UNIVERSITE DE PARIS VIII

ECOLE DOCTORALE COGNITION, LANGAGE, INTERACTION

Apport des mouvements buccaux, des mouvements extra-buccaux et du contexte facial à la perception de la parole chez l'enfant et chez l'adulte

Thèse de doctorat ès-sciences
Option : Psychologie cognitive

Présentée par
Grozdana ERJAVEC

Sous la direction de :
M. Denis LEGROS

Jury de soutenance :

- M. Farid EL-MASSIOUI, Professeur, Université de Paris VIII, CHART – EA 4004, président ;
- Mme Ranka BIJELJAC-BABIC, Maître de conférences-HDR, Université de Poitiers et Université de Paris V, Laboratoire Psychologie de la Perception – UMR 8242, rapporteur ;
- Mme Michèle MOLINA, Professeur, Université de Caen, Laboratoire PALM – EA 4649, rapporteur ;
- M. François JOUEN, Professeur, EPHE, Laboratoire CHART, examinateur ;
- M. Denis LEGROS, Professeur émérite, Université de Paris VIII, CHART – EA 4004, directeur de thèse.

Année universitaire 2014-2015

Remerciements

Know that the world is woven of interconnected threads. This is, because that is. This is not, because that is not. This is born, because that is born. This dies, because that dies. … The one contains the many and the many contains the one. Without the one, there cannot be the many. Without the many, there cannot be the one.

Buddha

À la fin de ce travail de thèse, j'aimerais remercier toutes les personnes ayant eu une influence, directe ou indirecte, sur sa réalisation. Mes plus vifs remerciements vont ainsi :

À mon directeur de thèse, M. Denis Legros, pour le courage de s'aventurer sur un terrain inconnu, pour ses conseils, son expertise, son soutien et sa patience tout le long du chemin.

Au directeur général de mon laboratoire de rattachement, M. Charles Tijus, pour m'avoir offert la chance de découvrir le monde de la recherche, pour son soutien, ses encouragements et pour le partage de son expertise scientifique.

Aux rapporteurs et aux membres du jury pour leur disponibilité, pour leur temps et pour m'avoir fait l'honneur d'accepter de lire, de commenter et de corriger ce travail.

Aux professeurs rencontrés le long de mon parcours, M. François Jouen, M. Thierry Baccino, M. Farid El-Massioui, M. Jean-François Lambert, qui m'ont apporté le savoir scientifique et celui de la vie qui m'a ouvert de nouvelles visions sur le monde et ses phénomènes.

À mes amis, Delphine, Rok, Marie, Emmanuel, Charlotte, Fatiha, Soraya, Kamila, Coralie, Anne, Viljem, Natasa et Zaklina pour m'avoir soutenue, conseillée, nourrie, hébergée et écoutée dans les moments les plus difficiles. (Merci Delphine.)

À M. Stéphane Natkin pour son soutien à la fin de ma thèse.

À Crt Erjavec, Soraya Bensalem, Yanka Camara et Maya Bizet pour leur engagement dans la réalisation de l'étude proposée dans le cadre de cette thèse. Merci pour votre expertise, votre temps et votre patience. (Merci Crt.)

À mes collègues du laboratoire pour leur soutien et pour leur aide technique et morale.

À toutes les personnes ayant pris le temps de participer aux expériences qui ont été menées.

À mes parents et à mes grands-parents qui m'ont tout donné.

Résumé

Le présent travail de thèse s'inscrit dans le domaine de recherche sur la perception audio-visuelle (AV) de la parole. Son objectif est de répondre aux questions suivantes : (i) Quelle est la nature du traitement de l'input visuel (holistique *vs* analytique) dans la perception AV de la parole ? (ii) Quelle est l'implication des régions faciales extra-buccales dans la perception AV de la parole ? (iii) Quel est le comportement oculaire des sujets lors de la perception AV de la parole ? (iv) Quelle est l'évolution de la perception AV de la parole dans les aspects (i), (ii) et (iii) au cours du développement. Le paradigme de dégradation de l'information auditive par le bruit a été utilisé dans deux expériences qui ont été menées avec des participants de quatre groupes d'âge, enfants, préadolescents, adolescents, adultes (16 participants par groupe). La tâche des participants consistait à répéter les syllabes de type consonne-voyelle /a/, faiblement et fortement dégradées par le bruit rose, présentées dans quatre conditions différentes. Ces conditions étaient les suivantes : une auditive (AU) et trois audio-visuelles (AV) (AV visage (AVV)), AV « bouche extraction » (AVB-E) (format bouche sans contrastes lumineux), AV « bouche-masquage » (AVB-M) (format bouche avec contrastes lumineux) pour l'expérience 1, et AVV, AV « bouche active » (AVV-BA) (format « visage » avec un contexte facial statique), AV « régions extra-buccales actives » (AVV-EBA) (format « visage » sans bouche) pour l'expérience 2. Le nombre total des répétitions correctes par condition (performance totale), la différence dans ce score entre chaque condition AV et la condition auditive (gain AV) et la durée totale des fixations oculaires dans la région buccale et les autres régions faciales (pour les formats AVV) ont été analysés. Les principaux résultats montrent que les mécanismes de traitement AV de la parole atteignent leur maturité avant l'enfance tardive. La vision du visage entier de l'orateur n'est pas avantageuse pour ce type de traitement. Elle semble même désavantageuse pour les adultes possiblement car elle déclenche le traitement d'autres aspects du visage (identité, expressions faciales) qui pourrait interférer avec le traitement des indices acoustiques relatifs à la parole. Pour les quatre groupes d'âge, la contribution des mouvements articulatoires dans les régions extra-buccales à la perception AV de la parole s'est avérée faible et limitée aux conditions de haute incertitude quant à l'information auditive. Pour les stimuli respectant les caractéristiques écologiques de l'information faciale, les patterns du comportement oculaire dans la perception bimodale de la parole varient en fonction du degré de dégradation de l'information auditive, mais semblent relativement stables durant la période allant de l'enfance à l'âge adulte. Finalement, les modalités de présentation de l'information visuelle localisée à la bouche ont affecté le

comportement oculaire chez les adultes, les pré-adolescents et les enfants. Ceci suggère que le traitement visuo-attentionnel dans le cadre de la perception AV de la parle est sensible aux caractéristiques de bas niveau des stimuli visuels chez ces populations. Les variations au niveau du traitement visuo-attentionnel s'accompagnent, dans une certaine mesure, de variations dans la perception AV de la parole.

