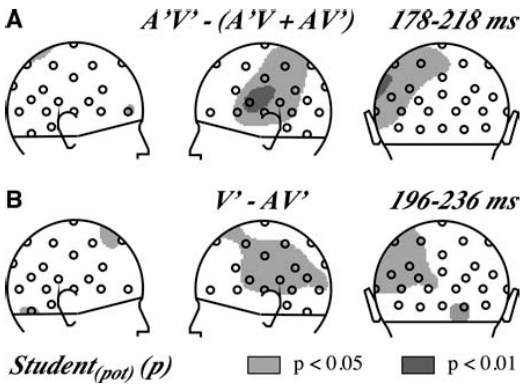

Fig. 4 **A** Test of the additivity of the three MMNs elicited in the audiovisual oddball paradigm presented as probability maps associated to Student *t*-tests in right, left, and back views. *Grey* areas include electrodes where the averaged potential in the indicated time-window significantly differs between the MMN elicited by A′V′ deviants and the sum of the MMNs elicited by A′V and AV′ deviants. **B** Comparison of the MMNs elicited in the audiovisual and the visual oddball paradigm presented as probability maps associated to Student *t*-tests in right, left, and back views. *Grey* areas include electrodes where the averaged potential in the indicated time-window significantly differs between the AV′ MMN and V′ MMN

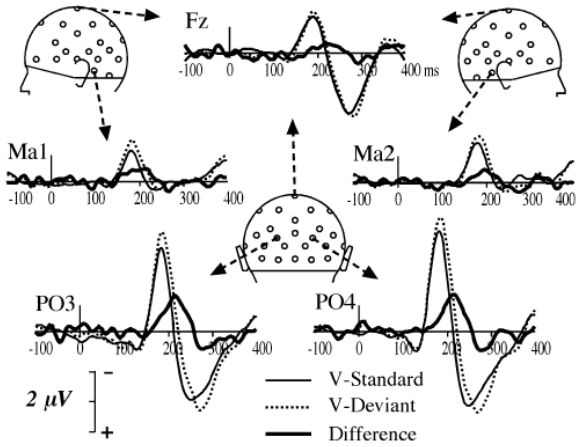

Fig. 5 ERPs elicited by V standards and V′ deviants and deviant minus standard difference waveforms at a subset of five electrodes (Fz, Ma1, Ma2, PO3 and PO4) from 100 ms pre-stimulus to 400 ms post-stimulus

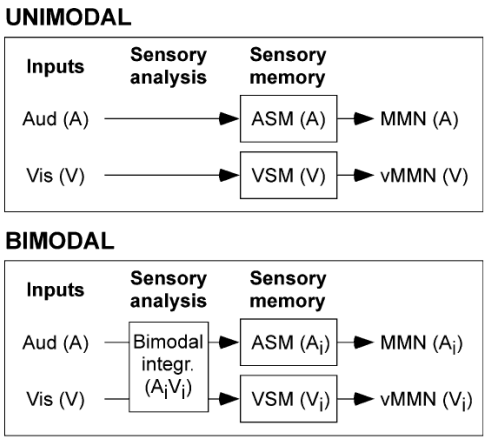

Fig. 6 A schematic model of the MMN processes for unimodal and bimodal inputs. In unimodal parameters, the model refers to that proposed by Näätänen (1992) for auditory MMN. When auditory (*A*) and visual (*V*) inputs are synchronously presented, crossmodal interactions underlying the construction of a multimodal percept can start at early stages of analysis in the afferent sensory systems and modify the input signals before they are encoded in the respective sensory memories (ASM and VSM). The auditory and visual MMN processes would operate on these new sensory signals (A_i and V_i). Note that, although the stage of crossmodal interactions is outlined exclusively before the MMN processes in this figure for simplification, these interactions may persist for several hundred of milliseconds

2. or, if the vMMN reflects a memory-based process (Czigler et al. 2002), an audiovisual event would be encoded differently from a visual-only event in that memory.

The first possibility is hardly supported by the topographies displayed in Fig. 3, and the scalp distribution of the significant differences between the V′ MMN and the AV′ MMN is difficult to interpret regarding either hypothesis.

Whichever hypothesis is correct, the difference between V′ MMN and AV′ MMN implies that the auditory and visual features of the bimodal input have been already partly combined in the afferent sensory systems before the MMN process occurs. This assumption fits with several recent observations that the construction of an integrated percept from bimodal inputs begins at very early stages of sensory analysis, well before the latency of the MMN processes (e.g. Giard and Peronnet 1999; Fort et al. 2002a; Molholm et al. 2002; Lebib et al. 2003; Besle et al. 2004). In addition, the fact that MMN is sensitive to the perceptual dimension of the stimulus (here its multimodal status) rather than to its physical dimension has been well documented in the auditory modality (review in Näätänen and Winkler 1999) and can also explain the existence of an auditory MMN in the McGurk illusion.

However, neither the auditory nor the visual sensory memory seems to encode an integrated trace of the audiovisual deviants. Indeed, in this case, assuming that the source of a MMN is likely to be close to or at the location of the memory upon which that MMN is based, the three kinds of deviant (A′V, AV′ and A′V′) should have elicited MMNs with similar topographies, because they would be all based on the same integrated memory representation. Rather, the three MMNs were found to have clearly different topographies with components typical of activities in the respective sensory-specific cortices, indicating that the MMN processes would operate mostly separately in each modality on the different sensory components of the multimodal representation under construction (Fig. 6).

Nonetheless, the hypothesis of complete independence of auditory and visual MMN processes is unlikely because the MMN to the double deviants in the audiovisual paradigm departs hardly from the mere addition of its unisensory components. (For the sake of simplification, this interpretation has not been shown in Fig. 6.) Future experimentation should be conducted to further assess the relationships between the auditory and visual sensory memories and MMN processes.

References

Alain C, Woods DL, Knight RT (1998) A distributed cortical network for auditory sensory memory in humans. Brain Res 812:23 37

Bental E, Dafny N, Feldman S (1968) Convergence of auditory and visual stimuli on single cells in the primary visual cortex of unanesthetized unrestrained cats. Exp Neurol 20:341 351

Bertelson P, Aschersleben G (1998) Automatic visual bias of perceived auditory location. Psychon Bull Rev 5:482 489

Berti S, Schroger E (2004) Distraction effects in vision: behavioral and event-related potential indices. Neuroreport 15:665 669

Besle J, Fort A, Delpuech C, Giard M-H (2004) Bimodal speech: early suppressive visual effects in the human auditory cortex. Eur J Neurosci 20:2225 2234

Bruneau N, Roux S, Garreau B, Martineau J, Lelord G (1990) Cortical evoked potentials as indicators of Auditory-Visual Cross-Modal Association in young adults. Pavlov J Biol Sci 25:189 204

Cahill L, Ohl F, Scheich H (1996) Alteration of auditory cortex activity with a visual stimulus through conditioning: a 2-deoxyglucose analysis. Neurobiol Learn Mem 65:213 222

Colin C, Radeau M, Soquet A, Demolin D, Colin F, Deltenre P (2002b) Mismatch negativity evoked by the McGurk MacDonald effect: a phonetic representation within short-term memory. Clin Neurophysiol 113:495 506

Colin C, Radeau M, Soquet A, Dachy B, Deltenre P (2002a) Electrophysiology of spatial scene analysis: the mismatch negativity (MMN) is sensitive to the ventriloquism illusion. Clin Neurophysiol 113:507 518

Colin C, Radeau M, Soquet A, Deltenre P (2004) Generalization of the generation of an MMN by illusory McGurk percepts: voiceless consonants. Clin Neurophysiol 115:1989 2000

Czigler I, Balazs L, Winkler I (2002) Memory-based detection of task-irrelevant visual changes. Psychophysiology 39:869 873

Fort A, Giard M-H (2004) Multiple electrophysiological mechanisms of audio-visual integration in human perception. In: Calvert G, Spence C, Stein B (eds) The handbook of multisensory processes. MIT Press, Cambridge

Fort A, Delpuech C, Pernier J, Giard MH (2002a) Dynamics of cortico-subcortical crossmodal operations involved in audio-visual object detection in humans. Cereb Cortex 12:1031 1039

Fort A, Delpuech C, Pernier J, Giard MH (2002b) Early auditory-visual interactions in human cortex during nonredundant target identification. Brain Res Cogn Brain Res 14:20 30

Giard MH, Peronnet F (1999) Auditory-visual integration during multimodal object recognition in humans: a behavioral and electrophysiological study. J Cogn Neurosci 11:473 490

Giard MH, Perrin F, Pernier J (1990) Brain generators implicated in processing of auditory stimulus deviance A topographic ERP study. Psychophysiology 27:627 640

Heslenfeld DJ (2003) Visual mismatch negativity. In: Polich J (ed) Detection of change: event-related potential and fMRI findings. Kluwer Academic Publishers, Dordrecht, pp 41 60

Kenemans JL, Jong TG, Verbaten MN (2003) Detection of visual change: mismatch or rareness?. Neuroreport 14:1239 1242

Kropotov JD, Näätänen R, Sevostianov AV, Alho K, Reinikainen K, Kropotova OV (1995) Mismatch negativity to auditory stimulus change recorded directly from the human temporal cortex. Psychophysiology 32:418 422

Lebib R, Papo D, de Bode S, Baudonniere PM (2003) Evidence of a visual-to-auditory cross-modal sensory gating phenomenon as reflected by the human P50 event-related brain potential modulation. Neurosci Lett 341:185 188

McGurk H, McDonald J (1976) Hearing lips and seeing voices. Nature 264:746 748

Molholm S, Ritter W, Murray MM, Javitt DC, Schroeder CE, Foxe JJ (2002) Multisensory auditory-visual interactions during early sensory processing in humans: a high-density electrical mapping study. Brain Res Cogn Brain Res 14:115 128

Möttönen R, Krause CM, Tiippana K, Sams M (2002) Processing of changes in visual speech in the human auditory cortex. Brain Res Cogn Brain Res 13:417 425

Näätänen R (1992) Attention and Brain Function. Hillsdale, NJ, USA

Näätänen R, Winkler I (1999) The concept of auditory stimulus representation in cognitive neuroscience. Psychol Bull 125:826 859

Nyman G, Alho K, laurinen P, Paavilainen P, Radil T, Rainikainen K, Sams M, Näätänen R (1990) Mismatch negativity (MMN) for sequences of auditory and visual stimuli: evidence for a mechanism specific to the auditory modality. Electroencephalogr Clin Neurophysiol 77:436 444

Pazo-Alvarez P, Cadaveira F, Amenedo E (2003) MMN in the visual modality: a review. Biol Psychol 63:199 236

Pernier J, Perrin F, Bertrand O (1988) Scalp current density fields: concept and properties. Electroencephalogr Clin Neurophysiol 69:385 389

Perrin F, Pernier J, Bertrand O, Giard M-H (1987) Mapping of scalp potentials by surface spline interpolation. Electroencephalogr Clin Neurophysiol 66:75 81

Perrin F, Pernier J, Bertrand O, Echallier JF (1989) Spherical splines for scalp potential and current density mapping. Electroencephalogr Clin Neurophysiol 72:184 187

Ritter W, Deacon D, Gomes H, Javitt DC, Vaughan HG Jr (1995) The mismatch negativity of event-related potentials as a probe of transient auditory memory: a review. Ear Hear 16:52 67

Rosenblum LD, Fowler CA (1991) Audiovisual investigation of the loudness-effort effect for speech and nonspeech events. J Exp Psychol Hum Percept Perform 17:976 985

Saldana HM, Rosenblum LD (1993) Visual influences on auditory pluck and bow judgments. Percept Psychophys 54:406 416

Sams M, Aulanko R, Hamalainen H, Hari R, Lounasmaa OV, Lu ST, Simola J (1991) Seeing speech: visual information from lip movements modifies activity in the human auditory cortex. Neurosci Lett 127:141 145

Soto-Faraco S, Navarra J, Alsius A (2004) Assessing automaticity in audiovisual speech integration: evidence from the speeded classification task. Cognition 92:B13 B23

Stagg C, Hindley P, Tales A, Butler S (2004) Visual mismatch negativity: the detection of stimulus change. Neuroreport 15:659 663

Stekelenburg JJ, Vroomen J, de Gelder B (2004) Illusory sound shifts induced by the ventriloquist illusion evoke the mismatch negativity. Neurosci Lett 357:163 166

Vroomen J, Bertelson P, de Gelder B (2001) The ventriloquist effect does not depend on the direction of automatic visual attention. Percept Psychophys 63:651 659

Audiovisual events in sensory memory

Julien Besle[1], Anne Caclin[1], Romaine Mayet[1], Françoise Bauchet[2], Claude Delpuech[2], Marie-Hélène Giard[1], Dominique Morlet[1].

[1]INSERM, U821, Brain dynamics and cognition, Lyon, F-69500, France;
[1]Institut Fédératif des Neurosciences, Lyon, F-69000, France;
[1]Université Lyon 1, Lyon, F-69000, France
[2]CERMEP – Centre MEG, Lyon, F69000, France

Article in press in *Journal of Psychophysiology*

Corresponding author :

Dominique Morlet
INSERM U281, Brain dynamics and cognition
Centre Hospitalier Le Vinatier
Bâtiment 452
69675 BRON Cedex
Tel: +33 (0)4 72 13 89 03
Fax: +33 (0)4 72 13 89 01

e-mail : morlet@lyon.inserm.fr

ABSTRACT

The functional properties of the auditory sensory memory have been extensively studied using the Mismatch Negativity (MMN) component of the auditory Event-Related Potential (ERP) and its magnetic counterpart recorded using Magneto-encephalography (MEG). It has been found that distinct auditory features (such as frequency or intensity) are encoded separately in sensory memory. Nevertheless, the conjunction of these features (auditory "gestalts") can also be encoded in auditory sensory memory.

Here we investigated how auditory and visual features of bimodal events are represented in sensory memory by recording audiovisual MMNs in two different audiovisual oddball paradigms. The results of a first ERP experiment showed that the sensory memory representations of auditory and visual features of audiovisual events lie within the temporal and occipital cortex respectively, yet with possible interactions between the processing of the unimodal features. In a subsequent MEG experiment, we found some evidence that audiovisual feature conjunctions could also be represented in sensory memory. These results thus extend to the audiovisual domain a number of properties of sensory memory already established within the auditory system.

INTRODUCTION

The Mismatch Negativity is elicited in the auditory cortex when incoming sounds are detected as deviating from a neuronal representation of acoustic regularities. This neuronal representation is likely to form the neurophysiological basis of the Auditory Sensory Memory (ASM) (e.g. Näätänen, 1992; Ritter, Deacon, Gomes, Javitt, & Vaughan, 1995). If one assumes that the mismatch process between the deviant input and the neural trace of the regular (« standard ») stimuli occurs where the deviating feature is stored, then the MMN can be used to study the functional organization of ASM and the representation of sounds in that ASM. For example, it has been shown that the MMNs to sounds deviating in frequency, intensity, or duration, or along different dimensions of timbre, originate from different locations in the auditory cortex, indicating that these different acoustic features are processed in separate registers in ASM (Caclin, Brattico, Tervaniemi, Näätänen, Morlet, Giard, & McAdams, 2006; Giard, Lavikainen, Reinikainen, Perrin, Bertrand, Pernier, & Näätänen, 1995; Rosburg, 2003). On the other hand, it has been found that ASM can also store, besides the separate acoustic features of a sound, the conjunction of those features, suggesting the existence of a « gestalt » representation of sounds in ASM (Gomes, Bernstein, Ritter, Vaughan, & Miller, 1997; Sussman, Gomes, Nousak, Ritter, & Vaughan, 1998; Takegata, Huotilainen, Rinne, Näätänen, & Winkler, 2001; Takegata, Paavilainen, Näätänen, & Winkler, 1999; Winkler, Czigler, Sussman, Horvath, & Balazs, 2005). The principle of these studies was to use several standard sounds created by combining different values of individual features (e.g., location and frequency), and one or several deviants having the very same individual features as the standards, but using different pairings (conjunctions) of these features. Hence the only difference between the standards and the deviants was the particular combination of otherwise identical individual features.

An important question regarding the functional organization of ASM is whether this memory encodes only acoustic features –separately and in conjunction– or if the memory traces can be affected by visual information. A number of studies have indeed established that there exist early interactions between the processing of simultaneous auditory and visual information (Besle, Fort, Delpuech, & Giard, 2004; Fort, Delpuech, Pernier, & Giard, 2002; Giard & Peronnet, 1999; Lebib, Papo, de Bode, & Baudonniere, 2003; Molholm, Ritter, Murray, Javitt, Schroeder, & Foxe, 2002; review in Fort & Giard, 2004), which opens the possibility that the content of ASM could be modified when a visual event accompanies the auditory event.

This last hypothesis is supported by several studies showing an influence of visual cues on the auditory MMN process in particular situations where the presence of visual information

gives rise to auditory perceptual illusions like the McGurk effect or the ventriloquist illusion. In the McGurk illusion (McGurk & McDonald, 1976), the very same physical sound of a syllable can be perceived differently depending on the lip movements that are simultaneously seen (e.g. auditory /ba/ associated with visual /ga/ is perceived as /da/). An auditory MMN can be elicited by audiovisual McGurk syllables deviating from standards only on their visual dimension (Colin, Radeau, Soquet, & Deltenre, in press; Colin, Radeau, Soquet, Demolin, Colin, & Deltenre, 2002b; Möttönen, Krause, Tiippana, & Sams, 2002; Sams, Aulanko, Hamalainen, Hari, Lounasmaa, Lu, & Simola, 1991). Several explanations have been proposed, that are related to the functional specificity of speech: either there would exist a phonetic MMN process that is sensitive to the phonetic nature of articulatory movements (Colin et al., 2002b), or visual speech cues could have a specific access to the MMN generators in auditory cortex because, like auditory speech, they carry time-varying information (Möttönen et al., 2002). Generation of an auditory MMN by visual-only deviants can also be observed with the ventriloquist illusion in which the perceived location of a sound is shifted by a spatially disparate visual stimulus (Colin, Radeau, Soquet, Dachy, & Deltenre, 2002a; Stekelenburg, Vroomen, & de Gelder, 2004). As underlined above however, these two phenomena are highly peculiar in that they give rise to irrepressible audiovisual illusions that seem to occur at a sensory level of representation (ventriloquist effect: Bertelson & Aschersleben, 1998; McGurk effect: Soto-Faraco, Navarra, & Alsius, 2004; Vroomen, Bertelson, & de Gelder, 2001).

Yet in everyday life, perceptual events often occur in multiple sensory systems at once, and the brain coordinates and integrates redundant information from different sensory – particularly auditory and visual – modalities to produce coherent and unified representations of the external world (Calvert, Spence, & Stein, 2004). The question thus arises of how, in the general case, are audiovisual events processed in sensory memory?