Mots clefs : perception audio-visuelle de la parole, mouvements articulatoires, région buccale, régions extra-buccales, visage, contexte facial, comportement oculaire, développement

Abstract

The present thesis work fits into the domain/is incorporated within the framework of research on audio-visual (AV) speech perception. Its objective is to answer the following questions: (i) What is the nature of visual input processing (holistic *vs* analytic) in AV speech perception? (ii) What is the implication of extra-oral facial movement in AV speech perception? (iii) What are the oculomotor patterns in AV speech perception? (iv) What are the developmental changes in the above-mentioned aspects (i), (ii) and (iii)? The classic noise degradation paradigm was applied in two experiments conducted in the framework of the present thesis. Each experiment were conducted on participants of 4 age groups, adults, adolescents, pre-adolescents and children. Each group consisted of 16 participants. Participants' task was to repeat consonant-vowel (/a/) syllables. The syllables were both mildly and strongly degraded by pink noise and were presented in four audio(-visual) conditions, one purely auditory (AO) and three audio-visual conditions. The AV conditions were the following: (i) AV face (AVF), (ii) AV « mouth extraction » (AVM-E ; mouth format without visual contrasts), (iii) AV « mouth window » (AVM-W ; mouth format with high visual contrasts) in experiment 1, and (i) AVF, (ii) AVF « mouth active (and facial frame static) » (AVF-MA), (iii) AVF « extra-oral regions active (and mouth absent) » (AVF-EOA) in experiment 2. The data relative to (i) the total number of correct repetitions (total performance), (ii) the difference in the correct repetitions score between each AV and the AO condition (AV gain), and (iii) the total fixations duration in the oral area and other facial areas (for the AV formats) were analyzed. The main results showed that the mechanisms involved in AV speech perception reach their maturity before late childhood. The vision of the talker's full face does not seem to be advantageous in this context. It seems that the vision of the talker's full face might perturb AV speech processing in adults, possibly because it triggers processing of other types of information (identity, facial expressions) which could in terms interfere with the processing of acoustic aspects of speech. The contribution of the extra-oral articulatory movement to AV speech perception was poor and limited to the condition of highly degraded auditory information. For ecologically presented facial information, the oculomotor patterns in AV speech perception varied as a function of the level of auditory information degradation, but appeared rather stable across the 4 groups. Finally, the modalities of the featural (mouth) facial information presentation affected the oculomotor behavior patterns in adults, pre-adolescents and children, thus suggesting a certain sensitivity of visuo-attentional processing to low-level visual stimuli characteristics in AV

speech perception. The variations in visuo-attentional processing seemed to be associated to a certain extent with variations in AV speech perception.

Key words: audio-visual speech perception, articulatory movement, oral region, extra-oral regions, face, facial context, eye movement, development

Sommaire

Liste des tableaux

Liste des figures

1 Sur la nature bimodale de la perception de la parole

1.1 Introduction

Nos expériences du monde sont pour la plupart multimodales, c'est-à-dire qu'elles reposent sur plusieurs inputs sensoriels différents en lien avec un même objet ou un même évènement. Lors du processus perceptif, les informations provenant des différents canaux sensoriels sont intégrées dans un percept cohérent, unifié et généralement stable, assurant ainsi une efficacité optimale (rapidité et précision) de la perception. Ce type de perception porte le nom de perception multimodale ou encore intermodale (Stein & Meredith, 1993 ; Stein; Stanford; & Rowland, 2009 ; Zmigrod & Hommel, 2013). Par exemple, chez un sujet entendant, la perception d'un évènement tel que la chute d'un arbre résultera d'une intégration de l'input visuel, relatif aux signaux lumineux provenant de l'environnement (la forme et le mouvement de l'arbre qui tombe), et de l'input auditif en lien avec les signaux acoustiques générés par le tronc qui se casse, les branches heurtant d'autres arbres ou encore l'arbre tout entier touchant le sol. L'exemple présent est celui de la perception bimodale, reposant sur l'intégration de deux inputs sensoriels différents. L'expérience perceptive peut également être multimodale à proprement parler, c'est-à-dire résultant d'une intégration de trois inputs sensoriels distincts ou plus. Par exemple, la perception de la dégustation d'un sandwich résulte d'une intégration de l'input visuel (on voit le sandwich qui sera consommé), de l'input gustatif (on en perçoit le goût), de l'input olfactif (on en perçoit l'odeur), de l'input somesthésique (on en perçoit la structure dans la bouche), de l'input kinesthésique (on perçoit également les mouvements effectués lors de la mastication) et éventuellement de l'input auditif (on entend des craquements de la croûte de la baguette en train d'être croquée). En bref, les objets et les évènements du monde sont susceptibles de fournir une expérience sensorielle multimodale dont les différents éléments sont ensuite intégrés, par différentes régions cérébrales, dans un percept unifié.

Un exemple à part de la perception multimodale, en raison de la variabilité dans, la complexité et la rapidité de délivrance du signal à traiter est la perception de la parole. Longtemps conçue unimodale, car reposant sur la seule modalité auditive (voir la Figure 1), la perception de la parole est cependant, en l'état actuel de la recherche, est unanimement reconnue comme un phénomène multimodal (pour une revue, voir Dohen, 2009 ; Peelle & Sommers, *in press* ; Rosenblum, 2008). En effet, la production de la parole peut être définie comme une succession de mouvements respiratoires, phonatoires et articulatoires qui modulent les ondes de l'air en un signal acoustique présentant des caractéristiques spécifiques (pour plus de détails, voir Clark, Yallop, & Fletcher, 2007 ; Gick, Wilson, & Derrick, 2013). De ce fait, la

perception de la parole, en situation de communication face à face, repose sur deux types d'input sensoriel, auditif (relatif au message acoustique encodé par notre système sensoriel) et visuel (en lien avec les mouvements articulatoires de notre interlocuteur qui peuvent être perçus). On parle alors de la perception bimodale de la parole.

Ce chapitre, ainsi que la suite du document, aborde la perception de la parole du point de vue du paradigme cognitiviste qui s'intéresse à la façon dont la parole est traitée – encodée en représentations cognitives, stockée et utilisée - par le système cognitif, humain et artificiel, afin d'en identifier les éléments, d'en comprendre le message et d'élaborer une réponse comportementale adaptée. L'objectif de ce premier chapitre est de présenter les principales connaissances produites et approches méthodologiques utilisées dans la recherche sur ce sujet en mettant l'accent sur le rôle de l'information visuelle et l'intégration audio-visuelle dans la perception de la parole. Figurent également quelques lignes directrices que la recherche future dans ce domaine est susceptible de suivre.

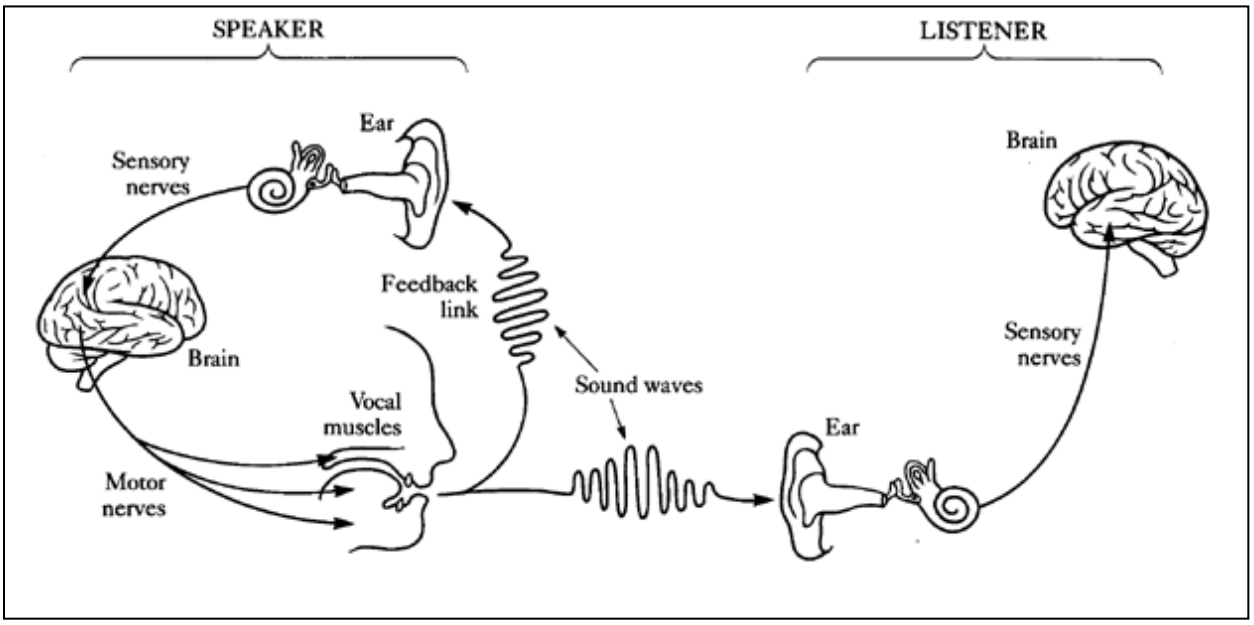

Figure 1. Chaîne de la communication parlée (Denese & Pinson, 1993).

L'Image de Denese et Pinson (1993) représente la perception de la parole comme un phénomène exclusivement auditif.

1.2 Input audio-visuel et son rôle facilitateur dans la perception de la parole

Dans les conditions écologiques de production d'un message parlé, ce dernier est souvent dégradé dans son aspect acoustique par le bruit environnant. La perception de la parole dans ce type de situations, appelées également les situations de *cocktail party* (Cheryy, 1953), peut ainsi poser des problèmes aussi bien pour les humains que pour les systèmes artificiels (Dupont & Luettin, 2000 ; Hauser, 1996 ; Kryter, 1996). Que l'ajout de l'information visuelle à l'information auditive puisse s'avérer particulièrement utile pour la perception de la parole dans ce type de conditions est un constat mis en évidence de manière empirique par Sumby et Pollack (1954). En effet, ces auteurs ont procédé à une dégradation acoustique d'une série d'items par le bruit blanc à des puissances différentes, créant ainsi différents ratios entre la puissance du message acoustique et le bruit, appelés également rapports signal-bruit (ou *signal-to-noise ratio (SNR*)). Les stimuli ainsi établis ont été présentés aux participants dans deux types de conditions, (i) sous forme acoustique uniquement (condition de présentation unimodale) et (ii) sous forme audio-visuelle, où l'information visuelle était congruente par rapport à et synchronisée avec l'information auditive (condition de présentation bimodale). Mesurant le nombre d'items correctement perçus dans chaque type de condition, les auteurs ont constaté que la perception du message était systématiquement meilleure (plus précise) dans les conditions de présentation bimodale. Partant du principe qu'une partie des performances perceptives des participants dans la condition bimodale reflétait directement leurs performances relatives au traitement de l'input auditif seul, Sumby et Pollack (1954) ont proposé que, pour chaque participant, la différence dans les scores perceptifs entre la condition audio-visuelle et la condition auditive seule représentait son gain audio-visuel, c'est-à-dire le gain dans la précision de la perception du message parlé relatif au traitement d'un input bimodal, audio-visuel.

Hormis quelques variations dans l'ampleur du gain audio-visuel, les observations de Sumby et Pollack (1954) ont été largement répliquées depuis (e.g., Grant, 2001 ; Binnie, Montgomery, & Jackson, 1974 ; Erber, 1969 ; Eramudugolla, Hendrson, & Matingley, 2010 ; Sommers, Tye-Murray, & Spehar, 2005 ; Stevenson & James, 2009). Aussi, il est actuellement communément admis que l'input bimodal, audio-visuel, facilite l'acuité perceptive (la reconnaissance des phonèmes, les plus petites unités discrètes et distinctives sur le plan phonologique), ainsi que la reconnaissance des mots (pour plus de détails, voir Borrie, 2015 ; voir également la section 1.3.2) dans les conditions où l'input auditif est dégradé.