Recently, a visual homologue of the auditory MMN, the vMMN has been observed on posterior scalp sites (Berti & Schröger, 2004) around the same latency range as the auditory MMN (review in Pazo-Alvarez, Cadaveira, & Amenedo, 2003). More specifically, a vMMN has been found in response to stimuli deviating from a regular visual sequence either in color, spatial frequency, stimulus contrast, motion direction, shape, line orientation, or stimulus location (review in Czigler, this issue). Although less extensively studied than in the auditory modality, the vMMN might also rely on memory-based processes (Czigler, Balazs, & Winkler, 2002; Stagg, Hindley, Tales, & Butler, 2004) (see however Kenemans, Jong, & Verbaten, 2003) ; Czigler, this issue) The questions therefore are: Is ASM sensitive to general visual information? Are the auditory and visual features of a bimodal event encoded

separately in the memory system underlying the MMN process? Or is a bimodal event processed in a Gestalt manner in this memory?

We have addressed these questions in two experiments, one using event-related potentials (ERPs), the other using Magnetoencephalography (MEG). The first (ERP) study, already published in detail elsewhere (Besle, Fort, & Giard, 2005), will only be briefly recalled here.

EXPERIMENT 1

In the first experiment, we used a bimodal oddball paradigm in which audiovisual deviant stimuli differed from audiovisual standards (AV) either on the visual dimension only (AV'), or on the auditory dimension only (A'V), or on both dimensions simultaneously (A'V'), in order to test the following non-mutually exclusive hypotheses: (i) if the visual dimension of a bimodal event is represented in ASM, then AV' deviants should elicit an auditory MMN (similarly to visual deviants in the McGurk or ventriloquist effects); (ii) if the visual and auditory features of a bimodal event are encoded in separate memory systems, then the MMN to A'V' deviants (A'V'-MMN) should present separate components over temporal and posterior scalp areas; and (iii) if the auditory and visual MMN processes are independent, then the A'V'-MMN should be equal to the sum of the MMNs to unimodal deviants (A'V-MMN + AV'-MMN).

The bimodal stimuli consisted in the deformation of a circle into an horizontal (standard V) or vertical (deviant V') ellipse associated with a synchronously presented rich tone (standard A, deviant A'). Standards (AV) were presented with a probability of 76%, and deviants (A'V, AV', A'V') with a probability of 8% each. In half of the experimental blocks, the visual and auditory features of the standards and deviants were exchanged to insure that the resulting MMNs could not be attributed to physical differences between the standard and deviants. The subjects' (N=15) task was to respond to short and unpredictable disappearance of the fixation cross at the centre of the circle. ERPs were recorded from 36 scalp electrodes with a nose reference. (See Besle et al., 2005, for a detailed description of the stimuli, paradigm, and ERP analysis).

In addition, we conducted as a control a visual oddball experiment in which the visual stimuli were identical to those used in the bimodal paradigm, in order to compare a genuine vMMN (elicited by deviants in a visual sequence) with the MMN elicited by AV' deviants (i.e., visual-only deviants in a bimodal sequence).

The main results were the following: (i) A'V deviants elicited an auditory MMN with a typical temporo-frontal topography; (ii) AV' deviants elicited a "visual MMN" with a bilateral occipital topography; (iii) A'V' deviants elicited an MMN with both temporal and occipital components;

(iv) A'V'-MMN significantly differed from the sum A'V-MMN + AV'-MMN at several temporo-parietal electrodes; and (v) the occipital topography of the unimodal vMMN (recorded in the visual-only paradigm) partly differed from that of AV'-MMN on the left hemiscalp. Figure 1 illustrates some of these results by showing the scalp current density distributions of A'V-MMN, AV'-MMN, A'V'-MMN, and vMMN.

These results already have several important consequences. First, the fact that the AV'-MMN had an occipital (and not temporo-frontal) topography in our protocol could indicate that, in the general case, the visual deviant of a bimodal stimulus elicits a visual MMN-like response; an auditory MMN would be elicited only if that visual deviance gives rise to the (illusory) perception of an auditory change.

However we also found that the visual MMNs elicited in the visual (vMMN) and the audiovisual (AV'-MMN) paradigms significantly differed while the only difference between these paradigms was the presence or absence of repetitive identical sounds associated with both the visual standards and deviants. If the visual MMN reflects a memory-based process (Czigler, this issue), this result may indicate that an audiovisual event is encoded differently than a visual-only event in that visual memory, and thus that the processing of the unisensory features of the bimodal input have already interacted in the afferent sensory systems before the vMMN process occurs. This interpretation fits with the repeated findings that the crossmodal operations underlying the construction of an integrated multimodal percept can begin at very early stages of sensory processing, well before the latency of the MMN processes (Besle et al., 2004; Fort et al., 2002; Giard & Peronnet, 1999; Lebib et al., 2003; Molholm et al., 2002). This is also in agreement with the fact that MMN is sensitive to the perceptual dimensions of the stimulus rather than to its physical dimensions (review in Näätänen & Winkler, 1999), a result which could also explain the elicitation of an auditory MMN in the McGurk illusion.

Nevertheless, the fact that the MMN to simultaneous deviances on both the auditory and visual dimensions of a bimodal event includes supratemporal and occipital components indicates that the MMN processes operate mostly separately in each modality; this would mean that the different sensory components of the multimodal representation under construction are encoded separately in the transient memory systems of their respective modality. These results fit with the repeated findings of separation of elementary feature encoding in sensory memory that have been established in the auditory modality (Caclin et al., 2006; Giard et al., 1995; Rosburg, 2003), and extend these findings to the case of the constituent elements of bimodal events.

However, the MMN elicited by the double auditory and visual deviants partly differed from the sum of the MMNs to single deviants (A'V-MMN + AV'-MMN). This non-additivity of the MMNs could be accounted for by two non-mutually exclusive hypotheses: either the two deviance-detection processes are not entirely independent, or the MMN generating processes can also access, besides the separate auditory and visual stimulus features, the conjunction of these features. Indeed, in the auditory modality, it has been shown that both single features and feature conjunctions may be processed by the MMN system (e.g. Takegata et al., 2001). If this principle holds for audiovisual regularities, one should be able to observe an MMN to the violation of the conjunction of the auditory and visual features of repetitive bimodal events. Note that such an interpretation could also account for the difference between the vMMN recorded in the visual-only paradigm and the AV'-MMN recorded in the bimodal paradigm.

EXPERIMENT 2

This second experiment thus aimed at testing whether, in bimodal events, the conjunction of auditory and visual features may also be encoded in the transient memory system used by the MMN processes. In addition, to further compare the MMN to audiovisual feature conjunction – if existing – to that of the « classical » MMN originating in the auditory cortex, we ran a control auditory-only experiment using the same sounds.

Method

Subjects

Ten right-handed adults (5 female, mean age 29 years) were paid to participate. All were free of any neurological disease, and had normal hearing and normal or corrected-to-normal vision. All participants gave a written informed consent prior to their inclusion in the study in accordance with the Code of Ethics of the World Medical Association (Declaration of Helsinki).

Audiovisual experiment: Stimuli and protocol

We ran an audiovisual oddball paradigm inspired from those used to study the memory representation of auditory feature conjunctions (e.g. Takegata et al., 1999). Four stimuli (two «standards»: A1V1, A2V2, and two «deviants»: A1V2, A2V1) were used with the following hypothesis: since the deviants and standards had the same auditory features (A1, A2) and the same visual features (V1, V2), the deviants should elicit an MMN only if the frequently occuring conjunctions of auditory and visual features have been detected and encoded in the memory representations involved in the MMN process.

The four stimuli were randomly delivered with a probability of 0.44 for each standard type and 0.06 for each deviant type. The visual features consisted in the deformation of a circle into an horizontal (V1) or a vertical (V2) ellipse formed by a 23% reduction of the horizontal (vertical) diameter of the circle. The basic circle had a diameter of 3 cm and was presented permanently on a dark screen placed 85 cm in front of the subjects' eyes (visual angle: 2°). A cross at its centre served as the fixation point. The auditory features consisted in rich harmonic tones (fundamental, 2nd and 4th harmonics) with the fundamental frequency rising linearly from 500 to 540 Hz (A1) or from 500 to 600 Hz (A2). The sounds were delivered binaurally through plastic tubes and earpieces with an intensity adjusted for each subject at 35 dB SL. The auditory and visual features were synchronously presented with a duration of 167 ms (including 7ms of rise/fall times for sounds).

The experiment included 10 blocks of 260 stimuli delivered with an interstimulus interval (onset to onset) of 583 ms. Each block began with the presentation of three standards, and each deviant was preceded by at least 3 standards. The subject's task was to ignore the stimuli and press a key to the disappearance of the fixation cross (pseudo-random disappearance of 120 ms occuring in about 10% of the trials, always during standard trials).

Control auditory experiment

The stimuli were the A1 and A2 sounds used in the audiovisual paradigm. The experiment included 4 blocks of 425 stimuli each delivered with an ISI of 590 ms. In two blocks, the standards (p= 0.88) were A1 and the deviants (p=0.12) were A2; in the other two blocks, the standards and deviants were exchanged. The subjects' task was to read a book of their choice. The auditory-only experiment was always run after the bimodal experiment.

Recordings and data analysis

Recordings were carried out in a magnetically shielded room with a whole-scalp 275 channel CTF system at the MEG-EEG CERMEP Department in Lyon. The magnetic signals were continuously acquired with a 150-Hz low-pass filter and a sampling rate of 600 Hz. EOG activity was recorded from a bipolar montage of two electrodes placed at the outer canthi of both eyes.

Data analysis was performed using the ELAN pack software developed at the INSERM U821 (former U280) laboratory (Lyon). The MEG signals were digitally filtered offline (1-40 Hz bidirectional Butterworth filter, slopes 12 dB/octave). The signals (event-related fields, ERFs) were then averaged separately for each stimulus type over a time period of 500 ms including a 100-ms prestimulus baseline. Responses to the first three standards of each block and to the standards immediately following a deviant were excluded from averaging. For each

subject, a signal rejection threshold was chosen so as to keep about 85% of the remaining trials for averaging.

The auditory MMN (A-MMN) and the MMN to audiovisual feature conjunctions (AV_{conj}-MMN) were measured in the differences between the brain responses to the deviant and the standard stimuli in the auditory-alone and audiovisual experiments, respectively. At the group level, the existence of an MMN (i.e., non-null amplitudes in the difference wave) was assessed at each channel in consecutive 10-ms periods in the time window usually found for the MMN (140-320 ms). We used permutation tests for paired data (the distribution of the deviant-minus-standard response amplitudes under the null hypothesis of an equal amplitude for standards and deviants is estimated by randomly permuting the standard and deviant ERPs within each subject). To further investigate how individual responses relate to the grand average, we also analyzed the significance of the MMNs at the individual level using comparable randomization (permutation) procedures.

Results

Subjects responded to the disappearance of the fixation cross with a mean reaction time of 418 ms (± 50 ms) and less than 1% of errors, showing that they performed the distractive task adequately.

Group analysis

Figure 2.A displays the superimposition of the grand average ERFs across the 10 subjects at all channels for standard and deviant stimuli in the auditory-only and audiovisual paradigms.

In the auditory-only paradigm, the ERF traces for standard and deviant stimuli present a first peak around 60 ms after stimulus onset and begin to differ from each other from about 165 ms with a maximum difference around 200 ms. Statistical analysis showed that the amplitudes of the deviant-minus-standard difference waves at temporal sensors were highly significant between about 155 and 250 ms of latency. These activities showed a stable topography within this time range with polarity reversals over the temporal sites of each hemiscalp, highlighting the presence of an auditory MMN with a main origin in the auditory cortex. Figure 2.B (left) illustrates the mean topography over a 10 ms-period around the MMN peak latency, together with the associated probability map showing the scalp sites of significant MMN amplitudes.

In the audiovisual paradigm, the differences between the responses to standard and deviant stimuli were globally much smaller (Fig 2.A, right) but detailed statistical analysis revealed several sensors presenting significant ERF amplitudes in the difference waveforms between about 210 and 300 ms. Although the scalp areas of significant amplitude spread out less and

the peak latency of the effect over temporal areas was later (280 ms) in the audiovisual paradigm than in the auditory-only condition, the topography of the deviant-minus-standard responses in the audiovisual paradigm resembled that of the auditory MMN on both the left and right temporal sites (Fig. 2.B right). Furthermore, compared to the auditory MMN, it presented an additional component over the occipital sites between about 235 and 265 ms latency with a peak around 250 ms (Fig. 2.B right, bottom line), suggesting the presence of an additional source in the audiovisual condition.

Individual subject analysis

For the auditory-only paradigm, all subjects presented a significant MMN with a topography typical of activities in the auditory cortex. Figure 2.C (left) illustrates the data for one subject (S10) and Table 1 gives the latency window of significant amplitudes for each subject.

The results were much more variable in the audiovisual paradigm: 3 subjects out of 10 presented significant amplitudes with a corresponding MMN topography on both temporal and occipital sites, 5 subjects had an instable and/or unilateral MMN topography with marginally significant amplitudes, and 2 subjects did not present any significant amplitude nor MMN topography (Table 1).

Discussion

The MEG experiment clearly evidenced the presence of an auditory MMN in all subjects when using a classical auditory oddball paradigm with frequency glides as standard and deviant stimuli. Although the relative position of the head within the MEG system varied from one subject to the other (from 5 to 40 mm, because of the variability in head sizes), the presence of a significant MMN signal in the grand-average data underlines the robustness of the auditory MMN process.

While the differences between the responses elicited by the standards and the conjunction deviants in the audiovisual paradigm were much less pronounced, the significant amplitudes and the topography of the grand-average deviant-minus-standard signals strongly suggest that an MMN-like response has been elicited by a change in the conjunction of auditory and visual features of bimodal events. Cross-modal feature conjunctions may thus be represented somehow in the transient memory system used by the MMN processes. Furthermore, this representation would include both temporal and occipital components as suggested by the topography of the AV_{conj}-MMN.

The weak amplitude, associated with a limited statistical significance, of the brain responses to the conjunction of audiovisual features may be explained by several factors. First, in the group analysis, the variability in the subjects' head size and position in the MEG system

might have had a greater effect on the grand-average AV_{conj}-MMN because of its smaller amplitude and the poorer spatial and temporal spreading out of significant field patterns in individual subjects (compared to the auditory-alone condition).

In addition, two other possible explanations can be found in previous MMN studies in the auditory modality. The amplitude and latency of the MMN strongly depend on the strength of the memory trace encoding an auditory regularity (Ritter et al., 1995), as well as on the subject's ability to discriminate sounds deviating from this auditory regularity (e.g. Pakarinen, Takegata, Rinne, Huotilainen, & Näätänen, 2007; Tiitinen, May, Reinikainen, & Näätänen, 1994): the more difficult the discrimination, the later the latency and the smaller the amplitude of the associated MMN. When subjects cannot detect auditory changes, no MMN is elicited. Concerning our experiment, these findings lead to two predictions.

First, the MMN elicited in the auditory-alone paradigm should present a larger amplitude and a shorter latency than the MMN to audiovisual feature conjunction since there is only one type of standard in the first paradigm vs. two in the audiovisual paradigm. Therefore a stronger memory trace is expected in the former than in the latter paradigm, this is indeed what we observed.

Second, a change in audiovisual feature conjunction should not elicit an MMN if the subjects cannot explicitly discriminate the deviants A1V2 and A2V1 from the standards A1V1 and A2V2. To assess whether the absence of an AV_{conj}-MMN in some subjects could be due to the difficulty in discriminating the deviants from the standard stimuli in our paradigm, we have performed an additional behavioural experiment in 6 subjects (mean age: 30 years), including 3 subjects that had participated in the MEG experiment. The experimental set-up was similar to that used in the MEG experiment except that the sounds were delivered through headphones. There were 6 sequences of 100 stimuli delivered in each paradigm (auditory-alone and audiovisual). The subject's task was to press a key as quickly as possible upon the detection of a deviant stimulus. Table 2 presents the results for each subject. In the auditory paradigm, the percentage of correct detections was on average of 98% with a mean response time of 411 ms. In the audiovisual paradigm, the mean percentage of correct responses was much lower (67%) and the response times longer (mean: 723 ms); 3 subjects out of 6 detected less than 53% of the deviants. These behavioural results confirm that the detection of the audiovisual-conjunction deviants was very difficult in our protocol and may thus explain the small amplitude or the absence of MMN elicited by these deviants. This hypothesis is further supported by the results from the 3 subjects who participated in both the MEG and behavioural experiments (S1, S10 and S9 in Table 1, corresponding to S'2, S'3 and S'6, respectively, in Table 2). The first two subjects could correctly discriminate the audiovisual deviants and presented a significant AV_{conj}-MMN

with a temporal and occipital topography, while the third subject did not present any MMN pattern at temporal sites.

To sum up, our experiments support the view that bimodal events are encoded in the transient memory system used by the MMN processes with anatomically separate representations in modality-specific cortices. In addition, the transient memory system could encode not only the single sensory features of bimodal events, but also their conjunction. These data would generalize some of the « rules » established in the auditory modality, namely that (i) both the single features of bimodal events and their conjunction are encoded in the transient memory system used by the MMN processes (Takegata et al, 1999, 2001); and (ii) an MMN would be elicited only if the deviants, whatever their nature can be detected by the subject. Further experiments using audiovisual feature conjunction deviants easier to discriminate from audiovisual standards should be conducted to confirm the existence and the topography of the MMN to audiovisual conjunctions.

Acknowledgements

We thank Claude Delpuech and Françoise Lecaignard at the MEG-EEG CERMEP Department in Lyon for their much appreciated help during the realization of the MEG experiment.

FIGURE LEGENDS

Figure 1

Summary of Experiment 1. Scalp current densities of the deviant-minus-standard ERPs at the latency of their respective maximum amplitude, for the different deviances in the audiovisual paradigm (A'V, AV', and A'V') and in the visual-only paradigm (V'). The range of the colour scale used is indicated below each figure.

Figure 2

Results of the auditory-only (left column) and audiovisual (right column) paradigms in Experiment 2.

A. Superimposition of the MEG responses at all the 275 channels for standard (green lines) and deviant (blue lines) stimuli in each paradigm.

B. Mean topographies of the deviant-minus-standard grand-average responses over a 10-ms window around the peak latency in each paradigm, with the corresponding statistical maps (randomization tests). Although the responses in the audiovisual paradigm are much less significant than in the auditory-only paradigm, the topographies in both paradigms present a polarity reversal over the temporal sites typical of activities in the auditory cortex. In addition, the responses in the audio-visual paradigm present an additional occipital component peaking about 30 ms earlier than the temporal component.

C. Same as in B for one subject (S10).

The range of the colour scale used is indicated below each figure. Significant areas ($p < 0.05$) are depicted in white in the statistical maps.

TABLE 1

	Auditory		Audiovisual Conjunction		
	Left temporal	Right temporal	Left temporal	Right temporal	Occipital
S1	150-240	150-250	190-265	195-265	215-235
S2	190-260	200-275	-	285-330	-
S3	160-230	160-230	-	-	-
S4	170-230	160-230	245-260	245-295	245-265
S5	170-250	170-240	265-275	270-300	280-305
S6	180-260	160-270	-	-	-
S7	180-250	180-270	220-270	175-205	245-255
S8	170-240	180-230	265-295	-	-
S9	205-215	200-250	-	-	230-270
S10	150-230	170-230	275-295	275-295	245-255

Latency windows (ms) of significant MMN amplitudes for each subject in Experiment 2, in each of the two paradigms.