Un point important de la recherche sur l'effet facilitateur de l'input bimodal dans la perception de la parole concerne la relation entre le gain audio-visuel et l'ampleur du rapport

signal-bruit qui détermine le degré de fiabilité du signal acoustique. En effet, il est bien établi que l'ampleur du gain audio-visuel varie en fonction du degré de dégradation de l'information auditive (e.g., Ma, Zhou, Ross, Foxe, & Parra, 2009 ; McCormick, 1979 ; O'Neil, 1954 ; Ross, Saint-Amour, Leavitt, Javitt, & Foxe, 2007 ; Sumby & Pollack, 1954 ; etc.). Les auteurs des premières études sur ce sujet ont suggéré que le gain audio-visuel augmente avec la baisse de l'intelligibilité de l'input auditif (Callan, Jones, Munhall, Callan, Kroos, & Vatikiotis-Bateson, 2003 ; Erber, 1969, 1971, 1975 ; Sumby & Pollack, 1954). Un tel pattern des résultats serait conforme à la prédiction basée sur le principe de l'efficacité inversée (*principle of inverse effectiveness* (*PoIE*)) (Meredith & Stein, 1983). Selon ce principe, une diminution des réponses du système cognitif aux stimuli unimodaux s'accompagne d'un renforcement du traitement multimodal. Néanmoins, les études de Ross et *al.* (2007) et Ma et *al*. (2009) ont montré que le gain audio-visuel est le plus important pour des dégradations moyennes à fortes (soit un rapport signal-bruit de -12dB). Il baisse pour des dégradations plus extrêmes, allant soit vers très faibles soit vers très fortes. La relation entre le gain audio-visuel et le degré de dégradation de l'information auditive est ainsi plus proche des prédictions du modèle Bayésien à condition qu'une gamme assez large d'items lexicaux soit prise en compte (voir Ma et *al*., 2009) et puisse être représentée graphiquement sous forme de courbe qui prend la forme d'une cloche (voir la Figure 2).

Malgré l'utilité évidente du paradigme proposé par Sumby et Pollack (1954), connu sous le nom de paradigme de dégradation de l'information auditive par le bruit, deux limites majeures ont pu être identifiées dans ses caractéristiques et son utilisation dans la majorité des études qui y ont eu recours. Premièrement, le principe de la dégradation de l'information auditive par le bruit restreint nécessairement l'étude de la facilitation de la perception de la parole par un input bimodal à des situations d'écoute à caractère aversif, nécessitant un certain effort cognitif. Deuxièmement, les stimuli les plus fréquemment utilisés dans les études qui s'appuient sur ce paradigme étant des syllabes et des mots, les résultats ne permettent pas d'évaluer l'apport d'un input bimodal à la perception d'unités linguistiques plus complexes telles que les phrases ou encore les textes. Troisièmement, la nature de la tâche limite l'étude de l'effet de l'input bimodal au domaine de la perception, et plus précisément encore, à l'acuité et à la reconnaissance perceptives. Certains auteurs se sont pourtant intéressés à d'autres dimensions cognitives du domaine. Par exemple, Reisberg, McLean et Goldfield (1987) ont établi qu'un avantage audio-visuel pouvait également apparaitre au niveau de la compréhension des phrases sémantiquement complexes, mais clairement audibles. Dans la même lignée, Arnold et Hill (2001) ont mis en évidence le rôle facilitateur de l'input bimodal, audio-visuel,

dans la compréhension des textes courts présentés dans des conditions ne comportant aucun bruit acoustique. L'étude de Wayne et Johnsrude (2012) a établi que l'input bimodal, audio-visuel, facilitait également l'apprentissage perceptif de la parole, notamment chez les patients porteurs d'implants cochléaires. D'autres auteurs encore ont montré que ce type d'input affecte (augmente) la rapidité de la reconnaissance des items lexicaux (e.g., Reisberg et *al*., 1987 ; Sekiyama, Soshi et Sakamoto, 2014) ou bien qu'il diminue l'effort cognitif alloué à la reconnaissance de tels items présentés dans des conditions d'écoute aversive (Gosselin & Gagné, 2011). Finalement, le champ des problématiques en lien avec l'effet facilitateur de l'input bimodal dans la perception de la parole a été étendu à d'autres types de conditions aversives où le message parlé peut être difficile à identifier à partir de l'input auditif seul, notamment quand ce dernier est produit dans une langue seconde (Davis & Kim, 2004 ; Hazan, Kim, & Chen, 2010 ; Navarra & Soto-Faraco, 2007)[1] ou encore par une personne présentant un accent étranger (e.g., Reisberg et *al.*, 1987). Dans toutes ces situations, l'input bimodal s'avère être un élément facilitateur de la perception de la parole et cet effet est d'autant plus important que l'orateur nous est connu (Kim & Davies, 2011).

Au vu des éléments empiriques, la communauté scientifique reconnaît unanimement que l'input bimodal, audio-visuel, facilite le traitement cognitif de la parole dans de nombreux aspects. Le paradigme de la dégradation de l'information auditive par le bruit a été spécifiquement conçu pour mettre en évidence et évaluer le phénomène en question pour les dimensions d'acuité et de reconnaissance perceptives. Il reste aujourd'hui un des paradigmes les plus largement utilisés dans le champ de la perception bimodale de la parole. Toutefois, l'interprétation du gain audio-visuel peut s'avérer quelque peu délicate. En effet, le principe d'additivité des performances qui seraient liées à un type de traitement (unimodal et bimodal) n'est pas clairement défini. Par exemple, certaines tentatives de modélisation des performances perceptives des sujets, pour la condition de présentation audio-visuelle, à partir de leurs performances obtenues dans les conditions unimodales -auditive seule et visuelle seule - laissent penser que la perception bimodale de la parole ne peut être expliquée par une simple addition des performances perceptives unimodales (voir Blamey, Cowan, Alcantara, Whitford, & Clark, 1989 ; Braida, 1991 ; Grant, 2002). Ceci qui suggère une certaine complexité des mécanismes dédiés à l'intégration des informations provenant de chaque input. Il n'est ainsi pas clairement

[1] L'effet facilitateur de l'input bimodal est cependant moindre dans de telles conditions (voir Yi, Phelps, Smiljanic, & Chandrasekaran, 2013).

établi dans quelle mesure le gain audio-visuel permet de mesurer spécifiquement la capacité de l'individu de traitement bimodal de la parole.