TABLE 2

	Correct deviance detection (%)		Mean response times (ms)	
	Auditory	Audiovisual	Auditory	Audiovisual
S'1	98	53	406	750
S'2 (=S1)	10	98	407	673
S'3 (=S10)	98	84	425	628
S'4	98	50	378	778
S'5	95	77	390	771
S'6 (=S9)	100	44	461	739
Mean ± sd	98 ± 2	67 ± 22	411 ± 29	723 ± 60

Results of the behavioural experiment complementing Experiment 2. Six subjects performed a speeded detection of the deviants in the auditory-only and audiovisual paradigms. The correspondence with the subjects' numbers in Experiment 2 (see Table 1) is indicated for the three subjects concerned.

REFERENCES

Bertelson, P., & Aschersleben, G. (1998). Automatic visual bias of perceived auditory location. *Psychonomic Bulletin & Review, 5*, 482-489.

Berti, S., & Schröger, E. (2004). Distraction effects in vision: behavioral and event-related potential indices. *NeuroReport, 15*(4), 665-669.

Besle, J., Fort, A., Delpuech, C., & Giard, M.-H. (2004). Bimodal speech: Early suppressive visual effects in the human auditory cortex. *European Journal of Neuroscience, 20*(8), 2225-2234.

Besle, J., Fort, A., & Giard, M.-H. (2005). Is the auditory sensory memory sensitive to visual information ? *Experimental Brain Research, 166*(3-4), 337-334.

Caclin, A., Brattico, E., Tervaniemi, M., Näätänen, R., Morlet, D., Giard, M. H., & McAdams, S. (2006). Separate neural processing of timbre dimensions in auditory sensory memory. *Journal of Cognitive Neuroscience, 18*(12), 1959-1972.

Calvert, G. A., Spence, C., & Stein, B. (Eds.). (2004). *The Handbook of Multisensory Processes*. Cambridge: The MIT Press.

Colin, C., Radeau, M., Soquet, A., Dachy, B., & Deltenre, P. (2002a). Electrophysiology of spatial scene analysis: the mismatch negativity (MMN) is sensitive to the ventriloquism illusion. *Clinical Neurophysiology, 113*(4), 507-518.

Colin, C., Radeau, M., Soquet, A., & Deltenre, P. (in press). Generalization of the generation of an MMN by illusory McGurk percepts: voiceless consonants. *Clin Neurophysiol.*

Colin, C., Radeau, M., Soquet, A., Demolin, D., Colin, F., & Deltenre, P. (2002b). Mismatch negativity evoked by the McGurk-MacDonald effect: a phonetic representation within short-term memory. *Clinical Neurophysiology, 113*(4), 495-506.

Czigler, I. Visual Mismatch Negativity: violation of non-attended environmental regulations. *(This issue).*

Czigler, I., Balazs, L., & Winkler, I. (2002). Memory-based detection of task-irrelevant visual changes. *Psychophysiology, 39*(6), 869-873.

Czigler, I., Balazs, L., & Winkler, I. (2002). Memory-based detection of task-irrelevant visual changes. *Psychophysiology, 39*(6), 869-873.

Fort, A., Delpuech, C., Pernier, J., & Giard, M. H. (2002). Dynamics of cortico-subcorical crossmodal operations involved in audio-visual object detection in humans. *Cerebral Cortex, 12*(10), 1031-1039.

Fort, A., & Giard, M.-H. (2004). Multiple electrophysiological mechanisms of audio-visual integration in human perception. In G. Calvert, C. Spence & B. Stein (Eds.), *The Handbook of Multisensory Processes* (pp. 503-514). Cambridge: MIT Press.

Giard, M. H., Lavikainen, J., Reinikainen, K., Perrin, F., Bertrand, O., Pernier, J., & Näätänen, R. (1995). Separate representations of stimulus frequency, intensity, and duration in auditory sensory memory: An Event-related potential and dipole-model analysis. *Journal of Cognitive Neuroscience, 7,2*, 133-143.

Giard, M. H., & Peronnet, F. (1999). Auditory-visual integration during multimodal object recognition in humans : a behavioral and electrophysiological study. *Journal of Cognitive Neuroscience, 11*(5), 473-490.

Gomes, H., Bernstein, R., Ritter, W., Vaughan, H. G., & Miller, J. (1997). Storage of feature conjonctions in transient auditory memory. *Psychophysiology, 34*, 712-716.

Kenemans, J. L., Jong, T. G., & Verbaten, M. N. (2003). Detection of visual change: mismatch or rareness? *NeuroReport, 14*(9), 1239-1242.

Lebib, R., Papo, D., de Bode, S., & Baudonniere, P. M. (2003). Evidence of a visual-to-auditory cross-modal sensory gating phenomenon as reflected by the human P50 event-related brain potential modulation. *Neuroscience letters, 341*(3), 185-188.

McGurk, H., & McDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*, 746-748.

Molholm, S., Ritter, W., Murray, M. M., Javitt, D. C., Schroeder, C. E., & Foxe, J. J. (2002). Multisensory auditory-visual interactions during early sensory processing in humans: a high-density electrical mapping study. *Cognitive Brain Research, 14*(1), 115-128.

Möttönen, R., Krause, C. M., Tiippana, K., & Sams, M. (2002). Processing of changes in visual speech in the human auditory cortex. *Cognitive Brain Research, 13*(3), 417-425.

Näätänen, R. (1992). *Attention and Brain Function*. Hillsdale, NJ.

Näätänen, R., & Winkler, I. (1999). The concept of auditory stimulus representation in cognitive neuroscience. *Psychological Bulletin, 125*(6), 826-859.

Pakarinen, S., Takegata, R., Rinne, T., Huotilainen, M., & Näätänen, R. (2007). Measurement of extensive auditory discrimination profiles using the mismatch negativity (MMN) of the auditory event-related potential (ERP). *Clin Neurophysiol, 118*(1), 177-185.

Pazo-Alvarez, P., Cadaveira, F., & Amenedo, E. (2003). MMN in the visual modality: a review. *Biological Psychology, 63*(3), 199-236.

Ritter, W., Deacon, D., Gomes, H., Javitt, D. C., & Vaughan, H. G., Jr. (1995). The mismatch negativity of event-related potentials as a probe of transient auditory memory: a review. *Ear and Hearing, 16*, 52-67.

Rosburg, T. (2003). Left hemispheric dipole locations of the neuromagnetic mismatch negativity to frequency, intensity and duration deviants. *Cognitive Brain Research, 16*(1), 83-90.

Sams, M., Aulanko, R., Hamalainen, H., Hari, R., Lounasmaa, O. V., Lu, S. T., & Simola, J. (1991). Seeing speech: Visual information from lip movements modifies activity in the human auditory cortex. *Neuroscience Letters, 127*, 141-145.

Soto-Faraco, S., Navarra, J., & Alsius, A. (2004). Assessing automaticity in audiovisual speech integration: evidence from the speeded classification task. *Cognition, 92*(3), B13-23.

Stagg, C., Hindley, P., Tales, A., & Butler, S. (2004). Visual mismatch negativity: the detection of stimulus change. *NeuroReport, 15*(4), 659-663.

Stekelenburg, J. J., Vroomen, J., & de Gelder, B. (2004). Illusory sound shifts induced by the ventriloquist illusion evoke the mismatch negativity. *Neuroscience Letters, 357*(3), 163-166.

Sussman, E., Gomes, H., Nousak, J. M., Ritter, W., & Vaughan, H. G., Jr. (1998). Feature conjunctions and auditory sensory memory. *Brain Research, 793*(1-2), 95-102.

Takegata, R., Huotilainen, M., Rinne, T., Näätänen, R., & Winkler, I. (2001). Changes in acoustic features and their conjunctions are processed by separate neuronal populations. *NeuroReport, 12*(3), 525-529.

Takegata, R., Paavilainen, P., Näätänen, R., & Winkler, I. (1999). Independent processing of changes in auditory single features and feature conjunctions in humans as indexed by the mismatch negativity. *Neuroscience Letters, 266*(2), 109-112.

Tiitinen, H., May, P., Reinikainen, K., & Näätänen, R. (1994). Attentive novelty detection in humans is governed by pre-attentive sensory memory. *Nature, 372*(6501), 90-92.

Vroomen, J., Bertelson, P., & de Gelder, B. (2001). The ventriloquist effect does not depend on the direction of automatic visual attention. *Perception and Psychophysics, 63*(4), 651-659.

Winkler, I., Czigler, I., Sussman, E., Horvath, J., & Balazs, L. (2005). Preattentive binding of auditory and visual stimulus features. *Journal of Cognitive Neuroscience, 17*(2), 320-339.

Bibliographie

Alho, K. (1992). Selective attention in auditory processing as reflected by event-related brain potentials. *Psychophysiology*, *29*, 247-263.

Alsius, A., Navarra, J., Campbell, R. & Soto-Faraco, S. (2005). Audiovisual integration of speech falters under high attention demands. *Current Biology*, *15*(9), 839-843.

Amassian, V. E. & Devito, R. V. (1954). Unit activity in reticular formation and nearby structures. *Journal of Neurophysiology*, *17*(6), 575-603.

Andersen, T. S., Tiippana, K. & Sams, M. (2004). Factors influencing audiovisual fission and fusion illusions. *Cognitive Brain Research*, *21*(3), 301-308.

Andreassi, J. L. & Greco, J. R. (1975). Effects of bisensory stimulation on reaction time and the evoked cortical potential. *Physiological Psychology*, *3*, 189-194.

Aoyama, A., Endo, H., Honda, S. & Takeda, T. (2006). Modulation of early auditory processing by visually based sound prediction. *Brain Research*, *1068*(1), 194-204.

Arden, G. B., Wolf, J. E. & Messiter, C. (2003). Electrical activity in visual cortex associated with combined auditory and visual stimulation in temporal sequences known to be associated with a visual illusion. *Vision Research*, *43*(23), 2469-2478.

Arndt, P. A. & Colonius, H. (2003). Two stages in crossmodal saccadic integration : evidence from a visual-auditory focused attention task. *Experimental Brain Research*, *150*(4), 417-426.

Arnold, P. & Hill, F. (2001). Bisensory augmentation : a speechreading advantage when speech is clearly audible and intact. *British Journal of Psychology*, *92*, 339-355.

Badgaiyan, R. D., Schacter, D. L. & Alpert, N. M. (1999). Auditory priming within and across modalities : Evidence from positron emission tomography. *Journal of Cognitive Neuroscience*, *11*(4), 337-348.

Barth, D. S., Goldberg, N., Brett, B. & Di, S. (1995). The spatiotemporal organization of auditory, visual and auditory-visual evoked potentials in rat cortex. *Brain Research*, *678*, 177-190.

Baynes, K., Funnell, M. G. & Fowler, C. A. (1994). Hemispheric contributions to the integration of visual and auditory information in speech perception. *Perception and Psychophysics*, *55*, 633-641.

Beauchamp, M. S., Argall, B. D., Bodurka, J., Duyn, J. H. & Martin, A. (2004). Unraveling multisensory integration : patchy organization within human STS multisensory cortex. *Nature Neuroscience*, *7*(11), 1190-1192.

Beauchamp, M. S., Lee, K. E., Argall, B. D. & Martin, A. (2004). Integration of auditory and visual information about objects in superior temporal sulcus. *Neuron*, *41*(5), 809-823.

Bell, A. H., Corneil, B. D., Meredith, M. A. & Munoz, D. P. (2001). The influence of stimulus properties on multisensory processing in the awake primate superior colliculus. *Canadian Journal of Experimental Psychology*, *55*, 123-132.

Bell, A. H., Meredith, M. A., Van Opstal, A. J. & Munoz, D. P. (2005). Crossmodal integration in the primate superior colliculus underlying the preparation and initiation of saccadic eye movements. *Journal of Neurophysiology*, *93*(6), 3659-3673.

Bell, C., Sierra, G., Buendia, N. & Segundo, J. P. (1964). Sensory Properties of Neurons in the Mesencephalic Reticular Formation. *Journal of Neurophysiology*, *27*, 961-987.

Ben-Artzi, E. & Marks, L. E. (1995). Visual-auditory interaction in speeded classification : role of stimulus difference. *Perception and Psychophysics*, *57*, 1151-1162.

Benedek, G., Eordegh, G., Chadaide, Z. & Nagy, A. (2004). Distributed population coding of multisensory spatial information in the associative cortex. *European Journal of Neuroscience*, *20*(2), 525-529.

Benedek, G., Fischer-Szatmari, L., Kovacs, G., Pereny, J. & Katoh, Y. Y. (1996). Visual, somatosensory and auditory modality properties along the feline suprageniculate-anterior ectosylvian sulcus/insular pathway. *Progress in Brain Research*, *112*, 325-334.

Benedek, G., Pereny, J., Kovacs, G., Fischer-Szatmari, L. & Katoh, Y. Y. (1997). Visual, somatosensory, auditory and nociceptive modality properties in the feline suprageniculate nucleus. *Neuroscience*, *78(1)*, 179-189.

Benoit, C., Mohamadi, T. & Kandel, S. (1994). Effects of phonetic context on audio-visual intelligibility of French. *Journal of Speech and Hearing Research*, *37*(5), 1195-203.

Bental, E. & Bihari, B. (1963). Evoked activity of single neurons in sensory association cortex of the cat. *Journal of Neurophysiology*, *26*, 207-214.

Bental, E., Dafny, N. & Feldman, S. (1968). Convergence of auditory and visual stimuli on single cells in the primary visual cortex of unanesthetized unrestrained cats. *Experimental Neurology*, *20*, 341-351.

Berman, A. L. (1961). Interaction of cortical responses to somatic and auditory stimuli in anterior ectosylvian gyrus of cat. *Journal of Neurophysiology*, *24*, 608-620.

Bermant, R. I. & Welch, R. B. (1976). Effect of degree of separation of visual-auditory stimulus and eye position upon spatial interaction of vision and audition. *Perceptual and Motor Skills*, *42*(43), 487-493.

Bernstein, I. H. (1970). Can we see and hear at the same time ? *Acta Psychologica*, *33*, 21-35.

Bernstein, I. H., Chu, P. K., Briggs, P. & Schurman, D. L. (1973). Stimulus intensity and foreperiod effects in intersensory facilitation. *Quarterly Journal of Experimental Pscychology*, *25*, 171-181.

Bernstein, I. H., Clark, M. H. & Edelstein, B. A. (1969a). Effects of an auditory signal on visual reaction time. *Journal of Experimental Psychology*, *80*(3), 567-569.

Bernstein, I. H., Clark, M. H. & Edelstein, B. A. (1969b). Intermodal effects in choice reaction time. *Journal of Experimental Psychology*, *81*(2), 405-407.

Bernstein, I. H. & Eason, T. R. (1970). Use of tone offset to facilitate reaction time to light onset. *Psychonomic Science*, *20*, 209-210.

Bernstein, I. H. & Edelstein, B. A. (1971). Effects of some variations in auditory input upon visual choice reaction time. *Journal of Experimental Psychology*, *87*(2), 241-247.

Bernstein, I. H., Rose, R. & Ashe, V. M. (1970a). Energy integration in intersensory facilitation. *Journal of Experimental Psychology*, *86*(2), 196-203.

Bernstein, I. H., Rose, R. & Ashe, V. M. (1970b). Preparatory State Effects in Intersensory Facilitation. *Psychonomic Science*, *19*(2), 113-114.

Bernstein, L. E., Auer, J., E. T. & Moore, J. K. (2004). Audiovisual speech binding : convergence or association ? dans G. A. Calvert, C. Spence & B. Stein (Eds.), *The Handbook of Multisensory Processes* (p. 203-224). Cambridge : The MIT Press.

Bernstein, L. E., Auer, J., E. T., Moore, J. K., Ponton, C. W., Don, M. & Singh, M. (2002). Visual speech perception without primary auditory cortex activation. *Neuroreport*, *13*(3), 311-315.

Bertelson, P. (1998). Starting from the ventriloquist : The perception of multimodal events. dans M. Sabourin, F. Craik & M. Robert (Eds.), *Advances in psychological science : Vol. 1. Biological and cognitive aspects approaches to human cognition* (Vol. 1, p. 419-439). Hove, UK : Psychology Press.

Bertelson, P. & Aschersleben, G. (1998). Automatic visual bias of perceived auditory location. *Psychonomic Bulletin & Review*, *5*, 482-489.

Bertelson, P. & Radeau, M. (1981). Cross-modal bias and perceptual fusion with auditory-visual spatial discordance. *Perception and Psychophysics*, *29*(6), 578-584.

Bertelson, P., Vroomen, J. & de Gelder, B. (2003). Visual recalibration of auditory speech identification : a McGurk aftereffect. *Psychological Science*, *14*(6), 592-597.

Berti, S. & Schröger, E. (2004). Distraction effects in vision : behavioral and event-related potential indices. *Neuroreport*, *15*(4), 665-669.

Besle, J., Caclin, A., Mayet, R., Bauchet, F., Delpuech, C., Giard, M. H. et coll. (sous presse). Audiovisual events in sensory memory. *Journal of Psychophysiology.*

Besle, J., Fort, A., Delpuech, C. & Giard, M. H. (2004). Bimodal speech : Early suppressive visual effects in the human auditory cortex. *European Journal of Neuroscience*, *20*(8), 2225-2234.

Besle, J., Fort, A. & Giard, M. H. (2004). Interest and validity of the additive model in electrophysiological studies of multisensory interactions. *Cognitive Processing*, *5*(3), 189-192.

Besle, J., Fort, A. & Giard, M. H. (2005). Is the auditory sensory memory sensitive to visual information ? *Experimental Brain Research*, *166*(3-4), 337-334.

Bignall, K. E. (1967). Effects of subcortical ablations on polysensory cortical responses and interactions in the cat. *Experimental Neurology*, *18*(1), 56-67.

Bignall, K. E. & Imbert, M. (1969). Polysensory and cortico-cortical projections to frontal lobe of squirrel and rhesus monkeys. *Electroencephalography and Clinical Neurophysiology*, *26*, 206-215.

Binnie, C. A., Montgomery, A. A. & Jackson, P. L. (1974). Auditory visual contributions perception consonants. *Journal of Speech and Hearing Research*, *17*, 619-630.

Blair, R. C. & Karniski, W. (1993). An alternative method for significance testing of waveform difference potentials. *Psychophysiology*, *30*(5), 518-524.

Blasi, V., Paulesu, E., Mantovani, F., Menoncello, L., De Giovanni, U., Sensolo, S. et coll. (1999). Ventral prefrontal areas specialised for lipreading : a PET activation study. *Neuroimage*, *9*(6), S1003.

Bolognini, N., Frassinetti, F., Serino, A. & Ladavas, E. (2005). "Acoustical vision" of below threshold stimuli : interaction among spatially converging audiovisual inputs.

Experimental Brain Research, *160*(3), 273-282.

Bonaventure, N. & Karli, P. (1968). Nouvelles données sur les potentiels d'origine auditive évoqués au niveau du cortex visuel chez la souris. *Comptes rendus des séances de la Société de biologie et de ses filiales*, *163*, 1705-1708.

Bothe, G. G. & Marks, L. E. (1970). Absolute sensitivity to white noise under auxiliary visual stimulation. *Perception and Psychophysics*, *8*(3), 176-178.

Braida, L. D. (1991). Crossmodal integration in the identification of consonant segments. *The Quarterly Journal of Experimental Psychology. A, Human Experimental Psychology*, *43*(3), 647-677.

Brancazio, L. (2004). Lexical influences in audiovisual speech perception. *Journal of Experimental Psychology : Human Perception and Performance*, *30*(3), 445-463.

Brancazio, L. & Miller, J. L. (2005). Use of visual information in speech perception : evidence for a visual rate effect both with and without a McGurk effect. *Perception and Psychophysics*, *67*(5), 759-769.

Brancazio, L., Miller, J. L. & Paré, M. A. (2003). Visual influences on the internal structure of phonetic categories. *Perception and Psychophysics*, *65*(4), 591-601.

Brosch, M., Selezneva, E. & Scheich, H. (2005). Nonauditory events of a behavioral procedure activate auditory cortex of highly trained monkeys. *The Journal of Neuroscience*, *25*(29), 6797-6806.

Brown, A. E. & Hopkins, H. K. (1967). Interaction of the auditory and visual sensory modalities. *The Journal of the Acoustical Society of America*, *41*(1), 1-6.

Buchtel, H. A. & Butter, C. M. (1988). Spatial attentional shifts : implications for the role of polysensory mechanisms. *Neuropsychologia*, *26*(4), 499-509.

Budinger, E., Heil, P. & Scheich, H. (2000). Functional organization of auditory cortex in the Mongolian gerbil (Meriones unguiculatus). III. Anatomical subdivisions and corticocortical connections. *European Journal of Neuroscience*, *12*(7), 2425-2451.

Bulkin, D. A. & Groh, J. M. (2006). Seeing sounds : visual and auditory interactions in the brain. *Current Opinion in Neurobiology*, *16*(4), 415-419.

Burnett, L. R., Stein, B. E., Chaponis, D. & Wallace, M. T. (2004). Superior colliculus lesions preferentially disrupt multisensory orientation. *Neuroscience*, *124*(3), 535-547.

Buser, P. & Rougeul, A. (1956). Réponses sensorielles corticales chez le Chat en préparation chronique. Leurs modifications lors de l'établissement de liaisons temporaires. *Revue Neurologique (Paris)*, *95*(6), 501-503.

Bushara, K. O., Grafman, J. & Hallett, M. (2001). Neural correlates of auditory-visual stimulus onset asynchrony detection. *The Journal of Neuroscience*, *21(1)*, 300-304.

Bushara, K. O., Hanakawa, T., Immisch, I., Toma, K., Kansaku, K. & Hallett, M. (2003). Neural correlates of cross-modal binding. *Nature Neuroscience*, *6*(2), 190-195.

Callan, D. E., Callan, A. M., Kroos, C. & Vatikiotis-Bateson, E. (2001). Multimodal contribution to speech perception revealed by independent component analysis : a single-sweep EEG case study. *Cognitive Brain Research*, *10*, 349-353.

Callan, D. E., Jones, J. A., Munhall, K. G., Callan, A. M., Kroos, C. & Vatikiotis-Bateson, E. (2003). Neural processes underlying perceptual enhancement by visual speech gestures. *Neuroreport*, *14*(17), 2213-2218.

Callan, D. E., Jones, J. A., Munhall, K. G., Kroos, C., Callan, A. M. & Vatikiotis-Bateson,

E. (2004). Multisensory integration sites identified by perception of spatial wavelet filtered visual speech gesture information. *Journal of Cognitive Neuroscience*, *16*(5), 805-816.

Calvert, G. A. (2001). Crossmodal processing in the human brain : insights from functional neuroimaging studies. *Cerebral Cortex*, *11*, 1110-1123.

Calvert, G. A., Brammer, M. J., Bullmore, E. T., Campbell, R., Iversen, S. D. & David, A. S. (1999). Response amplification in sensory-specific cortices during crossmodal binding. *Neuroreport*, *10*(12), 2619-2623.

Calvert, G. A., Bullmore, E. T., Brammer, M. J., Campbell, R., Williams, S. C., McGuire, P. K. et coll. (1997). Activation of auditory cortex during silent lipreading. *Science*, *276*(5312), 593-596.

Calvert, G. A. & Campbell, R. (2003). Reading speech from still and moving faces : the neural substrates of visible speech. *Journal of Cognitive Neuroscience*, *15*(1), 57-70.

Calvert, G. A., Campbell, R. & Brammer, M. J. (2000). Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. *Current Biology*, *10*, 649-657.

Calvert, G. A., Hansen, P. C., Iversen, S. D. & Brammer, M. J. (2001). Detection of audio-visual integration sites in humans by application of electrophysiological criteria to the BOLD effect. *Neuroimage*, *14*, 427-438.

Calvert, G. A. & Thesen, T. (2004). Multisensory integration : methodological approaches and emerging principles in the human brain. *Journal of Physiology (Paris)*, *98*(1-3), 191-205.

Campbell, R. (1992). The neuropsychology of lipreading. *Philosophical Transactions of the Royal Society of London. Series B : Biological Sciences*, *335*(1273), 39-45.

Campbell, R., Garwood, J., Franklin, S., Howard, D., Landis, T. & Regard, M. (1990). Neuropsychological studies of auditory-visual fusion illusions. Four case studies and their implications. *Neuropsychologia*, *28*, 787-802.

Campbell, R., Landis, T. & Regard, M. (1986). Face recognition and lipreading. A neurological dissociation. *Brain*, *109*(3), 509-21.

Campbell, R., MacSweeney, M., Surguladze, S., Calvert, G. A., McGuire, P. K., Suckling, J. et coll. (2001). Cortical substrates for the perception of face actions : an fMRI study of the specificity of activation for seen speech and for meaningless lower-face acts (gurning). *Cognitive Brain Research*, *12*(2), 233-243.

Canon, L. K. (1970). Intermodality inconsistency of input and directed attention as determinants of the nature of adaptation. *Journal of Experimental Psychology*, *84*(1), 141-147.

Canon, L. K. (1971). Directed attention and maladaptive "adaptation" to displacement of the visual field. *Journal of Experimental Psychology*, *88*(3), 403-408.

Cappe, C. & Barone, P. (2005). Heteromodal connections supporting multisensory integration at low levels of cortical processing in the monkey. *European Journal of Neuroscience*, *22*(11), 2886-2902.

Cathiard, M. A. & Tiberghien, G. (1994). Le visage de la parole : une cohérence bimodale temporelle ou configurationnelle. *Psychologie francaise*, *39*(4), 357-374.

Chalupa, L. M. & Rhoades, R. W. (1977). Responses of visual, somatosensory, and auditory

neurones in the golden hamster's superior colliculus. *Journal of Physiology*, *270*(3), 595-626.

Child, I. L. & Wendt, G. R. (1938). The temporal course of the influence of visual stimulation upon auditory threshold. *Journal of Experimental Psychology*, *23*(2), 109-127.

Choe, C. S., Welch, R. B., Guilford, R. M. & Juola, J. F. (1975). The ventriloquist effect : Visual dominance or response bias. *Perception and Psychophysics*, *18*, 55-60.

Ciganek, L. (1966). Evoked potentials in man : interaction of sound and light. *Electroencephalography and Clinical Neurophysiology*, *21*, 28-33.

Clavagnier, S., Falchier, A. & Kennedy, H. (2004). Long-distance feedback projections to area V1 : implications for multisensory integration, spatial awareness, and visual consciousness. *Cognitive Affective and Behavioral Neuroscience*, *4*(2), 117-126.

Cohen, N. E. (1934). Equivalence of brightnesses accross modalities. *The American Journal of Psychology*, *46*, 117-119.

Colin, C., Radeau, M., Soquet, A., Dachy, B. & Deltenre, P. (2002). Electrophysiology of spatial scene analysis : the mismatch negativity (MMN) is sensitive to the ventriloquism illusion. *Clinical Neurophysiology*, *113*(4), 507-518.

Colin, C., Radeau, M., Soquet, A. & Deltenre, P. (2004). Generalization of the generation of an MMN by illusory McGurk percepts : voiceless consonants. *Clinical Neurophysiology*, *115*, 1989-2000.

Colin, C., Radeau, M., Soquet, A., Demolin, D., Colin, F. & Deltenre, P. (2002). Mismatch negativity evoked by the McGurk-MacDonald effect : a phonetic representation within short-term memory. *Clinical Neurophysiology*, *113*(4), 495-506.

Colonius, H. (1990). Possibly dependent probability summation of reaction time. *Journal of Mathematical Psychology*, *34*(1), 253-275.

Colonius, H. & Diederich, A. (2006). The race model inequality : interpreting a geometric measure of the amount of violation. *Psychological Review*, *113*(1), 148-154.

Conrey, B. & Pisoni, D. B. (2006). Auditory-visual speech perception and synchrony detection for speech and nonspeech signals. *The Journal of the Acoustical Society of America*, *119*(6), 4065-4073.

Cooper, B. G., Miya, D. Y. & Mizumori, S. J. Y. (1998). Superior colliculus and active navigation : Role of visual and non- visual cues in controlling cellular representations of space. *Hippocampus*, *8*(4), 340-372.

Cotter, J. R. (1976). Visual and nonvisual units recorded from the optic tectum of Gallus domesticus. *Brain, Behavior and Evolution*, *13*(1), 1-21.

Cotton, J. C. (1935). Normal "visual hearing". *Science*, *82*, 592-593.

Cowan, N., Winkler, I., Teder, W. & Näätänen, R. (1993). Memory prerequisites of mismatch negativity in the auditory event-related potential (ERP). *Journal of Experimental Psychology : Learning, Memory and Cognition*, *19*, 909-921.

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

Czigler, I. (sous presse). Visual Mismatch Negativity : violation of non-attended environmental regulations. *Journal of Psychophysiology*.

Czigler, I., Balazs, L. & Winkler, I. (2002). Memory-based detection of task-irrelevant

visual changes. *Psychophysiology*, *39*(6), 869-873.

Czigler, I. & Winkler, I. (1996). Preattentive auditory change detection relies on unitary sensory memory representations. *Neuroreport*, *7*(15-17), 2413-2417.

Davis, E. T. (1966). Heteromodal effects upon visual threshold. *Psychological Monographs*, *80 (24, Whole No 632)*.

Davis, H., Osterhammel, P. A., Wier, C. C. & Gjerdingen, D. B. (1972). Slow vertex potentials : interactions among auditory, tactile, electric and visual stimuli. *Electroencephalography and Clinical Neurophysiology*, *33*, 537-545.

de Gelder, B., Bocker, K. B. E., Tuomainen, J., Hensen, M. & Vroomen, J. (1999). The combined perception of emotion from voice and face : early interaction revealed by human electric brain responses. *Neuroscience Letters*, *260*(2), 133-136.

de Gelder, B. & Vroomen, J. (2000). The perception of emotions by ear and eye. *Cognition and Emotion*, *14*(3), 289-311.

de Gelder, B., Vroomen, J. & Bertelson, P. (1998). Upright but not inverted faces modify the perception of emotion in the voice. *Current Psychology of Cognition*, *17(4-5)*, 1021-1031.

Dekle, D. J., Fowler, C. A. & Funnell, M. G. (1992). Audiovisual integration in perception of real words. *Perception and Psychophysics*, *51*(4), 355-362.

Diederich, A. & Colonius, H. (1987). Intersensory facilitation in the motor component ? *Psychological Research*, *49*, 23-29.

Diederich, A. & Colonius, H. (1991). A further test of the superposition model for the redundant-signals effect in bimodal detection [comment]. *Perception and Psychophysics*, *50*, 83-86.

Diesch, E. (1995). Left and right hemifield advantages of fusions and combinations in audiovisual speech perception. *The Quarterly Journal of Experimental Psychology. A, Human Experimental Psychology*, *48*, 320-333.

Dittmann-Balcar, A., Thienel, R. & Schall, U. (1999). Attention-dependent allocation of auditory processing resources as measured by mismatch negativity. *Neuroreport*, *10(18)*, 3749-3753.

Dixon, N. F. & Spitz, L. (1980). The detection of auditory visual desynchrony. *Perception*, *9*, 719-721.

Dolan, R. J., Morris, J. S. & de Gelder, B. (2001). Crossmodal binding of fear in voice and face. *Proceedings of The National Academy of Science*, *98*, 10006-10010.

Dorfman, D. D. & Miller, R. (1966). The effect of light on sound intensity generalization after two stimulus discrimination training. *Psychonomic Science*, *4*, 337-338.

Doubell, T. P., Baron, J., Skaliora, I. & King, A. J. (2000). Topographical projection from the superior colliculus to the nucleus of the brachium of the inferior colliculus in the ferret : convergence of visual and auditory information. *European Journal of Neuroscience*, *12*(12), 4290-4308.

Dräger, U. C. & Hubel, D. H. (1975). Responses to visual stimulation and relationship between visual, auditory, and somatosensory inputs in mouse superior colliculus. *Journal of Neurophysiology*, *38*(3), 690-713.

Driver, J. (1996). Enhancement of selective listening by illusory mislocation of speech sounds due to lip-reading. *Nature*, *381*, 66-68.

Driver, J. & Spence, C. (2000). Multisensory perception : beyond modularity and convergence in crossmodal integration. *Current Biology*, *10*, R731-R735.

Dubner, R. & Rutledge, L. T. (1964). Recording and analysis of converging input upon neurons in cat association cortex. *Journal of Neurophysiology*, *27*, 620-34.

Easton, R. D. & Basala, M. (1982). Perceptual dominance during lipreading. *Perception and Psychophysics*, *32*(6), 562-570.

Echallier, J. F., Perrin, F. & Pernier, J. (1992). Computer-assisted placement of electrodes on the human head. *Electroencephalography and Clinical Neurophysiology*, *82*, 160-163.

Edgington, E. S. (1995). *Randomization tests : Third edition : revised and expanded* (Vol. 147). New York : Marcel Dekker.

Edwards, S. B., Ginsburg, C. L., Henkel, C. K. & Stein, B. E. (1979). Sources of subcortical projections to the superior colliculus in the cat. *The Journal of Comparative Neurology*, *184*(2), 309-330.

Efron, B. & Tibshirani, R. J. (1993). *An introduction to the Bootstrap.* Boca Raton : Chapman & Hall/CRC.

Erber, N. P. (1969). Interaction of audition and vision in the recognition of oral speech stimuli. *Journal of Speech and Hearing Research*, *12*(2), 423-425.

Erber, N. P. (1975). Auditory-visual perception of speech. *Journal of Speech and Hearing Disorders*, *40*(4), 481-492.

Eriksen, C. W. (1988). A source of error in attempts to distinguish coactivation from separate activation in the perception of redundant targets. *Perception and Psychophysics*, *44*(2), 191-193.

Eriksen, C. W., Goettl, B., St James, J. D. & Fournier, L. R. (1989). Processing redundant signals : coactivation, divided attention, or what ? *Perception and Psychophysics*, *45*(4), 356-370.

Falchier, A., Clavagnier, S., Barone, P. & Kennedy, H. (2002). Anatomical evidence of multimodal integration in primate striate cortex. *The Journal of Neuroscience*, *22*(13), 5749-5759.

Felleman, D. J. & Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, *1*, 1-47.

Fidell, S. (1970). Sensory function in multimodal signal detection. *The Journal of the Acoustical Society of America*, *47*(4), 1009-1015.

Fishman, M. C. & Michael, C. R. (1973). Integration of auditory information in the cat visual cortex. *Vision Research*, *13*, 1415-1419.

Fort, A., Delpuech, C., Pernier, J. & Giard, M. H. (2002a). Dynamics of cortico-subcorical crossmodal operations involved in audio-visual object detection in humans. *Cerebral Cortex*, *12*(10), 1031-1039.

Fort, A., Delpuech, C., Pernier, J. & Giard, M. H. (2002b). Early auditory-visual interactions in human cortex during nonredundant target identification. *Cognitive Brain Research*, *14*, 20-30.

Fort, A. & Giard, M. H. (2004). Multiple electrophysiological mechanisms of audio-visual integration in human perception. dans G. Calvert, C. Spence & B. Stein (Eds.), *The Handbook of Multisensory Processes* (p. 503-514). Cambridge : MIT Press.

Fowler, C. A. & Rosenblum, L. D. (1991). Perception of the phonetic gesture. dans I. G. Mattingly & M. Studdert-Kennedy (Eds.), *Modularity and the motor theory of speech perception, proceedings of a conference to honor Alvin M. Liberman* (p. 33-59). Hillsdale, NJ : Lawrence Erlbaum Associates.

Foxe, J. J. & Schroeder, C. E. (2005). The case for feedforward multisensory convergence during early cortical processing. *Neuroreport*, *16*(5), 419-423.

Frassinetti, F., Bolognini, N. & Ladavas, E. (2002). Enhancement of visual perception by crossmodal visuo-auditory interaction. *Experimental Brain Research*, *147*(3), 332-343.

Frens, M. A. & Van Opstal, A. J. (1998). Visual-auditory interactions modulate saccade-related activity in monkey superior colliculus. *Brain Research Bulletin*, *46*(3), 211-224.

Frens, M. A., Van Opstal, A. J. & Willigen, R. F. Van der. (1995). Spatial and temporal factors determine auditory-visual interactions in human saccadic eye movements. *Perception and Psychophysics*, *57*, 802-816.

Garner, W. R. (1976). Interaction of stimulus dimensions in concept and choice processes. *Cognitive Psychology*, *8*(1), 98-123.

Gebhard, J. W. & Mowbray, G. H. (1959). On discriminating the rate of visual flicker and auditory flutter. *American Journal of Psychology*, *72*, 521-529.

Ghazanfar, A. A., Maier, J. X., Hoffman, K. L. & Logothetis, N. K. (2005). Multisensory integration of dynamic faces and voices in rhesus monkey auditory cortex. *The Journal of Neuroscience*, *25*(20), 5004-5012.

Ghazanfar, A. A. & Schroeder, C. E. (2006). Is neocortex essentially multisensory ? *Trends in Cognitive Sciences*, *10*(6), 278-85.

Giard, M. H., Fort, A., Mouchetant-Rostaing, Y. & Pernier, J. (2000). Neurophysiological mechanisms of auditory selective attention in humans. *Frontiers in Bioscience*, *5*, 84-94.

Giard, M. H., Lavikainen, J., Reinikainen, K., Perrin, F., Bertrand, O., Pernier, J. et coll. (1995). Separate representations of stimulus frequency, intensity, and duration in auditory sensory memory : An Event-related potential and dipole-model analysis. *Journal of Cognitive Neuroscience*, *7,2*, 133-143.

Giard, M. H. & Peronnet, F. (1999). Auditory-visual integration during multimodal object recognition in humans : a behavioral and electrophysiological study. *Journal of Cognitive Neuroscience*, *11*(5), 473-490.