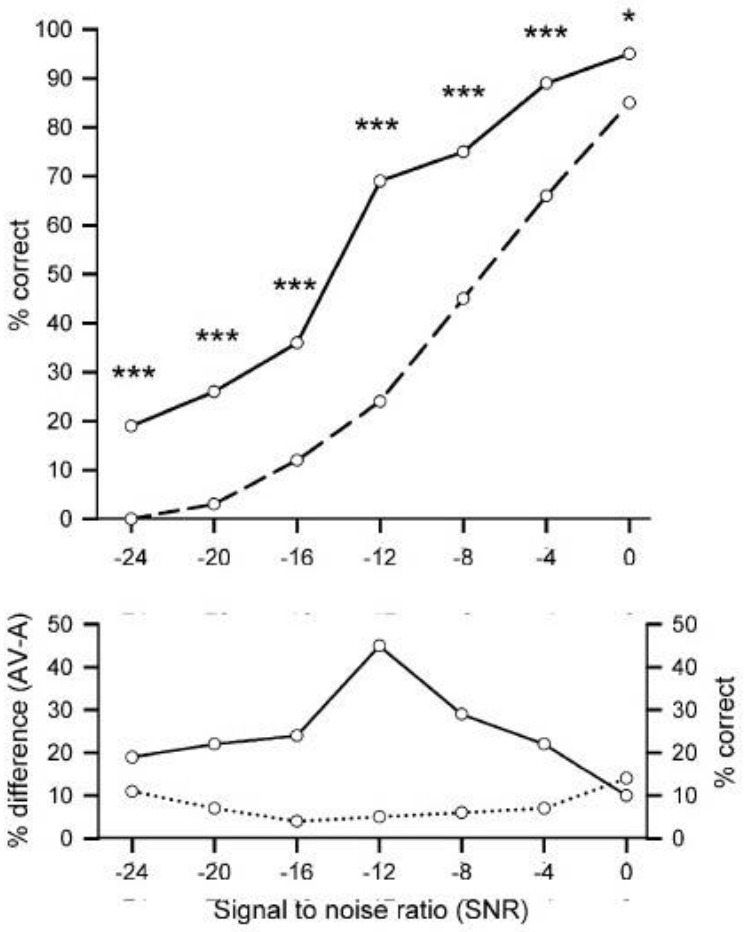

Figure 2. Représentation graphique des relations entre le degré de dégradation du message acoustique et les différentes mesures des performances perceptives recueillies dans le cadre du paradigme de dégradation de l'information auditive par le bruit (Ross et *al.*, 2007).

Le graphique du haut de la Figure 2 illustre le rapport entre les différents SNR et le pourcentage des mots correctement identifiés (*% correct*) dans la condition auditive (ligne à traits) et la condition audio-visuelle (ligne pleine). Les différences significatives entre les deux conditions sont indiquées par des étoiles (* pour p<0,05 et *** pour p<0,01). Le graphique du bas représente le rapport entre les différents SNR et le gain audio-visuel (*% difference (AV-A)*) (pour la ligne pleine) et le pourcentage d'items correctement identifiés (*% correct*) dans la condition visuelle seule (pour la ligne à traits). (L'image de Ross et *al*. (2007).)

1.3 Indices fournis par l'input visuel dans la perception de la parole

Dans la mesure où la parole est produite par des mouvements de différents organes du tract vocal, le message acoustique est forcément relatif aux mouvements articulatoires, phonatoires (les mouvements des cordes vocales, moins visibles cependant) et même respiratoires qui en sont l'origine. Comme nous venons de le voir, l'input visuel ajouté à l'input auditif correspondant facilite la perception de la parole. Mais quelles sont les informations exactes véhiculées par l'input visuel et comment sont-elles traitées par le système cognitif humain pour reconnaître les unités phonologiques, élémentaires ou plus complexes, ayant été produites ?

1.3.1 Indices sur les aspects temporels de l'input auditif

La première catégorie d'indices véhiculés par l'input visuel lors de la perception de la parole est informative par rapport à l'occurrence temporelle du message acoustique. Par exemple, l'initiation articulatoire d'un phonème tel que « p » (/p/) dans « papa » (/pa-pa/), en l'occurrence, il s'agit d'un mouvement d'ouverture suivi de celui de fermeture de la bouche, précède la production du signal acoustique lui-même. Elle peut ainsi nous informer sur la suite de l'évènement, c'est-à-dire sur l'arrivée prochaine du signal acoustique (voir la Figure 3).

Ces gestes préparatoires, ou indices pré-vocaux (*prevoicing features*), semblent faciliter la détection du signal acoustique. En effet, dans une étude phare abordant ce type de problématiques, Grant et Seitz (2000) ont observé une réduction de 0,8 à 2,2 dB du seuil de détection des phrases dégradées acoustiquement dans lesquelles les inputs visuel et auditif étaient congruents et présentés de manière synchrone. Ayant procédé à une analyse de corrélation entre la zone de l'ouverture de la bouche et les fluctuations dans l'enveloppe sonore[2] pour les fréquences de base définissant la qualité d'un son (formant F1, formant F2 formant F3 et aussi bande large), les chercheurs en question ont observé que les corrélations les plus importantes étaient associées à des réductions plus grandes du seuil de détection des phrases. Cette observation révèle ainsi que la cohérence temporelle entre l'input visuel et l'input auditif a joué un rôle majeur dans l'élévation du niveau de détection du signal acoustique par les indices visuels (Grant, 2001 ; Kim & Davis, 2003). Ainsi, il n'est pas surprenant qu'en cas de présentation d'inputs synchronisés, mais non congruents (c'est le cas lorsque le message visuel

[2] L'enveloppe sonore est relative la courbe qui correspond à l'évolution d'une caractéristique du son en fonction de temps.

ne correspond pas au message acoustique), les performances des participants dans la détection des phrases ne différaient pas de celles obtenues dans la condition auditive seule (Grant & Seitz, 2000). Utilisant des stimuli moins complexes tels que des syllabes ou des mots, d'autres études montrent même que le seuil de détection de ces unités est augmenté dans les conditions où l'input visuel ne correspond pas à l'input auditif ou n'est pas synchronisé avec celui-ci (e.g., Bernstein, Auer, & Takayanagi, 2004 ; Kim & Davies, 2004 ; Thomas & Jordan, 2004)[3].

Une autre étude qui a également abouti au constat que la parole dans sa modalité visuelle renforce notre sensibilité au signal acoustique est celle de Schwartz, Berthomier et Savariaux (2004). Ayant utilisé le paradigme de la dégradation de l'information visuelle par le bruit et des stimuli qui différaient dans leur aspect acoustique (les voyelles /u/ (comme dans « mou ») et /y/ (comme dans « du »)), mais pas dans leur aspect visuel (la même vidéo d'une bouche articulant une forme arrondie avant de se fermer a été présentée dans les deux cas), ces auteurs ont constaté que le taux de reconnaissances correctes était plus élevé dans la condition de présentation audio-visuelle que dans la condition auditive seule. Aussi, les indices de nature strictement temporelle, qui correspondraient aux mouvements d'ouverture et de fermeture de la bouche et précéderaient les fluctuations dans l'enveloppe sonore du message acoustique, véhiculés par l'input visuel semblent faciliter non seulement la détection du message acoustique correspondant, mais également la reconnaissance de ce dernier.