Gielen, S. C., Schmidt, R. A. & Van den Heuvel, P. J. (1983). On the nature of intersensory facilitation of reaction time. *Perception and Psychophysics*, *34*(2), 161-168.

Gilbert, G. M. (1941). Inter-sensory facilitation and inhibition. *The Journal of General Psychology*, *24*, 381-407.

Giray, M. & Ulrich, R. (1993). Motor coactivation revealed by response force in divided and focused attention. *Journal of Experimental Psychology : Human Perception and Performance*, *19*, 1278-1291.

Godey, B., Schwartz, D., Graaf, J. B. de, Chauvel, P. & Liégeois-Chauvel, C. (2001). Neuromagnetic source localization of auditory evoked fields and intracerebral evoked potentials : a comparison of data in the same patients. *Clinical Neurophysiology*,

112(10), 1850-1859.

Gomes, H., Bernstein, R., Ritter, W., Vaughan, H. G. & Miller, J. (1997). Storage of feature conjonctions in transient auditory memory. *Psychophysiology*, *34*, 712-716.

Gondan, M., Lange, K., Rösler, F. & Röder, B. (2004). The redundant target effect is affected by modality switch costs. *Psychonomic Bulletin & Review*, *11*(2), 307-313.

Gondan, M., Niederhaus, B., Rösler, F. & Röder, B. (2005). Multisensory processing in the redundant-target effect : a behavioral and event-related potential study. *Perception and Psychophysics*, *67*(4), 713-726.

Gordon, B. G. (1973). Receptive fields in the deep layers of the cat superior colliculus. *Journal of Neurophysiology*, *36*, 157-178.

Grant, K. W. (2001). The effect of speechreading on masked detection thresholds for filtered speech. *The Journal of the Acoustical Society of America*, *109*(5), 2272-2275.

Grant, K. W. & Braida, L. D. (1991). Evaluating the articulation index for auditory-visual input. *The Journal of the Acoustical Society of America*, *89*(6), 2952-2960.

Grant, K. W. & Seitz, P. F. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. *The Journal of the Acoustical Society of America*, *108*(3), 1197-1208.

Grant, K. W., van Wassenhove, V. & Poeppel, D. (2004). Detection of auditory (cross-spectral) and auditory-visual (cross-modal) synchrony. *Speech Communication*, *44*, 43-53.

Grant, K. W. & Walden, B. E. (1996). Evaluating the articulation index for auditory-visual consonant recognition. *The Journal of the Acoustical Society of America*, *100*(4), 2415-2424.

Green, D. M. & Swets, J. A. (1966). *Signal Detection Theory and Psychophysics*. New York : Wiley.

Green, K. P. & Gerdeman, A. (1995). Cross-modal discrepancies in coarticulation and the integration of speech information : the McGurk effect with mismatched vowels. *Journal of Experimental Psychology : Human Perception and Performance*, *21*(6), 1409-1426.

Green, K. P. & Kuhl, P. K. (1989). The role of visual information in the processing of place and manner features in speech perception. *Perception and Psychophysics*, *45*(1), 34-42.

Green, K. P. & Kuhl, P. K. (1991). Integral processing of visual place and auditory voicing information during phonetic perception. *Journal of Experimental Psychology : Human Perception and Performance*, *17*(1), 278-288.

Green, K. P., Kuhl, P. K., Meltzoff, A. N. & Stevens, E. B. (1991). Integrating speech information across talkers, gender, and sensory modality : female faces and male voices in the McGurk effect. *Perception and Psychophysics*, *50*(6), 524-536.

Green, K. P. & Miller, J. L. (1985). On the role of visual rate information in phonetic perception. *Perception and Psychophysics*, *38*(3), 269-276.

Gregg, L. W. & Brogden, W. J. (1952). The effect of simultaneous visual stimulation on absolute auditory sensitivity. *Journal of Experimental Psychology*, *43*, 179-186.

Grossenbacher, P. G. & Lovelace, C. T. (2001). Mechanisms of synesthesia : cognitive and physiological constraints. *Trends in Cognitive Sciences*, *5*(1), 36-41.

Gulick, W. L. & Smith, F. L. (1959). The effect of intensity of visual stimulation upon auditory acuity. *The Psychological Record*, *9*, 29-32.

Guthrie, D. & Buchwald, J. S. (1991). Significance testing of difference potentials. *Psychophysiology*, *28*(2), 240-244.

Harrington, L. K. & Peck, C. K. (1998). Spatial disparity affects visual-auditory interactions in human sensorimotor processing. *Experimental Brain Research*, *122*, 247-252.

Harris, L. R. (1980). The superior colliculus and movements of the head and eyes in cats. *Journal of Physiology*, *300*, 367-391.

Harris, L. R., Blakemore, C. & Donaghy, M. (1980). Integration of visual and auditory space in the mammalian superior colliculus. *Nature*, *288*(5786), 59-66.

Hartmann, G. W. (1933). II Changes in visual acuity through simultaneous stimlulation of other sense organs. *Journal of Experimental Psychology*, *16*(3), 393-407.

Hartmann, G. W. (1934). The facilitating effect of strong general illumination upon the discrimination of pitch and intensity differences. *Journal of Experimental Psychology*, *17*(6), 813-822.

Hawkins, H. L. & Presson, J. (1986). Auditory information processing. dans K. Boff & L. Kaufman (Eds.), *Handbook of perception and human performance* (p. 1-64). New York : John Wiley & Sons.

Haxby, J. V., Horwitz, B., Ungerleider, L. G., Maisog, J. M., Pietrini, P. & Grady, C. L. (1994). The functional organization of human extrastriate cortex : a PET-rCBF study of selective attention to faces and locations. *The Journal of Neuroscience*, *14*(11), 6336-6353.

Hershenson, M. (1962). Reaction time as a measure of intersensory facilitation. *Journal of Experimental Psychology*, *63*, 289-293.

Heslenfeld, D. J. (2003). Visual mismatch negativity. dans J. Polich (Ed.), *Detection of change : event-related potential and fMRI findings* (p. 41-60). Dordrecht : Kluwer Academic Publishers.

Hickok, G. & Poeppel, D. (2004). Dorsal and ventral streams : a framework for understanding aspects of the functional anatomy of language. *Cognition*, *92*(1-2), 67-99.

Hietanen, J. K., Leppänen, J. M. & Illi, M. (2004). Evidence for the integration of audiovisual emotional information at the perceptual level of processing. *European Journal of Cognitive Psychology*, *16*(6), 769-790.

Hillyard, S. A., Teder-Sälejärvi, W. A. & Munte, T. F. (1998). Temporal dynamics of early perceptual processing. *Current Opinion in Neurobiology*, *8*(2), 202-210.

Hirsh, I. J. & Sherrick, C. E. (1961). Perceived order in different sense modalities. *Journal of Experimental Psychology*, *62*(5), 423-432.

Hishida, R., Hoshino, K., Kudoh, M., Norita, M. & Shibuki, K. (2003). Anisotropic functional connections between the auditory cortex and area 18a in rat cerebral slices. *Neuroscience Research*, *46*(2), 171-182.

Holcomb, P. J. & Anderson, J. E. (1993). cross-modal semantic priming - a time-course analysis using event-related brain potentials. *Language and Cognitive Processes*, *8*, 379-411.

Holcomb, P. J. & Neville, H. J. (1990). Auditory and Visual Semantic Priming in Lexical Decision : A Comparison Using Event-Related Brain Potentials. *Language and*

Cognitive Processes, *5*, 281-312.

Horn, G. & Hill, R. M. (1966). Responsiveness to sensory stimulation of units in the superior colliculus and subjacent tectotegmental regions of the rabbit. *Experimental Neurology*, *14*(2), 199-223.

Horvath, J., Czigler, I., Sussman, E. & Winkler, I. (2001). Simultaneously active pre-attentive representations of local and global rules for sound sequences in the human brain. *Cognitive Brain Research*, *12*(1), 131-144.

Hotta, T. & Kameda, K. (1963). Interaction between somatic and visual or auditory responses in the thalamus of the cat. *Experimental Neurology*, *8*, 1-13.

Howarth, C. I. & Treisman, M. (1958). Lowering of an auditory threshold produced by a light signal occurring after the threshold stimulus. *Nature*, *182*(4642), 1093-1094.

Hubbard, T. L. (1996). Synesthesia-like mappings of lightness, pitch, and melodic interval. *The American Journal of Psychology*, *109*, 219-238.

Hughes, H. C., Nelson, M. D. & Aronchick, D. M. (1998). Spatial characteristics of visual-auditory summation in human saccades. *Vision Research*, *38*, 3955-3963.

Hughes, H. C., Reuter-Lorenz, P. A., Nozawa, G. & Fendrich, R. (1994). Visual-Auditory Interactions in Sensorimotor Processing - Saccades Versus Manual Responses. *Journal of Experimental Psychology : Human Perception and Performance*, *20*, 131-153.

Jaaskelainen, I. P., Ojanen, V., Ahveninen, J., Auranen, T., Levänen, S., Möttönen, R. et coll. (2004). Adaptation of neuromagnetic N1 responses to phonetic stimuli by visual speech in humans. *Neuroreport*, *15*(18), 2741-2744.

Jabbur, S. J., Atweh, S. F., To'mey, G. F. & Banna, N. R. (1971). Visual and auditory inputs into the cuneate nucleus. *Science*, *174*(14), 1146-1147.

Jack, C. E. & Thurlow, W. R. (1973). Effects of degree of visual association and angle of displacement on the "ventriloquism" effect. *Perceptual and Motor Skills*, *37*(3), 967-979.

Jackson, C. V. (1953). Visual factors in auditory localization. *Quarterly Journal of Experimental Psychology*, *5*, 52-65.

Jay, M. F. & Sparks, D. L. (1984). Auditory receptive fields in primate superior colliculus shift with changes in eye position. *Nature*, *309*, 345-347.

Jay, M. F. & Sparks, D. L. (1987). Sensorimotor integration in the primate superior colliculus. II. Coordinates of auditory signals. *Journal of Neurophysiology*, *57*(1), 35-55.

Jiang, W., Jiang, H. & Stein, B. E. (2002). Two corticotectal areas facilitate multisensory orientation behavior. *Journal of Cognitive Neuroscience*, *14*(8), 1240-1255.

Jiang, W. & Stein, B. E. (2003). Cortex controls multisensory depression in superior colliculus. *Journal of Neurophysiology*, *90*(4), 2123-2135.

Jiang, W., Wallace, M. T., Jiang, H., Vaughan, W. & Stein, B. E. (2001). Two cortical areas mediate multisensory integration in superior colliculus neurons. *Journal of Neurophysiology*, *85*, 506-522.

John, I. D. (1964). The role of extraneaous stimuli in responsiveness to signals : refractoriness or facilitation ? *Australian Journal of Psychology*, *16*, 97-96.

Johnson, H. M. (1920). The dynamogenic influence of light on tactile discrimination. *Psychobiology*, *2*, 351-374.

Jones, E. G. (2001). The thalamic matrix and thalamocortical synchrony. *Trends in Neuroscience*, *24*(10), 595-601.

Jones, E. G. & Powell, T. P. S. (1970). An anatomical study of converging sensory pathways within the cerebral cortex of the monkey. *Brain*, *93*, 793-820.

Jones, J. A. & Callan, D. E. (2003). Brain activity during audiovisual speech perception : an fMRI study of the McGurk effect. *Neuroreport*, *14*(8), 1129-1133.

Jones, J. A. & Jarick, M. (2006). Multisensory integration of speech signals : the relationship between space and time. *Experimental Brain Research*, *174*(3), 588-594.

Jones, J. A. & Munhall, K. G. (1997). The effects of separating auditory and visual sources on audiovisual integration of speech. *Canadian Acoustics*, *25*(4), 13-19.

Karlovich, R. S. (1968). Sensory interaction : perception of loudness during visual stimulation. *The Journal of the Acoustical Society of America*, *44*(2), 570-575.

Karlovich, R. S. (1969). Auditory thresholds during stroboscopic visual stimulation. *The Journal of the Acoustical Society of America*, *45*(6), 1470-1473.

Kawashima, R., O'Sullivan, B. T. & Roland, P. E. (1995). Positron-emission tomography studies of cross-modality inhibition in selective attentional tasks : closing the "mind's eye". *Proceedings of The National Academy of Science*, *92*, 5969-5972.

Kenemans, J. L., Jong, T. G. & Verbaten, M. N. (2003). Detection of visual change : mismatch or rareness ? *Neuroreport*, *14*(9), 1239-1242.

Kim, J. & Davis, C. (2003). Hearing foreign voices : does knowing what is said affect visual-masked-speech detection ? *Perception*, *32*(1), 111-120.

Kim, J. & Davis, C. (2004). Investigating the audiovisual speech detection advantage. *Speech Communication*, *44*, 19-30.

King, A. J. & Palmer, A. R. (1983). Cells responsive to free-field auditory stimuli in guinea-pig superior colliculus : distribution and response properties. *Journal of Physiology*, *342*, 361-381.

King, A. J. & Palmer, A. R. (1985). Integration of visual and auditory information in bimodal neurons in the guinea-pig superior colliculus. *Experimental Brain Research*, *60*, 492-500.

Kirchner, H. & Colonius, H. (2005). Interstimulus contingency facilitates saccadic responses in a bimodal go/no-go task. *Cognitive Brain Research*, *25*(1), 261-272.

Klemm, O. (1909). Lokalisation von Sinneseindrüchen bei disparaten Nebenreizen. *Psychologische Studien (Wundt)*, *5*, 73-161.

Klucharev, V., Möttönen, R. & Sams, M. (2003). Electrophysiological indicators of phonetic and non-phonetic multisensory interactions during audiovisual speech perception. *Cognitive Brain Research*, *18*(1), 65-75.

Knox, G. W. (1945). Investigation of flicker and fusion : III. Effect of audio stimulations on visual critical flicker frequency. *Journal of General Psychology*, *33*, 139-143.

Knudsen, E. I. (1982). Auditory and visual maps of space in the optic tectum of the owl. *The Journal of Neuroscience*, *2*(9), 1177-1194.

Komura, Y., Tamura, R., Uwano, T., Nishijo, H. & Ono, T. (2005). Auditory thalamus integrates visual inputs into behavioral gains. *Nature Neuroscience*, *8*(9), 1203-1209.

Korzyukov, O. A., Winkler, I., Gumenyuk, V. I. & Alho, K. (2003). Processing abstract auditory features in the human auditory cortex. *Neuroimage*, *20*(4), 2245-2258.

Kravkov, S. W. (1934). Changes of visual acuity in one eye under the influence of the illumination of the other or of acoustic stimuli. *Journal of Experimental Psychology*, *17*(6), 805-812.

Kravkov, S. W. (1936). The influence of sound upon light and color sensitivity of the eye. *Acta Ophtalmologica*, *14*, 348-360.

Kropotov, J. D., Alho, K., Näätänen, R., Ponomarev, V. A., Kropotova, O. V., Anichkov, A. D. et coll. (2000). Human auditory-cortex mechanisms of preattentive sound discrimination. *Neuroscience Letters*, *280*(2), 87-90.

Lachaux, J.-P., Rudrauf, D. & Kahane, P. (2003). Intracranial EEG and human brain mapping. *Journal of Physiology (Paris)*, *97*(4-6), 613-628.

Lakatos, P., Chen, C. M., O'Connell, M. N., Mills, A. & Schroeder, C. E. (2007). Neural oscillations and multisensory integration in primary auditory cortex. *Neuron*, *53*(3), 279-292.

Laurienti, P. J., Burdette, J. H., Wallace, M. T., Yen, Y. F., Field, A. S. & Stein, B. E. (2002). Deactivation of sensory-specific cortex by cross-modal stimuli. *Journal of Cognitive Neuroscience*, *14*(3), 420-429.

Laurienti, P. J., Kraft, R. A., Maldjian, J. A., Burdette, J. H. & Wallace, M. T. (2004). Semantic congruence is a critical factor in multisensory behavioral performance. *Experimental Brain Research*, *158*(4), 405-414.

Laurienti, P. J., Perrault, T. J., Stanford, T. R., Wallace, M. T. & Stein, B. E. (2005). On the use of superadditivity as a metric for characterizing multisensory integration in functional neuroimaging studies. *Experimental Brain Research*, *166*(3-4), 289-297.

Laurienti, P. J., Wallace, M. T., Maldjian, J. A., Susi, C. M., Stein, B. E. & Burdette, J. H. (2003). Cross-modal sensory processing in the anterior cingulate and medial prefrontal cortices. *Human Brain Mapping*, *19*(4), 213-223.

Liberman, A. M. & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition*, *21*(1), 1-36.

Liégeois-Chauvel, C., Musolino, A., Badier, J. M., Marquis, P. & Chauvel, P. (1994). Evoked potentials recorded from the auditory cortex in man : evaluation and topography of the middle latency components. *Electroencephalography and Clinical Neurophysiology*, *92*(3), 204-214.

Lisker, L. & Rossi, M. (1992). Auditory and visual cueing of the +/- rounded feature of vowels. *Language and Speech*, *35*(4), 391-417.

Logothetis, N. K. (2003). The underpinnings of the BOLD functional magnetic resonance imaging signal. *The Journal of Neuroscience*, *23*(10), 3963-3971.

Logothetis, N. K., Pauls, J., Augath, M., Trinath, T. & Oeltermann, A. (2001). Neurophysiological investigation of the basis of the fMRI signal. *Nature*, *412*(6843), 150-157.

London, I. D. (1954). Research on sensory interaction in the Soviet Union. *Psychological Bulletin*, *51*(6), 531-568.

Lovelace, C. T., Stein, B. E. & Wallace, M. T. (2003). An irrelevant light enhances auditory detection in humans : a psychophysical analysis of multisensory integration in stimulus detection. *Cognitive Brain Research*, *17*(2), 447-453.

Loveless, N. E., Brebner, J. & Hamilton, P. (1970). Bisensory presentation of information. *Psychological Bulletin*, *73*, 161-199.

Lu, Z. L., Williamson, S. J. & Kaufman, L. (1992a). Behavioral lifetime of human auditory sensory memory predicted by physiological measures. *Science*, *258*, 1668-1670.

Lu, Z. L., Williamson, S. J. & Kaufman, L. (1992b). Human auditory primary and association cortex have differing lifetimes for activation traces. *Brain Research*, *572*, 236-241.

Luck, S. J. (2005). *An introduction to the Event-Related Potential Technique*. Cambridge : The MIT Press.

Ludman, C. N., Summerfield, A. Q., Hall, D., Elliott, M., Foster, J., Hykin, J. L. et coll. (2000). Lip-reading ability and patterns of cortical activation studied using fMRI. *British Journal of Audiology*, *34* (4), 225-230.

Lueck, C. J., Crawford, T. J., Savage, C. J. & Kennard, C. (1990). Auditory-visual interaction in the generation of saccades in man. *Experimental Brain Research*, *82*, 149-157.

Macaluso, E., George, N., Dolan, R., Spence, C. & Driver, J. (2004). Spatial and temporal factors during processing of audiovisual speech : a PET study. *Neuroimage*, *21* (2), 725-732.

MacDonald, J. & McGurk, H. (1978). Visual influences on speech perception processes. *Perception and Psychophysics*, *24* (3), 253-257.