Le phénomène de l'élévation du niveau de la sensibilité au signal acoustique induit par des indices visuels correspondants a également été étudié à l'aide d'indices neurophysiologiques (voir la section 1.5.1 pour plus de détails sur la question). D'une part, les résultats des études sur l'activité neuronale globale du A1 montrent que les stimuli visuels sont susceptibles d'influencer les oscillations dans l'excitabilité neuronale de cette région en remettant en phase ces dernières, de façon à s'accorder à la structure rythmique des stimuli (Kayser, Petkov, & Logothetis, 2008 ; Perrodin, Kayser, Logothetis, & Petkov, 2015), modulant ainsi l'efficacité de traitement des stimuli à venir (pour plus de détails, voir la section 1.5.1). D'autre part, les études utilisant la méthode des potentiels évoqués ont mis en évidence une diminution de l'amplitude et de la latence (ce qui marque une facilitation de traitement) des ondes N1 et P2 au niveau du A1 qui reflètent le traitement d'un stimulus donné par la région en question (Alsius, Möttönen, Sams, Soto-Faraco, & Tiippana, 2014 ; Besle, Fort, Delpuech, &

[3] Toutefois, voir la section 1.4 sur la fusion audio-visuelle. Dans certaines conditions, l'incogruence et l'asynchronie audio-visuelles ne sont pas un obstacle au traitement bimodal de la parole.

Giard, 2004 ; van Wassenhove, Grant, & Poeppel, 2005). Ces résultats sont généralement interprétés comme un effet des indices visuels de la parole sur le traitement attentionnel du signal acoustique. Les indices visuels moduleraient les mécanismes neurophysiologiques sous-tendant le traitement attentionnel de l'input auditif et ainsi l'efficacité même du traitement du signal acoustique (pour plus de détails, voir la section 1.5.1).

Pour finir, il est important de souligner qu'un consensus relativement général a été établi au sujet des relations temporelles entre l'input visuel et l'input auditif correspondant ; le premier précéderait d'environ 200 à 150 ms le second (pour plus de détails, voir Chandrasekaran, Trubanova, Stillittano, Caplier, & Ghazanfar, 2009). Schwartz et Savariaux (2014) contestent cependant cette supposition en avançant que ceci peut être vrai pour les gestes préparatoires, au début de la production d'une phrase, ou encore à l'intérieur d'une phrase pour des consonnes occlusives (par exemple, /p/, /k/, /g/, etc.), marquées, sur le plan acoustique, par une courte pause. Ces gestes préparatoires ont une valeur prédictive dans le sens où une ouverture de la bouche plus ou moins prononcée mais non accompagnée par un son nous laisse deviner qu'il y aura une suite à ce mouvement, une fermeture, puis une nouvelle ouverture qui se soldera par l'émission d'un son. Toutefois, quand il s'agit d'une parole coarticulée ou continue, c'est-à-dire quand on parle des mouvements d'ouverture et de fermeture de la bouche à l'intérieur d'une suite continue de phonèmes, tout geste articulatoire est souvent accompagné de l'émission d'un son et relativement bien synchronisé avec le son produit. Ces gestes, que Schwartz et Savariaux (2014) appellent des gestes comodulatoires, présenteraient donc une relation plus simple en termes d'évènements (un mouvement articulatoire pour un évènement acoustique) et en termes de leur occurrence temporelle (une synchronie relativement importante) que les gestes préparatoires (voir la Figure 3).

Les relations temporelles entre les gestes articulatoires et les évènements acoustiques semblent ainsi relativement complexes. On pourrait supposer que les gestes préparatoires ont un rôle crucial pour le traitement de la suite de la chaîne parlée, coarticulée, dans le sens où la parole possède par elle-même une structure rythmique, dans laquelle les gestes articulatoires (et notamment les ouvertures de la bouche) sont corrélés avec les pics d'amplitude dans l'enveloppe acoustique du message parlé. Cette structure rythmique pourrait être plus ou moins anticipée/prédite par la personne qui reçoit le message parlé. Les gestes préparatoires correspondent peut-être au coup d'envoi du traitement d'un évènement complexe, mais rythmique, donc pouvant être anticipé. Le point de départ pourrait ainsi s'avérer être déterminant pour le traitement de la suite. Toutefois, une telle supposition doit nécessairement

être vérifiée de manière empirique. En effet, les études neurophysiologiques citées précédemment, ayant utilisé essentiellement des stimuli monosyllabiques ou encore des mots isolés séparés par un certain intervalle temporel, semblent montrer qu'une structure rythmique régulière facilite le traitement du signal acoustique et que les indices visuels sur l'occurrence d'un évènement acoustique facilitent son traitement dans la modalité visuelle. Toutefois, vu le choix et les modalités de l'administration des stimuli, ces études ne peuvent pas être conclusives quant au rôle des deux types de gestes articulatoires, préparatoires et comodulatoires, dans le traitement de la parole telle qu'elle est produite dans des conditions écologiques.