MacLeod, A. & Summerfield, A. Q. (1987). Quantifying the contribution of vision to speech perception in noise. *British Journal of Audiology*, *21* (2), 131-141.

MacSweeney, M., Amaro, E., Calvert, G. A., Campbell, R., David, A. S., McGuire, P. K. et coll. (2000). Silent speechreading in the absence of scanner noise : an event-related fMRI study. *Neuroreport*, *11*, 1729-1733.

MacSweeney, M., Calvert, G. A., Campbell, R., McGuire, P. K., David, A. S., Williams, S. C. et coll. (2002). Speechreading circuits in people born deaf. *Neuropsychologia*, *40* (7), 801-807.

MacSweeney, M., Campbell, R., Calvert, G. A., McGuire, P. K., David, A. S., Suckling, J. et coll. (2001). Dispersed activation in the left temporal cortex for speech-reading in congenitally deaf people. *Philosophical Transactions of the Royal Society of London. Series B : Biological Sciences*, *268*, 451-447.

Magariños-Ascone, C., Garcia-Austt, E. & Buno, W. (1994). Polymodal sensory and motor convergence in substantia nigra neurons of the awake monkey. *Brain Research*, *646* (2), 299-302.

Maier, B., Bevan, W. & Behar, I. (1961). The effect of auditory stimulation upon the critical flicker frequency for different regions of the visible spectrum. *The American Journal of Psychology*, *74*, 67-73.

Mangun, G. R. (1995). Neural mechanisms of visual selective attention. *Psychophysiology*, *32*, 4-18.

Manly, B. J. F., McAlevey, L. & Stevens, D. (1986). A randomization procedure for comparing group means on multiple measurements. *British Journal of Mathematical and Statistical Psychology*, *39*, 183-189.

Marks, L. E. (1974). On associations of light and sound : The mediation of birghtness, pitch and loudness. *The American Journal of Psychology*, *87*, 173-188.

Marks, L. E. (1975). On colored-hearing synesthesia : Crossmodal translations of sensory

dimensions. *Psychological Bulletin*, *82*(3), 303-331.

Marks, L. E. (1987). On cross-modal similarity : auditory-visual interactions in speeded discrimination. *Journal of Experimental Psychology : Human Perception and Performance*, *13*(3), 384-394.

Marks, L. E. (1989). On cross-modal similarity : the perceptual structure of pitch, loudness, and brightness. *Journal of Experimental Psychology : Human Perception and Performance*, *15*(3), 586-602.

Maruyama, K. (1959). Effect of intersensory tone stimulation on absolute light threshold. *Tohoku Psychologica Folia*, *17*, 51-81.

Maruyama, K. (1961). "Contralateral relationship" between the ears and the halves of the visual field in sensory interaction. *Tohoku Psychologica Folia*, *19*, 81-92.

Massaro, D. W. (1987). Speech Perception by Ear and Eye. dans B. Dodd & R. Campbell (Eds.), *Hearing by eye : The psychology of lipreading.* (p. 53-83). London : Lawrence Erlbaum Associates.

Massaro, D. W. (1993). Perceiving asynchronous bimodal speech in consonant-vowel and vowel syllables. *Speech Communication*, *13*, 127-134.

Massaro, D. W. & Cohen, M. M. (1983). Evaluation and integration of visual and auditory information in speech perception. *Journal of Experimental Psychology : Human Perception and Performance*, *9*(5), 753-771.

Massaro, D. W., Cohen, M. M. & Smeele, P. M. T. (1996). Perception of asynchronous and conflicting visual and auditory speech. *The Journal of the Acoustical Society of America*, *100*(3), 1777-1786.

Massaro, D. W. & Egan, P. B. (1996). Perceiving affect from the voice and the face. *Psychonomic Bulletin & Review*, *3*(2), 215-221.

Mathiak, K., Hertrich, I., Zvyagintsev, M., Lutzenberger, W. & Ackermann, H. (2005). Selective influences of cross-modal spatial-cues on preattentive auditory processing : a whole-head magnetoencephalography study. *Neuroimage*, *28*(3), 627-634.

McDonald, J. J., Teder-Sälejärvi, W. A., Heraldez, D. & Hillyard, S. A. (2001). Electrophysiological evidence for the "missing link" in crossmodal attention. *Canadian Journal of Experimental Psychology*, *55*, 141-149.

McEvoy, L., Levänen, S. & Loveless, N. E. (1997). Temporal characteristics of auditory sensory memory : Neuromagnetic evidence. *Psychophysiology*, *34*, 308-316.

McGrath, M. & Summerfield, A. Q. (1985). Intermodal timing relations and audio-visual speech recognition by normal-hearing adults. *The Journal of the Acoustical Society of America*, *77(2)*, 678-685.

McGurk, H. & McDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*, 746-748.

Melara, R. D. (1989). Dimensional interaction between color and pitch. *Journal of Experimental Psychology : Human Perception and Performance*, *15,1*, 69-79.

Melara, R. D. & Marks, L. E. (1990). Processes underlying dimensional interactions : correspondences between linguistic and nonlinguistic dimensions. *Memory and Cognition*, *18*(5), 477-95.

Melara, R. D. & O'Brien, T. P. (1987). Interaction between synesthetically corresponding dimensions. *Journal of Experimental Psychology : General*, *116*(4), 323-326.

Melara, R. D. & O'Brien, T. P. (1990). Effects of cuing on cross-modal congruity. *Journal*

of Memory and Language, *29*(6), 655-686.

Meredith, M. A. (1999). The frontal eye fields target multisensory neurons in cat superior colliculus. *Experimental Brain Research*, *128*(4), 460-470.

Meredith, M. A., Nemitz, J. W. & Stein, B. E. (1987). Determinants of multisensory integration in superior colliculus neurons. 1. Temporal factors. *The Journal of Neuroscience*, *10*, 3215-3229.

Meredith, M. A. & Stein, B. E. (1983). Interactions among converging sensory inputs in the superior colliculus. *Science*, *221*, 389-391.

Meredith, M. A. & Stein, B. E. (1985). Descending efferents from superior colliculus relay integrated multisensory information. *Science*, *227*, 657-659.

Meredith, M. A. & Stein, B. E. (1986a). Spatial factors determine the activity of multisensory neurons in cat superior colliculus. *Brain Research*, *365*, 350-354.

Meredith, M. A. & Stein, B. E. (1986b). Visual, auditory, and somatosensory convergence on cells in superior colliculus results in multisensory integration. *Journal of Neurophysiology*, *56*, 640-662.

Meredith, M. A., Wallace, M. T. & Stein, B. E. (1992). Visual, auditory and somatosensory convergence in output neurons of the cat superior colliculus : multisensory properties of the tecto-reticulo-spinal projection. *Experimental Brain Research*, *88*, 181-186.

Mesulam, M. M. (1998). From sensation to cognition. *Brain*, *121*(6), 1013-1052.

Miki, K., Watanabe, S. & Kakigi, R. (2004). Interaction between auditory and visual stimulus relating to the vowel sounds in the auditory cortex in humans : a magnetoencephalographic study. *Neuroscience Letters*, *357*(3), 199-202.

Miller, J. O. (1982). Divided attention : Evidence for coactivation with redundant signals. *Cognitive Psychology*, *14*, 247-279.

Miller, J. O. (1986). Time course of coactivation in bimodal divided attention. *Perception and Psychophysics*, *40*(5), 331-343.

Miller, J. O. (1991). Channel interaction and the redundant targets effect in bimodal divided attention. *Journal of Experimental Psychology : Human Perception and Performance*, *17*, 160-169.

Miller, J. O. & Lopes, A. (1991). Bias produced by fast guessing in distribution-based tests of race models. *Perception and Psychophysics*, *50*(6), 584-590.

Miller, J. O. & Ulrich, R. (2003). Simple reaction time and statistical facilitation : a parallel grains model. *Cognitive Psychology*, *46*(2), 101-151.

Miller, J. O., Ulrich, R. & Lamarre, Y. (2001). Locus of the redundant-signals effect in bimodal divided attention : a neurophysiological analysis. *Perception and Psychophysics*, *63*(3), 555-562.

Miller, L. M. & D'Esposito, M. (2005). Perceptual fusion and stimulus coincidence in the cross-modal integration of speech. *The Journal of Neuroscience*, *25*(25), 5884-5893.

Molholm, S., Ritter, W., Javitt, D. C. & Foxe, J. J. (2004). Multisensory Visual-Auditory Object Recognition in Humans : a High-density Electrical Mapping Study. *Cerebral Cortex*, *14*(4), 452-465.

Molholm, S., Ritter, W., Murray, M. M., Javitt, D. C., Schroeder, C. E. & Foxe, J. J. (2002). Multisensory auditory-visual interactions during early sensory processing in humans : a high-density electrical mapping study. *Cognitive Brain Research*, *14*(1),

115-128.

Molholm, S., Sehatpour, P., Mehta, A. D., Shpaner, M., Gomez-Ramirez, M., Ortigue, S. et coll. (2006). Audio-visual multisensory integration in superior parietal lobule revealed by human intracranial recordings. *Journal of Neurophysiology*, *96*(2), 721-729.

Moore, E. J. & Karlovich, R. S. (1970). Auditory thresholds during visual stimulation as a function of signal bandwidth. *The Journal of the Acoustical Society of America*, *47*(2), 659-660.

Morais, J. (1975). The effect of ventriloquism on the right-side advantage for verbal material. *Cognition*, *3*(2), 127-139.

Mordkoff, J. T. & Yantis, S. (1991). An interactive race model of divided attention. *Journal of Experimental Psychology : Human Perception and Performance*, *17*, 520-538.

Morrell, F. (1972). Visual system's view of acoustic space. *Nature*, *238*, 44-46.

Morrell, L. K. (1967). Intersensory facilitation of reaction time. *Psychonomic Science*, *8*(2), 77-78.

Morrell, L. K. (1968a). Cross-modality effects upon choice reaction time. *Psychonomic Science*, *11*(4), 129-130.

Morrell, L. K. (1968b). Sensory interactions : evoked potentials observations in man. *Experimental Brain Research*, *6*, 146-155.

Morrell, L. K. (1968c). Temporal characteristics of sensory interaction in choice reaction time. *Journal of Experimental Psychology*, *77*(1), 14-18.

Morton, J. (1967). Comments on "Interaction of the Auditory and Visual Sensory Modalities". *The Journal of the Acoustical Society of America*, *42*(6), 1342.

Möttönen, R., Krause, C. M., Tiippana, K. & Sams, M. (2002). Processing of changes in visual speech in the human auditory cortex. *Cognitive Brain Research*, *13*(3), 417-425.

Möttönen, R., Schurmann, M. & Sams, M. (2004). Time course of multisensory interactions during audiovisual speech perception in humans : a magnetoencephalographic study. *Neuroscience Letters*, *363*(2), 112-115.

Moul, E. R. (1930). an experimental study of visual and auditory thickness. *The American Journal of Psychology*, *42*, 544-560.

Mudd, S. A. (1963). Spatial stereotypes of four dimensions of pure tone. *Journal of Experimental Psychology*, *66*, 347-352.

Muller-Gass, A., Stelmack, R. M. & Campbell, K. B. (2006). The effect of visual task difficulty and attentional direction on the detection of acoustic change as indexed by the Mismatch Negativity. *Brain Research*, *1078*(1), 112-130.

Mulligan, R. M. & Shaw, M. L. (1980). Multimodal signal detection : independent decisions vs. integration. *Perception and Psychophysics*, *28*(5), 471-478.

Mulvenna, C. M. & Walsh, V. (2006). Synaesthesia : supernormal integration ? *Trends in Cognitive Sciences*, *10*(8), 350-352.

Munhall, K. G., Gribble, P., Sacco, L. & Ward, M. (1996). Temporal constraints on the McGurk effect. *Perception and Psychophysics*, *58*, 351-362.

Munhall, K. G., Kroos, C., Jozan, G. & Vatikiotis-Bateson, E. (2004). Spatial frequency requirements for audiovisual speech perception. *Perception and Psychophysics*, *66*(4), 574-583.

Murata, K., Cramer, H. & Rita, P. Bach-y. (1965). Neuronal convergence of noxious, acoustic and visual stimluli in the visual cortex of the cat. *Journal of Neurophysiology*, *28*, 1223-1240.

Musacchia, G., Sams, M., Nicol, T. & Kraus, N. (2006). Seeing speech affects acoustic information processing in the human brainstem. *Experimental Brain Research*, *168*(1-2), 1-10.

Myers, A. K., Cotton, B. & Hilp, H. A. (1981). Matching the rate of concurrent tone bursts and light flashes as a function of flash surround luminance. *Perception and Psychophysics*, *30*(1), 33-38.

Näätänen, R. (1992). *Attention and Brain Function.* Hillsdale : LEA, Inc.

Näätänen, R., Gaillard, A. W. K. & Mantysalo, S. (1978). Early selective-attention effect on evoked potential reinterpreted. *Acta Psychologica*, *42*, 313-329.

Näätänen, R. & Picton, T. W. (1987). The N1 wave of the Human electric and magnetic response to sound : a review and an analysis of the component structure. *Psychophysiology*, *24*, 375-425.

Näätänen, R., Tervaniemi, M., Sussman, E., Paavilainen, P. & Winkler, I. (2001). "Primitive intelligence" in the auditory cortex. *Trends in Neuroscience*, *24*(5), 283-288.

Näätänen, R. & Winkler, I. (1999). The concept of auditory stimulus representation in cognitive neuroscience. *Psychological Bulletin*, *125*(6), 826-859.

Nagy, A., Eordegh, G., Paroczy, Z., Markus, Z. & Benedek, G. (2006). Multisensory integration in the basal ganglia. *European Journal of Neuroscience*, *24*(3), 917-924.

Nagy, A., Paroczy, Z., Norita, M. & Benedek, G. (2005). Multisensory responses and receptive field properties of neurons in the substantia nigra and in the caudate nucleus. *European Journal of Neuroscience*, *22*(2), 419-424.

Neely, K. K. (1956). Effect of Visual Factors on the Intelligibility of Speech. *The Journal of the Acoustical Society of America*, *28*(6), 1275-1277.

Nickerson, R. S. (1973). Intersensory facilitation of reaction time : energy summation or preparation enhancement ? *Psychological Review*, *80*, 489-509.

Noesselt, T., Fendrich, R., Bonath, B., Tyll, S. & Heinze, H. J. (2005). Closer in time when farther in space - Spatial factors in audiovisual temporal integration. *Cognitive Brain Research*, *25*(2), 443-458.

Nousak, J. M., Deacon, D., Ritter, W. & Vaughan, H. G. (1996). Storage of information in transient auditory memory. *Cognitive Brain Research*, *4*(4), 305-317.

Novitski, N., Tervaniemi, M., Huotilainen, M. & Näätänen, R. (2004). Frequency discrimination at different frequency levels as indexed by electrophysiological and behavioral measures. *Cognitive Brain Research*, *20*(1), 26-36.

Nyman, g., Alho, K., laurinen, P., Paavilainen, P., Radil, T., Reinikainen, K. et coll. (1990). Mimatch Negativity (MMN) for sequences of auditory and visual stimuli : Evidence for a mechanism specific to the auditory modality. *Electroencephalography and Clinical Neurophysiology*, *77*, 436-444.

Odgaard, E. C., Arieh, Y. & Marks, L. E. (2003). Cross-modal enhancement of perceived brightness : sensory interaction versus response bias. *Perception and Psychophysics*, *65*(1), 123-132.

Odgaard, E. C., Arieh, Y. & Marks, L. E. (2004). Brighter noise : sensory enhance-

ment of perceived loudness by concurrent visual stimulation. *Cognitive Affective and Behavioral Neuroscience*, *4*(2), 127-132.

O'Hare, J. J. (1956). Intersensory effect of Visual stimluli on the minimum Audible Threshold. *The Journal of General Psychology*, *54*, 167-170.

Ojanen, V., Möttönen, R., Pekkola, J., Jaaskelainen, I. P., Joensuu, R., Autti, T. et coll. (2005). Processing of audiovisual speech in Broca's area. *Neuroimage*, *25*(2), 333-338.

Olivetti Belardinelli, M., Sestieri, C., Di Matteo, R., Delogu, F., Del Gratta, C., Ferreti, A. et coll. (2004). Audio-visual corssmodal interactions in environmental perception : an fMRI investigation. *Cognitive Processing*, *5*, 167-174.

Olson, I. R., Gatenby, J. C. & Gore, J. C. (2002). A comparison of bound and unbound audio-visual information processing in the human cerebral cortex. *Cognitive Brain Research*, *14*(1), 129-138.

Osborn, W. C., Sheldon, R. W. & Baker, R. A. (1963). Vigilance performance under conditions of redundant and non redundant signal presentation. *Journal of Applied Psychology*, *47*, 130-134.

Otten, L. J., Alain, C. & Picton, T. W. (2000). Effects of visual attentional load on auditory processing. *Neuroreport*, *11*(4), 875-880.

Paavilainen, P., Simola, J., Jaramillo, M., Näätänen, R. & Winkler, I. (2001). Preattentive extraction of abstract feature conjunctions from auditory stimulation as reflected by the mismatch negativity. *Psychophysiology*, *38*, 359-365.

Pandey, P. C., Kunov, H. & Abel, S. M. (1986). Disruptive effects of auditory signal delay on speech perception with lipreading. *The Journal of Auditory Research*, *26*, 27-41.

Pandya, D. N., Hallett, M. & Kmukherjee, S. K. (1969). Intra- and interhemispheric connections of the neocortical auditory system in the rhesus monkey. *Brain Research*, *14*(1), 49-65.

Pandya, D. N. & Seltzer, B. (1982). Association areas of the cerebral cortex. *Trends in Neuroscience*, *5*, 386-390.

Paré, M. A., Richler, R. C., ten Hove, M. & Munhall, K. G. (2003). Gaze behavior in audiovisual speech perception : the influence of ocular fixations on the McGurk effect. *Perception and Psychophysics*, *65*(4), 553-567.

Patching, G. R. & Quinlan, P. T. (2002). Garner and congruence effects in the speeded classification of bimodal signals. *Journal of Experimental Psychology : Human Perception and Performance*, *28*(4), 755-775.

Paulesu, E., Perani, D., Blasi, V., Silani, G., Borghese, N. A., De Giovanni, U. et coll. (2003). A functional-anatomical model for lipreading. *Journal of Neurophysiology*, *90*(3), 2005-2013.

Pazo-Alvarez, P., Amenedo, E. & Cadaveira, F. (2004). Automatic detection of motion direction changes in the human brain. *European Journal of Neuroscience*, *19*(7), 1978-1986.

Pazo-Alvarez, P., Cadaveira, F. & Amenedo, E. (2003). MMN in the visual modality : a review. *Biological Psychology*, *63*(3), 199-236.

Peck, C. K. (1987). Visual-auditory interactions in cat superior colliculus : their role in the control of gaze. *Brain Research*, *420*(1), 162-166.

Pekkola, J., Ojanen, V., Autti, T., Jaaskelainen, I. P., Möttönen, R., Tarkiainen, A. et coll.

(2005). Primary auditory cortex activation by visual speech : an fMRI study at 3 T. *Neuroreport*, *16*(2), 125-128.

Pernier, J. & Bertrand, O. (1997). L'électro- et la magnéto-encéphalographie. dans S. Dehaene (Ed.), *Le cerveau en action "Imagerie cérébrale fonctionnelle en psychologie cognitive"*. Paris : PUF.

Peronnet, F. & Gerin, P. (1972). Potentiels évoqués auditifs et visuels : Topographie et interactions. dans *Activités évoquées et leur conditionnement chez l'homme normal et en pathologie mentale* (35 ed., p. 35-55). Paris : INSERM.

Perrault, T. J., Vaughan, J. W., Stein, B. E. & Wallace, M. T. (2003). Neuron-specific response characteristics predict the magnitude of multisensory integration. *Journal of Neurophysiology*, *90*(6), 4022-4026.

Perrault, T. J., Vaughan, J. W., Stein, B. E. & Wallace, M. T. (2005). Superior colliculus neurons use distinct operational modes in the integration of multisensory stimuli. *Journal of Neurophysiology*, *93*(5), 2575-2586.

Perrin, F., Bertrand, O. & Pernier, J. (1987). Scalp current density mapping : value and estimation from potential data. *IEEE Transactions on Bio-medical Engineering*, *34*(4), 283-288.

Perrin, F., Pernier, J., Bertrand, O. & Echallier, J. F. (1989). Spherical splines for scalp potential and current density mapping. *Electroencephalography and Clinical Neurophysiology*, *72*, 184-187.

Pick, J., H. L., Warren, D. H. & Hay, J. C. (1969). Sensory conflict in judgements of spatial direction. *Perception and Psychophysics*, *6*, 203-205.

Populin, L. C. & Yin, T. C. (2002). Bimodal interactions in the superior colliculus of the behaving cat. *The Journal of Neuroscience*, *22*(7), 2826-2834.

Posner, M. I., Nissen, M. J. & Klein, R. M. (1976). Visual dominance : an information-processing account of its origins and significance. *Psychological Review*, *83*, 157-171.

Pourtois, G., Debatisse, D., Despland, P. A. & de Gelder, B. (2002). Facial expressions modulate the time course of long latency auditory brain potentials. *Cognitive Brain Research*, *14*(1), 99-105.

Pourtois, G., de Gelder, B., Vroomen, J., Rossion, B. & Crommelinck, M. (2000). The time-course of intermodal binding between seeing and hearing affective information. *Neuroreport*, *11*(6), 1329-1333.

Pratt, C. C. (1936). Interaction across modalities : simultaneous stimulation. *Proceedings of The National Academy of Science*, *22*(9), 562-566.

Puce, A. & Allison, T. (1999). Differential processing of mobile and static faces by temporal cortex. *Neuroimage*, *6*, S801.

Puce, A., Allison, T., Bentin, S., Gore, J. C. & McCarthy, G. (1998). Temporal cortex activation in humans viewing eye and mouth movements. *The Journal of Neuroscience*, *18*(6), 2188-2199.

Raab, D. H. (1962). Statistical facilitation of simple reaction times. *Transactions of the New York Academy of Sciences*, *24*, 574-590.

Radeau, M. (1976). L'adaptation au déplacement de l'espace visuel : Revue critique. *Archives de psychologie*, *44*(supp 4), 1-91.

Radeau, M. (1985). Signal intensity, task context, and auditory-visual interactions. *Per-*

ception, *14*(5), 571-577.

Radeau, M. (1992). Cognitive impenetrability in auditory-visual interaction. dans J. Alegria, D. Holender, J. Morais & M. Radeau (Eds.), *Analytic approaches to human cognition* (p. 41-55). Amsterdam : Elsevier Science Publishers.

Radeau, M. (1994a). Auditory-visual spatial interaction and modularity. *Cahiers de psychologie cognitive*, *13*, 3-51.

Radeau, M. (1994b). Ventriloquism against audio-visual speech - Or, where japanese-speaking barn owls might help. *Cahiers de psychologie cognitive*, *13*, 124-140.

Radeau, M. & Bertelson, P. (1974). The after-effects of ventriloquism. *Quarterly Journal of Experimental Pscychology*, *26*(1), 63-71.

Radeau, M. & Bertelson, P. (1977). Adaptation to auditory-visual discordance and ventriloquism in semi-realistic situations. *Perception and Psychophysics*, *22*, 137-146.

Radeau, M. & Bertelson, P. (1978). Cognitive factors and adaptation to auditory-visual discordance. *Perception and Psychophysics*, *23*, 341-343.

Radeau, M. & Bertelson, P. (1987). Auditory-visual interaction and the timing of inputs. Thomas (1941) revisited. *Psychological Research*, *49*(1), 17-22.

Ratcliff, R. (1979). Group reaction time distributions and an analysis of distribution statistics. *Psychological Bulletin*, *86*(3), 446-461.

Regan, D. & Spekreijse, H. (1977). Auditory-visual interactions and the correspondence between perceived auditory space and perceived visual space. *Perception*, *6*(2), 133-138.

Reisberg, D., McLean, J. & Goldfield, A. (1987). Easy to hear but hard to understand : a lipreading advantage with intact auditory stimuli. dans B. Dodd & R. Campbell (Eds.), *Hearing by eye : The psychology of lipreading.* (p. 93-113). London : Lawrence Erlbaum Associates.

Rich, A. N. & Mattingley, J. B. (2002). Anomalous perception in synaesthesia : a cognitive neuroscience perspective. *Nature Reviews. Neuroscience*, *3*(1), 43-52.

Robert-Ribes, J., Schwartz, J. L., Lallouache, T. & Escudier, P. (1998). Complementarity and synergy in bimodal speech : Auditory, visual, and audio-visual identification of French oral vowels in noise. *The Journal of the Acoustical Society of America*, *103*(6), 3677-3689.

Roberts, M. (1987). Audio-visual speech perception and selective adaptation. dans B. Dodd & R. Campbell (Eds.), *Hearing by eye : The psychology of lipreading.* (p. 87-96). London : Lawrence Erlbaum Associates.

Roberts, M. & Summerfield, A. Q. (1981). Audiovisual presentation demonstrates that selective adaptation to speech is purely auditory. *Perception and Psychophysics*, *30*(4), 309-314.

Rockland, K. S. & Ojima, H. (2003). Multisensory convergence in calcarine visual areas in macaque monkey. *International Journal of Psychophysiology*, *50*(1-2), 19-26.

Roffler, S. K. & Butler, R. A. (1967). Localization of tonal stimuli in the vertical plane. *The Journal of the Acoustical Society of America*, *43*(6), 1260-1266.

Rosburg, T. (2003). Left hemispheric dipole locations of the neuromagnetic mismatch negativity to frequency, intensity and duration deviants. *Cognitive Brain Research*, *16*(1), 83-90.

Rosenblum, L. D. & Fowler, C. A. (1991). Audiovisual investigation of the loudness-effort effect for speech and nonspeech events. *Journal of Experimental Psychology : Human Perception and Performance*, *17*(4), 976-985.

Rosenblum, L. D. & Saldaña, H. M. (1992). Discrimination tests of visually influenced syllables. *Perception and Psychophysics*, *52*(4), 461-473.

Rosenblum, L. D. & Saldaña, H. M. (1996). An audiovisual test of kinematic primitives for visual speech perception. *Journal of Experimental Psychology : Human Perception and Performance*, *22*, 318-331.

Ross, L. A., Saint-Amour, D., Leavitt, V. M., Javitt, D. C. & Foxe, J. J. (sous presse). Do You See What I Am Saying ? Exploring Visual Enhancement of Speech Comprehension in Noisy Environments. *Cerebral Cortex*.

Rudmann, D. S., McCarley, J. S. & Kramer, A. F. (2003). Bimodal displays improve speech comprehension in environments with multiple speakers. *Human Factors*, *45*(2), 329-336.

Rutledge, L. T. (1963). Interactions of Peripherally and Centrally Originating Input to Association Cortex. *Electroencephalography and Clinical Neurophysiology*, *15*, 958-968.

Ryan, T. A. (1940). Interrelations of the sensory systems in perception. *Psychological Bulletin*, *37*(9), 659-698.

Saint-Amour, D., De Sanctis, P., Molholm, S., Ritter, W. & Foxe, J. J. (2007). Seeing voices : High-density electrical mapping and source-analysis of the multisensory mismatch negativity evoked during the McGurk illusion. *Neuropsychologia*, *45*, 587-597.

Saldaña, H. M. & Rosenblum, L. D. (1993). Visual influences on auditory pluck and bow judgments. *Perception and Psychophysics*, *54*(3), 406-416.

Saldaña, H. M. & Rosenblum, L. D. (1994). Selective adaptation in speech perception using a compelling audiovisual adaptor. *The Journal of the Acoustical Society of America*, *95*(6), 3658-3661.

Sams, M., Aulanko, R., Hamalainen, H., Hari, R., Lounasmaa, O. V., Lu, S. T. et coll. (1991). Seeing speech : Visual information from lip movements modifies activity in the human auditory cortex. *Neuroscience Letters*, *127*, 141-145.

Sams, M. & Levänen, S. (1998). A neuromagnetic study of the integration of audiovisual speech in the brain. dans Y. Koga, K. Nagata & K. Hirita (Eds.), *Brain topography today* (p. 47-53). Amsterdam : Elsevier Science.

Sams, M., Manninen, P., Surakka, V., Helin, P. & Kättö, R. (1998). McGurk effect in Finnish syllables, isolated words, and words in sentences : Effects of word meaning and sentence context. *Speech Communication*, *26*, 75-87.

Sanabria, D., Correa, A., Lupianez, J. & Spence, C. (2004). Bouncing or streaming ? Exploring the influence of auditory cues on the interpretation of ambiguous visual motion. *Experimental Brain Research*, *157*(4), 537-541.

Schneider, A. S. & Davis, J. L. (1974). Interactions of the evoked responses to visual, somatic, and auditory stimuli in polysensory areas of the cat cortex. *Physiology & Behavior*, *13*(3), 365-372.

Schneider, G. E. (1969). Two visual systems. *Science*, *163*(870), 895-902.

Schroeder, C. E. & Foxe, J. J. (2002). The timing and laminar profile of converging inputs

to multisensory areas of the macaque neocortex. *Cognitive Brain Research*, *14*(1), 187-198.

Schroeder, C. E., Molholm, S., Lakatos, P., Ritter, W. & Foxe, J. J. (2004). Human-simian correspondence in the early cortical processing of multisensory Cues. *Cognitive Processing*, *5*(3), 140-151.

Schroeder, C. E., Smiley, J., Fu, K. G., McGinnis, T., O'Connell, M. N. & Hackett, T. A. (2003). Anatomical mechanisms and functional implications of multisensory convergence in early cortical processing. *International Journal of Psychophysiology*, *50*(1-2), 5-17.

Schröger, E. (1997). On the detection of auditory deviations : A pre-attentive activation model. *Psychophysiology*, *34*(3), 245-257.

Schröger, E. & Widmann, A. (1998). Speeded responses to audiovisual signal changes result from bimodal integration. *Psychophysiology*, *35*, 755-759.

Schröger, E. & Wolff, C. (1996). Mismatch response of the human brain to changes in sound location. *Neuroreport*, *7*(18), 3005-3008.

Schwartz, J. L. (2003). Why the FMLP should not be applied to McGurk data. dans J. Schwartz, F. Berthommier, M. Cathiard & D. Sodoyer (Eds.), *AVSP* (p. 77-82). Saint-Jorioz.

Schwartz, J. L., Berthommier, F. & Savariaux, C. (2004). Seeing to hear better : evidence for early audio-visual interactions in speech identification. *Cognition*, *93*(2), B69-78.

Schwartz, J. L., Robert-Ribes, J. & Escudier, P. (1998). Ten years after Summerfield : a taxonomy for audio-visual fusion in speech perception. dans R. Campbell, B. Dodd & D. Burnham (Eds.), *Hearing by Eye II* (p. 85-108). Hove : Psychology Press.

Schwarz, W. (1989). A new model to explain the redundant-signals effect. *Perception and Psychophysics*, *46*, 498-500.

Sekiyama, K., Kanno, I., Miura, S. & Sugita, Y. (2003). Auditory-visual speech perception examined by fMRI and PET. *Neuroscience Research*, *47*(3), 277-287.

Sekuler, R., Sekuler, A. B. & Lau, R. (1997). Sounds alter visual motion perception. *Nature*, *385*, 308.

Seltzer, B. & Pandya, D. N. (1978). Afferent cortical connections and architectonics of the superior temporal sulcus and surrounding cortex in the rhesus monkey. *Brain Research*, *149*(1), 1-24.

Senkowski, D., Molholm, S., Gomez-Ramirez, M. & Foxe, J. J. (2006). Oscillatory Beta Activity Predicts Response Speed during a Multisensory Audiovisual Reaction Time Task : A High-Density Electrical Mapping Study. *Cerebral Cortex*, *16*(11), 1556-65.

Serrat, W. D. & Karwoski, T. (1936). An investigation of the effect of auditory stimulation on visual sensitivity. *Journal of Experimental Psychology*, *19*(5), 604-611.

Shams, L., Kamitani, Y. & Shimojo, S. (2000). What you see is what you hear. *Nature*, *408*, 788.

Shams, L., Kamitani, Y., Thompson, S. & Shimojo, S. (2001). Sound alters visual evoked potentials in humans. *Neuroreport*, *12*(17), 3849-3852.

Sheridan, J. A., Cimbalo, R. S., Sills, J. A. & Alluisi, E. A. (1966). Effects of darkness, constant illumination and synchronized photic stimulation on auditory sensitivity to pulsed tones. *Psychonomic Science*, *5*, 311-312.

Shigeno, S. (2002). Anchoring effects in audiovisual speech perception. *The Journal of the Acoustical Society of America*, *111*(6), 2853-2861.

Shipley, T. (1964). Auditory flutter-driving of visual-flicker. *Science*, *145*, 1328-1330.

Simon, J. R. (1982). Effect of any auditory stimulus on the processing of a visual stimulus single- and dual-tasks conditions. *Acta Psychologica*, *51*(1), 61-73.

Simon, J. R. & Craft, J. L. (1970). Effects of an irrelevant auditory stimulus on visual choice reaction time. *Journal of Experimental Psychology*, *86*(2), 272-274.

Soto-Faraco, S., Navarra, J. & Alsius, A. (2004). Assessing automaticity in audiovisual speech integration : evidence from the speeded classification task. *Cognition*, *92*(3), B13-23.

Spence, C. J. & Driver, J. (1997). Audiovisual links in exogenous covert spatial orienting. *Perception and Psychophysics*, *59*, 1-22.

Spinelli, D. N., Starr, A. & Barrett, T. W. (1968). Auditory specificity in unit recordings from cat's visual cortex. *Experimental Neurology*, *22*, 75-84.

Squires, N. K., Donchin, E., Squires, K. C. & Grossberg, S. (1977). Bisensory stimulation : inferring decision-related processes from P300 component. *Journal of Experimental Psychology*, *3*(2), 299-315.

Stanford, T. R., Quessy, S. & Stein, B. E. (2005). Evaluating the operations underlying multisensory integration in the cat superior colliculus. *The Journal of Neuroscience*, *25*(28), 6499-6508.

Stein, B. E. & Gaither, N. S. (1983). Receptive-field properties in reptilian optic tectum : some comparisons with mammals. *Journal of Neurophysiology*, *50*(1), 102-124.

Stein, B. E., Huneycutt, W. S. & Meredith, M. A. (1988). Neurons and behavior : The same rules of multisensory integration apply. *Brain Research*, *448*, 355-358.

Stein, B. E., London, N., Wilkinson, L. K. & Price, D. D. (1996). Enhancement of perceived visual intensity by auditory stimuli : a psychophysical analysis. *Journal of Cognitive Neuroscience*, *8*, 497-506.

Stein, B. E. & Meredith, M. A. (1993). *The merging of the senses* (1 ed.). Cambridge, MA : The MIT Press.

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

Stekelenburg, J. J., Vroomen, J. & de Gelder, B. (2004). Illusory sound shifts induced by the ventriloquist illusion evoke the mismatch negativity. *Neuroscience Letters*, *357*(3), 163-166.

Stewart, D. L. & Starr, A. (1970). Absence of visually influenced cells in auditory cortex of normal and congenitally deaf cats. *Experimental Neurology*, *28*(3), 525-528.

Stoffels, E. J. & van der Molen, M. W. (1988). Effects of visual and auditory noise on visual choice reaction time in a continuous-flow paradigm. *Perception and Psychophysics*, *44*(1), 7-14.

Stoffels, E. J., van der Molen, M. W. & Keuss, P. J. (1985). Intersensory facilitation and inhibition : immediate arousal and location effects of auditory noise on visual choice reaction time. *Acta Psychologica*, *58*(1), 45-62.

Stoffels, E. J., van der Molen, M. W. & Keuss, P. J. (1989). An additive factors analysis of

the effect(s) of location cues associated with auditory stimuli on stages of information processing. *Acta Psychologica*, *70*(2), 161-97.

Sumby, W. H. & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *The Journal of the Acoustical Society of America*, *26*, 212-215.

Summerfield, A. Q. (1979). Use of visual information for phonetic perception. *Phonetica*, *36*(4-5), 314-331.

Summerfield, A. Q. (1987). Some Preliminaries to a Comprehensive Account of Audio-visual Speech Perception. dans B. Dodd & R. Campbell (Eds.), *Hearing by eye : The psychology of lipreading.* (p. 3-52). London : Lawrence Erlbaum Associates.

Summerfield, A. Q. & MacGrath, M. (1984). Detection and resolution of audio-visual incompatibility in the perception of vowels. *The Quarterly Journal of Experimental Psychology. A, Human Experimental Psychology*, *36*(A), 51-74.

Sundara, M., Namasivayam, A. K. & Chen, R. (2001). Observation-execution matching system for speech : a magnetic stimulation study. *Neuroreport*, *12*(7), 1341-1344.

Surakka, V., Tenhunen-Eskelinen, M., Hietanen, J. K. & Sams, M. (1998). Modulation of human auditory information processing by emotional visual stimuli. *Cognitive Brain Research*, *7*, 159-163.

Sussman, E., Gomes, H., Nousak, J. M., Ritter, W. & Vaughan, J., H. G. (1998). Feature conjunctions and auditory sensory memory. *Brain Research*, *793*(1-2), 95-102.

Symons, J. R. (1963). The effect of various heteromodal stimuli on visual sensitivity. *Quarterly Journal of Experimental Psychology*, *15*, 234-251.

Takegata, R., Huotilainen, M., Rinne, T., Näätänen, R. & Winkler, I. (2001). Changes in acoustic features and their conjunctions are processed by separate neuronal populations. *Neuroreport*, *12*(3), 525-529.

Takegata, R., Paavilainen, P., Näätänen, R. & Winkler, I. (1999). Independent processing of changes in auditory single features and feature conjunctions in humans as indexed by the mismatch negativity. *Neuroscience Letters*, *266*(2), 109-112.