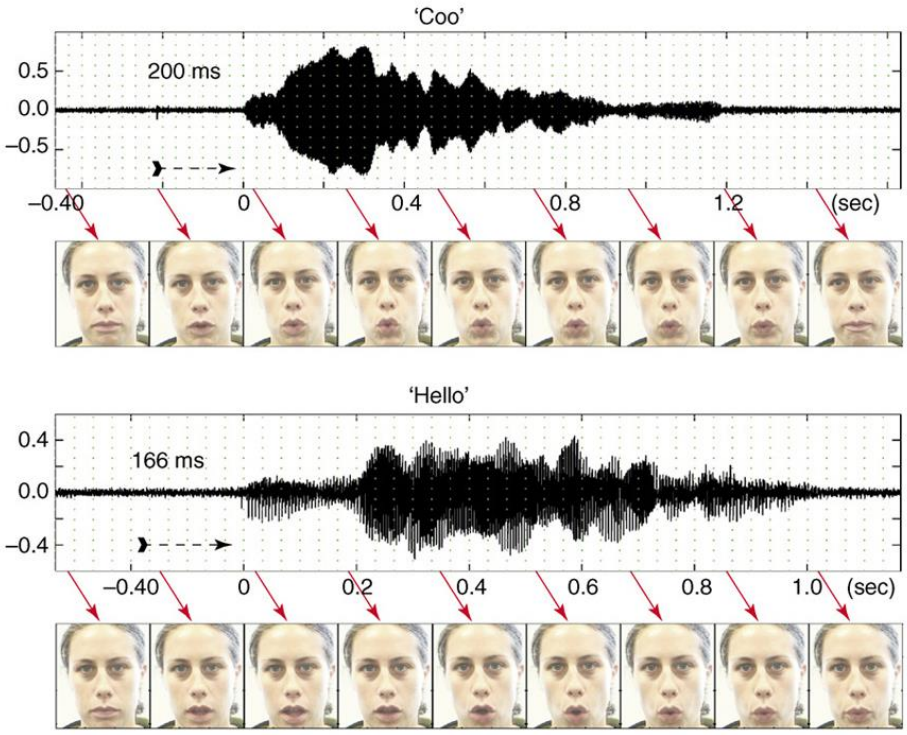

Figure 3. Représentation schématique des relations temporelles entre les mouvements articulatoires et les variations dans l'amplitude des sons produits pour les mots « Coo » /ku :/ et « Hello » /hʌˈləʊ/ (Schroeder, Lakatos, Kajikawa, Partan, & Puce, 2008).

L'oscillogramme, illustrant le décours temporel et l'amplitude des éléments acoustiques, est mis en relation avec les endroits clefs dans la production motrice. On constate ainsi que les mouvements articulatoires préparatoires précèdent de quelques millisecondes (200 ms pour « Coo » et 166 ms pour « Hello ») le moment de voisement (*voice onset time (V.O.T.)*). En revanche, les relations temporelles sont beaucoup plus rapprochées pour la suite de la chaîne acoustique et les mouvements articulatoires comodulatoires avec, notamment, les pics dans

l'amplitude des éléments acoustiques correspondant aux mouvements d'une ouverture importante de la bouche. (L'image de Schroeder et *al*. (2008).)

1.3.2 Indices sur le contenu de l'input auditif

Outre les indices sur l'occurrence temporelle des évènements acoustiques (sons), l'input visuel fournit également des indices sur le contenu même des sons produits, notamment sur la composition spectrale définissant leurs qualités phonétiques. En effet, il est largement connu que certaines personnes malentendantes peuvent identifier les consonnes et les mots produits par un locuteur directement à partir de ses mouvements articulatoires. On dit alors que la parole a été lue sur les lèvres de celui qui parle. Selon certaines données (Schwartz, 2004), cette lecture sur les lèvres nous permettrait à elle seule d'identifier un bon nombre de phonèmes (de 40% à 60%) et de mots (de 10% à 20% pouvant aller jusqu'à 60%).

Pour comprendre comment on réussi à identifier des phonèmes à partir du signal visuel correspondant, il est nécessaire de s'intéresser de plus près à la façon dont la parole est produite, ce qui relève du champ de la phonétique articulatoire. Comme il a été dit plus haut, la production de la parole peut être considérée comme une suite de mouvements ayant pour fonction de produire un courant d'air (mouvements respiratoires assurés par le diaphragme et les poumons) et de le moduler. Sur son chemin vers les poumons ou des poumons vers l'extérieur (plus courant dans les langues européennes), le courant d'air, qui emprunte les voies du système respiratoire, est modulé au niveau du larynx par les cordes vocales. On parle de mouvements phonatoires. L'activité des cordes vocales donne au courant d'air sa première caractéristique appelée le voisement. En cas de vibrations des cordes vocales, le son produit est défini comme sonore, dans le cas contraire comme sourd. Par la suite, les ondes de l'air passent par le système supra-laryngien, qui consiste en une série de cavités (pharyngienne, orale et nasale), de muscles, d'os et de dents qui modulent les ondes sonores davantage par des mouvements articulatoires. A ces mouvements articulatoires on attribue deux autres caractéristiques qui définissent la qualité acoustique des sons produits, notamment le point et le mode d'articulation. Le point d'articulation est relatif à l'endroit où se produit l'obstruction dans le courant d'air. Par exemple, dans le cas d'un phonème glottal tel que /g/, l'obstruction se produit au niveau de la glotte. Le mode d'articulation d'un son, en revanche, est relatif à la façon dont le courant d'air est obstrué tout le long du tract vocal. Par exemple, pour un phonème occlusif tel que /t/, l'obstruction de l'air est totale et suivie par une libération totale de l'air. Notons que le point et

le mode d'articulation s'appliquent uniquement aux consonnes, car la production des voyelles n'implique pas d'obstruction du courant d'air.

Les mouvements et les différentes positions des cordes vocales ainsi que des articulateurs façonnent ainsi le courant d'air et lui fournissent différentes caractéristiques acoustiques telles que les fréquences résonnantes, connues sous le nom de formats, ainsi que l'amplitude. Par exemple, une grande ouverture de la bouche, comme dans la voyelle /a/, permet à l'air de sortir avec beaucoup de puissance attribuant ainsi au son une amplitude importante (pour plus de détails, voir Clark et *al.*, 2007 ; Gick et *al.*, 2013).

Si les mouvements phonatoires sont peu visibles, les mouvements articulatoires le sont davantage, essentiellement en ce qui concerne la bouche, la langue et les dents. En effet, comme le notent certains auteurs (Jiang, Alwan, Keating, Auer, & Bernstein, 2002 ; Yehia, Rubin, & Vatikiotis-Bateson, 2002), chez l'humain, la production de la parole est associée à des mouvements prévisibles provoquant des déformations de la région buccale, mais aussi des régions extra-buccales qui sont clairement visibles. Ce sont ainsi ces éléments qui sont susceptibles de fournir des informations à une personne recevant un message parlé quant aux sons produits. Cette supposition trouve également des arguments empiriques. En effet, dans une étude de Munhall et Vatikiotis-Bateson (2004), une corrélation très forte entre les mouvements du devant du visage et l'amplitude, ainsi que les paramètres spectraux[4] des sons produits a été trouvée. Il semblerait que ces indices sont utilisés avec succès dans la perception bimodale de la parole Par exemple, dans une étude de Kuhl, Williams et Meltzoff, (1991), les participants adultes associaient systématiquement des vocalises d'une hâteur élevée à des images présentant une bouche ouverte et aplatie telle qu'on trouve dans la production du son /i/ possédant lui-même des fréquences élevées, et les vocalises d'une hâteur basse à l'image d'une bouche grand-ouverte telle qu'on trouve dans la production du son /a/ caractérisé par des fréquences basses.