Talairach, J. & Szikla, G. (1967). *Atlas d'anatomie stéréotaxique du téléencéphale. Etude anatomo-radiologiques.* Paris : Masson.

Talairach, J. & Tournoux, P. (1988). *Co-Planar Stereotaxic Atlas of the Human Brain.* New York : Thieme Medical Publishers.

Taylor, R. L. (1974). An analysis of sensory interaction. *Neuropsychologia*, *12*, 65-71.

Taylor, R. L. & Campbell, G. T. (1976). Sensory interaction : vision is modulated by hearing. *Perception*, *5*(4), 467-477.

Teder-Sälejärvi, W. A., Di Russo, F., McDonald, J. J. & Hillyard, S. A. (2005). Effects of spatial congruity on audio-visual multimodal integration. *Journal of Cognitive Neuroscience*, *17*(9), 1396-1409.

Teder-Sälejärvi, W. A., McDonald, J. J., Di Russo, F. & Hillyard, S. A. (2002). An analysis of audio-visual crossmodal integration by means of event-related potential (ERP) recordings. *Cognitive Brain Research*, *14*(1), 106-114.

Tervaniemi, M., Maury, S. & Näätänen, R. (1994). Neural representations of abstract stimulus features in the human brain as reflected by the mismatch negativity. *Neuroreport*, *5*(7), 844-846.

Thiel, C. M., Zilles, K. & Fink, G. R. (2004). Cerebral correlates of alerting, orienting

and reorienting of visuospatial attention : an event-related fMRI study. *Neuroimage*, *21*(1), 318-328.

Thomas, G. J. (1941). Experimental Study of the influence of vision on sound localization. *Journal of Experimental Psychology*, *28*, 163-177.

Thompson, R. F., Johnson, R. H. & Hoopes, J. J. (1963). Organization of auditory, somatic sensory, and visual projection to association fields of cerebral cortex in the cat. *Journal of Neurophysiology*, *26*, 343-364.

Thompson, R. F. & Shaw, J. A. (1965). Behavioral correlates of evoked activity recorded from association areas of the cerebral cortex. *Journal of Comparative and Physiological Psychology*, *60*(3), 329-339.

Thompson, R. F., Smith, H. E. & Bliss, D. (1963). Auditory, somatic sensory, and visual response interactions and interrelations in association and primary cortical fields of the cat. *Journal of Neurophysiology*, *26*, 365-378.

Thurlow, W. R. & Jack, C. E. (1973). Certain determinants of the "ventriloquism effect". *Perceptual and Motor Skills*, *36*(3), 1171-1184.

Tiippana, K. & Andersen, T. S. (2004). Visual attention modulates audiovisual speech perception. *European Journal of Cognitive Psychology*, *16*(3), 457-472.

Tiitinen, H., May, P., Reinikainen, K. & Näätänen, R. (1994). Attentive novelty detection in humans is governed by pre-attentive sensory memory. *Nature*, *372*(6501), 90-92.

Todd, J. W. (1912). Reaction to multiple stimuli. *Archives of Psychology*, *3*(25).

Toldi, J., Fehér, O. & Gerő, L. (1980). The existence of two polysensory systems in the suprasylvian gyrus of the cat. *Acta Physiologica Academiae Scientarum Hungaricae*, *55,3*, 181-187.

Townsend, J. T. (1997). Serial Exhaustive Models Can Violate the Race Model Inequality : Implications for Architecture and Capacity. *Psychological Review*, *104*(3), 595-602.

Tuomainen, J., Andersen, T. S., Tiippana, K. & Sams, M. (2005). Audio-visual speech perception is special. *Cognition*, *96*(1), B13-22.

Turatto, M., Benso, F., Galfano, G. & Umilta, C. (2002). Nonspatial attentional shifts between audition and vision. *Journal of Experimental Psychology : Human Perception and Performance*, *28*(3), 628-639.

Ullsperger, P., Erdmann, U., Freude, G. & Dehoff, W. (2006). When sound and picture do not fit : Mismatch negativity and sensory interaction. *International Journal of Psychophysiology*, *59*(1), 3-7.

Valtonen, J., May, P., Makinen, V. & Tiitinen, H. (2003). Visual short-term memory load affects sensory processing of irrelevant sounds in human auditory cortex. *Cognitive Brain Research*, *17*(2), 358-367.

van Wassenhove, V., Grant, K. W. & Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proceedings of The National Academy of Science*, *102*(4), 1181-1186.

van Wassenhove, V., Grant, K. W. & Poeppel, D. (2007). Temporal window of integration in auditory-visual speech perception. *Neuropsychologia*, *45*(3), 598-607.

Vatakis, A. & Spence, C. (2006a). Audiovisual synchrony perception for music, speech, and object actions. *Brain Research*, *1111*(1), 134-142.

Vatakis, A. & Spence, C. (2006b). Audiovisual synchrony perception for speech and music

assessed using a temporal order judgment task. *Neuroscience Letters*, *393*(1), 40-44.

Vaughan, H. G. & Ritter, W. (1970). The sources of auditory evoked responses recorded from the human scalp. *Electroencephalography and Clinical Neurophysiology*, *28*, 360-367.

Vincent, S. B. (1912). The function of the vibrissae in the behavior of the white rat. *Behavior Monographs*, *1*(5).

Vogel, E. K. & Luck, S. J. (2000). The visual N1 component as an index of a discrimination process. *Psychophysiology*, *37*, 190-203.

von Hornbostel, E. M. (1931). Uber Geruchshelligkeit. *Pflügers Archiv für die Gesamte Physiologie des Menschen und der Tiere*, *227*, 517-538.

von Schiller, P. (1935). Interrelation of different senses in perception. *British Journal of Psychology*, *25*, 465-469.

Vroomen, J. & de Gelder, B. (2000). Crossmodal integration : a good fit is no criterion. *Trends in Cognitive Sciences*, *4*(2), 37-38.

Vroomen, J., Driver, J. & de Gelder, B. (2001). Is cross-modal integration of emotional expressions independent of attentional resources ? *Cognitive Affective and Behavioral Neuroscience*, *1*(4), 382-387.

Vroomen, J., Linden, S. van, de Gelder, B. & Bertelson, P. (2007). Visual recalibration and selective adaptation in audiotory-visual speech perception : Contrasting build-up courses. *Neuropsychologia*, *45*(3), 572-577.

Vroomen, J., Linden, S. van, Keetels, M., de Gelder, B. & Bertelson, P. (2004). Selective adaptation and recalibration of auditory speech by lipread information : dissipation. *Speech Communication*, *44*, 55-61.

Walker, J. T., Irion, A. L. & Gordon, D. G. (1981). Simple and contingent aftereffects of perceived duration in vision and audition. *Perception and Psychophysics*, *29*(5), 475-486.

Walker, P. & Smith, S. (1986). The basis of Stroop interference involving the multimodal correlates of auditory pitch. *Perception*, *15*(4), 491-496.

Walker, S., Bruce, V. & Omalley, C. (1995). Facial-identity and facial-speech processing : Familiar faces and voices in the McGurk effect. *Perception and Psychophysics*, *57*, 1124-1133.

Wallace, M. T., Meredith, M. A. & Stein, B. E. (1992). Integration of multiple sensory modalities in cat cortex. *Experimental Brain Research*, *91*, 484-488.

Wallace, M. T., Meredith, M. A. & Stein, B. E. (1993). Converging Influences from Visual, Auditory, and Somatosensory Cortices Onto Output Neurons of the Superior Colliculus. *Journal of Neurophysiology*, *69*, 1797-1809.

Wallace, M. T., Meredith, M. A. & Stein, B. E. (1998). Multisensory integration in the superior colliculus of the alert cat. *Journal of Neurophysiology*, *80*, 1006-1010.

Wallace, M. T. & Stein, B. E. (1994). Cross-modal synthesis in the midbrain depends on input from cortex. *Journal of Neurophysiology*, *71,1*, 429-432.

Walter, W. G. (1964). The convergence and interaction of visual, auditory and tactile responses in human non-specific cortex. *Annals of The New York Academy of Sciences*, *122*, 320-361.

Ward, L. M. (1994). Supramodal and Modality-Specific Mechanisms for Stimulus- Dri-

ven Shifts of Auditory and Visual Attention. *Canadian Journal of Experimental Psychology*, *48*, 242-259.

Ward, L. M., McDonald, J. J. & Lin, D. (2000). On asymmetries in cross-modal spatial attention orienting. *Perception and Psychophysics*, *62*, 1258-1264.

Warren, D. H. (1979). Spatial localization under conflict conditions : is there a single explanation ? *Perception*, *8*(3), 323-337.

Warren, D. H., Welch, R. B. & McCarthy, T. (1981). The role of visual-auditory "compellingness" in the ventriloquism effect : Implications for transitivity among the spatial senses. *Perception and Psychophysics*, *30(6)*, 557-564.

Watanabe, K. & Shimojo, S. (2001). When sound affects vision : effects of auditory grouping on visual motion perception. *Psychological Science*, *12*(2), 109-116.

Watkins, K. E., Strafella, A. P. & Paus, T. (2003). Seeing and hearing speech excites the motor system involved in speech production. *Neuropsychologia*, *41*(8), 989-994.

Watkins, W. H. & Feehrer, C. E. (1965). Acoustic Facilitation of Visual Detection. *Journal of Experimental Psychology*, *70*(3), 332-333.

Weerts, T. C. & Thurlow, W. R. (1971). The effects of eye position and expectation on sound localisation. *Perception and Psychophysics*, *9*, 35-39.

Welch, R. B., DuttonHurt, L. D. & Warren, D. H. (1986). Contributions of audition and vision to temporal rate perception. *Perception and Psychophysics*, *39*(4), 294-300.

Welch, R. B. & Warren, D. H. (1980). Immediate perceptual response to intersensory discrepancy. *Psychological Bulletin*, *88*(3), 638-667.

Welch, R. B. & Warren, D. H. (1986). Intersensory interactions. dans K. Boff, L. Kaufman & J. Thomas (Eds.), *Handbook of Perception and Human Performance, Volume I : Sensory Processes and Perception* (p. 25.1-25.36). New York : Wiley.

Werner, H. (1934). L'unité des sens. *Journal de psychologie*, *31*, 190-205.

Wickelgren, B. G. (1971). Superior colliculus : some receptive field properties of bimodally responsive cells. *Science*, *173*(991), 69-72.

Widmann, A., Kujala, T., Tervaniemi, M., Kujala, A. & Schröger, E. (2004). From symbols to sounds : Visual symbolic information activates sound representations. *Psychophysiology*, *41*(5), 709-715.

Wilkinson, L. K., Meredith, M. A. & Stein, B. E. (1996). The role of anterior ectosylvian cortex in cross-modality orientation and approach behavior. *Experimental Brain Research*, *112*, 1-10.

Wilson, S. M., Saygin, A. P., Sereno, M. I. & Iacoboni, M. (2004). Listening to speech activates motor areas involved in speech production. *Nature Neuroscience*, *7*(7), 701-702.

Windmann, S. (2004). Effects of sentence context and expectation on the McGurk illusion. *Journal of Memory and Language*, *20*, 212-230.

Winkler, I., Cowan, N., Csepe, V., Czigler, I. & Näätänen, R. (1996). Interactions between transient and long-term auditory memory as reflected by the mismatch negativity. *Journal of Cognitive Neuroscience*, *8*, 403-415.

Winkler, I., Czigler, I., Jaramillo, M., Paavilainen, P. & Näätänen, R. (1998). Temporal constraints of auditory event synthesis : evidence from ERPs. *Neuroreport*, *9*, 495-499.

Winkler, I., Czigler, I., Sussman, E., Horvath, J. & Balazs, L. (2005). Preattentive binding

of auditory and visual stimulus features. *Journal of Cognitive Neuroscience*, *17*(2), 320-339.

Winkler, I., Karmos, G. & Näätänen, R. (1996). Adaptive modeling of the unattended acoustic environment reflected in the mismatch negativity event-related potential. *Brain Research*, *742*(1-2), 239-252.

Winkler, I., Paavilainen, P., Alho, K., Reinikainen, K., Sams, M. & Näätänen, R. (1990). The effect of small variation of the frequent auditory stimulus on the event-related brain potential to the infrequent stimulus. *Psychophysiology*, *27,2*, 228-235.

Winkler, I., Paavilainen, P. & Näätänen, R. (1992). Can echoic memory store two traces simultaneously ? A study of event- related brain potentials. *Psychophysiology*, *29*, 337-349.

Winkler, I., Reinikainen, K. & Näätänen, R. (1993). Event-related brain potentials reflect traces of echoic memory in humans. *Perception and Psychophysics*, *53*, 443-449.

Witkin, H. A., Wapner, S. & Leventhal, T. (1952). Sound localization with conflicting visual and auditory cues. *Journal of Experimental Psychology*, *43*(1), 58-67.

Woods, D. L., Alho, K. & Algazi, A. (1992). Intermodal selective attention. I. Effects on event-related potentials to lateralized auditory and visual stimuli. *Electroencephalography and Clinical Neurophysiology*, *82*, 341-355.

Wright, T. M., Pelphrey, K. A., Allison, T., McKeown, M. J. & McCarthy, G. (2003). Polysensory interactions along lateral temporal regions evoked by audiovisual speech. *Cerebral Cortex*, *13*(10), 1034-1043.

Yaka, R., Notkin, N., Yinon, U. & Wollberg, Z. (2000). Visual, auditory, and bimodal activity in the banks of the lateral suprasylvian sulcus in the cat. *Rossiiskii Fiziologicheskii Zhurnal Imeni I.M. Sechenova / Rossiiskaia Akademiia Nauk*, *86*(7), 877-883.

Yumoto, M., Uno, A., Itoh, K., Karino, S., Saitoh, O., Kaneko, Y. et coll. (2005). Audiovisual phonological mismatch produces early negativity in auditory cortex. *Neuroreport*, *16*(8), 803-806.

Yvert, B., Fischer, C., Bertrand, O. & Pernier, J. (2005). Localization of human supratemporal auditory areas from intracerebral auditory evoked potentials using distributed source models. *Neuroimage*, *28*(1), 140-153.

Zhang, P., Chen, X., Yuan, P., Zhang, D. & He, S. (2006). The effect of visuospatial attentional load on the processing of irrelevant acoustic distractors. *Neuroimage*, *33*(2), 715-724.

Zietz, K. (1931). Gegenseitige Beeinflussung von Farb- und Tonerlebsinen : Studien über experimentell erzeugte Synästhesie. *Zeitschrift für Psychologie*, *121*, 257-356.

Résumé Dans le modèle classique d'organisation des systèmes sensoriels, les informations de différentes modalités sont censées converger à des étapes relativement tardives de traitement (après leur analyse dans les cortex sensoriels spécifiques), dans un nombre limité d'aires corticales, dites polysensorielles associatives. Or, dès les débuts de l'étude du système nerveux central, d'autres modes d'interactions intersensorielles ont été mis en évidence, telles que la convergence sous-corticale, ou l'influence d'informations d'une modalité sur l'activité d'un cortex spécifique d'une autre modalité sensorielle. Par ailleurs, de nombreuses études en psychologie expérimentale ont montré l'influence que pouvaient avoir les informations d'une modalité sur la perception sensorielle dans une autre modalité. Grâce à l'utilisation de techniques de neuroimagerie non invasives et à l'intégration de mesures comportementales et neurophysiologiques, des interactions intersensorielles "précoces" ont pu être mises en évidence plus récemment chez l'homme.

Les travaux de cette thèse ont concerné l'influence que peuvent avoir des informations visuelles dans deux phénomènes perceptifs mettant principalement en jeu le cortex auditif : la perception de la parole et la représentation en mémoire sensorielle auditive.

Concernant la perception de la parole, nous avons montré, dans une première étude en potentiels évoqués de surface chez le sujet sain, d'une part, que le temps de réponse pour réaliser une tâche de discrimination phonologique de syllabes est plus rapide lorsque ces syllabes sont accompagnées des mouvements articulatoires des lèvres qui les produisent et, d'autre part, que cette facilitation comportementale est associée à une diminution de l'activité auditive entre 120 et 200 ms après la présentation du son. Afin de mieux caractériser ces interactions audiovisuelles précoces, nous avons mené le même protocole expérimental sur un groupe de patients épileptiques porteurs d'électrodes implantées dans le cortex temporal. Les résultats de cette deuxième étude ont montré que la vision des mouvements articulatoires pouvait à elle seule activer le cortex auditif (principalement les cortex secondaires). Cette activation visuelle du cortex auditif pouvait entrainer une diminution de l'activité de traitement de la syllabe auditive entre 50 et 200ms, dont une partie seulement était visible sur le scalp dans la première étude. Les résultats de ces deux études peuvent s'expliquer soit par un effet d'indiçage temporel intersensoriel, dû au fait que les indices visuels précédaient toujours les indices auditifs dans les syllabes utilisées, soit par une véritable intégration des information phonétiques auditives et visuelles. Dans une troisième étude comportementale, nous avons montré que l'effet d'indiçage temporel intersensoriel suffisait à expliquer une diminution du temps de traitement des syllabes, mais uniquement dans des conditions d'écoute bruitées, ce qui suggère que cet effet n'est pas à l'origine de celui observé dans les deux premières études.

Pour étudier les représentations en mémoire sensorielle, nous avons utilisé la *Mismatch Negativity* (MMN, Négativité de discordance), une onde des potentiels évoqués générée par la détection automatique et pré-attentionnelle de la violation d'une régularité sensorielle. La MMN est générée dans les cortex sensoriels spécifiques (auditif ou visuel), et serait due à un processus de discordance neuronale entre la représentation de la régularité en mémoire sensorielle et l'entrée d'un stimulus déviant violant cette régularité. Dans une première étude comportementale, nous avons montré que la détection d'un événement déviant dans une suite d'événements audiovisuels standards était plus rapide lorsque cette déviance portait à la fois sur les traits auditifs et visuels plutôt que sur un seul des traits auditif ou visuel. Dans une deuxième étude, en potentiels évoqués de surface chez le sujet sain, nous avons montré que les interactions audiovisuelles vraisemblablement à l'origine de cette facilitation comportementale opéraient sur les processus liés aux MMN visuelle et auditive. Par ailleurs, la MMN visuelle générée par la déviance visuelle d'une régularité audiovisuelle différait de la MMN générée par la même déviance dans un contexte purement visuel. Dans une troisième étude, nous avons montré, réciproquement, que la déviance auditive d'une régularité audiovisuelle générait une MMN auditive différente de celle générée par la même déviance dans un contexte purement auditif. Ces deux derniers résultats indiquent que les représentations d'une régularité audiovisuelle dans les mémoires sensorielles auditive et visuelle incluent respectivement des informations visuelles et auditives. En revanche nous avons échoué à montrer, dans une dernière étude en magnétoencéphalographie, que la violation de la conjonction régulière de deux traits auditif et visuel suffisait à générer une MMN.

L'ensemble des ces résultats montrent que les traitements auditifs et visuels dans les cortex sensoriels spécifiques peuvent interagir à des étapes relativement précoces d'analyse. Les voies anatomiques pouvant expliquer ces effets précoces sont discutées.