Le fait que les indices visuels puissent être particulièrement utiles pour la reconnaissance des unités phonologiques correspondantes, les phonèmes, a été mis en évidence par Summerfield (1987). Cet auteur a en effet établi que les confusions dans la perception des phonèmes pouvaient être différentes entre la modalité visuelle seule et la modalité auditive seule, soulignant ainsi la relation de complémentarité en termes d'informations entre les deux types d'inputs (voir aussi Robert-Ribes, Schwartz, Lallouache, & Escudier, 1998). Plus

[4] Les paramètres spectraux sont relatifs aux changements dynamiques dans la composition en fréquences et en amplitude.

précisément, Summerfield (1987) suggère que l'information relative au point d'articulation, très vulnérable à des distorsions acoustiques quand elle est présentée dans sa modalité auditive, serait extraite à partir d'indices visuels, notamment à partir de la forme et de l'ouverture de la bouche, ainsi que de la position de la langue et des dents. En revanche, l'information en lien avec le voisement et le mode d'articulation serait extraite à partir des caractéristiques acoustiques d'une consonne. Par exemple, visuellement, les phonèmes /p/ et /b/ sont pratiquement identiques. Ils ont le même point d'articulation ; ce sont des consonnes bilabiales. Or, sur le plan auditif, on arrive à les différencier grâce aux différences présentes au niveau du formant F2 en lien avec le voisement de chaque consonne (/p/ étant une consonne sourde, /b /une consonne sonores). Par opposition à l'exemple précédent, on peut s'intéresser aux consonnes /m/ et /n/. En effet, si une confusion auditive entre les deux est fréquente car elles ont le même mode d'articulation (il s'agit de consonnes nasales), ainsi que le même type de voisement (ce sont des consonnes sonores), visuellement la différenciation entre les deux est marquée et facilement visible (/m/ étant une consonne bilabiale, /n/ une consonne alvéolaire).

Cette différence et cette complémentarité entre les unités phonologiques de base, les phonèmes, et les plus petites unités discrètes et distinctives en modalité visuelle, appelées les visèmes, présenteraient un avantage non seulement pour la reconnaissance des sons élémentaires d'une langue, mais également des unités plus complexes, les mots. En effet, ayant adapté le modèle de voisinage phonologique, impliquant le principe d'activation-compétition dans l'explication de la reconnaissance auditive des mots, à la modalité visuelle, Tye-Murray, Sommers et Spehar (2007) ont introduit le terme de voisinage visuel. Si le voisinage phonologique est défini comme le nombre de mots différant du mot cible au niveau d'un seul phonème (Luce & Pisoni, 1998 ; Marslen-Wilson & Tyler, 1980), le voisinage visuel d'un mot correspond au nombre de mots qui diffèrent du mot cible au niveau d'un seul visème. Lors de la reconnaissance d'un mot, le mot cible ainsi que ses voisins seraient activés, et le mot cible devrait être choisi au terme d'un processus de compétition entre l'ensemble des items activés. Un phénomène important apparaît alors ; la taille du voisinage phonologique complique la reconnaissance d'un mot. En effet, les performances des participants dans la reconnaissance d'un mot sont négativement corrélées avec la taille de son voisinage (Luce & Pisoni, 1998 ; Vitevitch, Stamer, & Sereno, 2008). Ce même phénomène a également été observé pour les mots présentés dans leur modalité visuelle seule (Auer, 2002 ; Feld & Sommers, 2011 ; Mattys, Bernstein, & Auer, 2002), validant ainsi le concept de voisinage visuel. Selon la même logique, la présentation d'un mot dans sa forme audio-visuelle introduit un nouveau type de voisinage,

le voisinage bimodal/audio-visuel, qui correspond à l'intersection des voisinages phonologique et visuel. Un phénomène intéressant est alors susceptible de se produire, le voisinage bimodal d'un mot peut être considérablement réduit par rapport à chacun de ses voisinages unimodaux, ce qui facilite sa reconnaissance. En somme, l'input visuel peut apporter des contraintes supplémentaires dans le processus de sélection de l'item cible facilitant ainsi la reconnaissance de ce dernier (Tye-Murray, Sommers, & Spehar, 2007b). (Pour une illustration du phénomène, voir la Figure 4 ; pour plus de détails, voir Yao, 2011.)

Tous les éléments étant pris en compte, il convient de souligner que les indices fournis par l'input visuel pour la perception de la parole peuvent être redondants et complémentaires à ceux véhiculés par l'input auditif. D'une part, l'input visuel peut être considéré comme redondant par rapport à l'input auditif, c'est-à-dire être bien corrélé avec et reproduisant en quelque sorte ce dernier. Il s'agit ici des correspondances dans l'intensité, la durée, le tempo et le rythme des deux inputs. En effet, les fluctuations du signal acoustique dans ces dimensions sont associées aux paramètres de la réalisation motrice tels que le début et la fin des gestes articulatoires, la fréquence et l'amplitude du mouvement d'ouverture et de fermeture de la bouche. Ce type d'indices, non spécifique d'une modalité sensorielle et relativement bien synchronisé à travers les différents canaux sensoriels qui le véhiculent, est qualifié d'amodal (e.g., Bahrick & Lickliter, 2014 ; Brenna, Nava, Turati, Montirosso, Cavallini, & Borgatti, 2015 ; Flom & Bahrick, 2007). D'autre part, certains indices visuels et acoustiques traités lors de la perception bimodale de la parole sont également spécifiques des modalités sensorielles, visuelle et auditive respectivement. En effet, l'input visuel apporte des informations qui lui sont spécifiques sur les patterns dynamiques de la réalisation motrice d'une ou plusieurs unités sonores. Ces informations sont relatives aux changements dans la forme, la position et le mode des mouvements des articulateurs. Les indices spécifiquement visuels, les visèmes, trouvent leur correspondance dans les indices spécifiquement auditifs, les phonèmes qui concernent les patterns dynamiques des caractéristiques phonétiques des unités produites telles que, par exemple, l'occurrence des fréquences caractéristiques d'un phonème. Comme le note Summerfield (1987), l'information spécifique de la modalité visuelle est différente et ainsi complémentaire à celle spécifique de la modalité auditive. En effet, les indices acoustiques les plus saillants apportés par l'input visuel sont en lien avec le point d'articulation, alors que les indices acoustiques les plus saillants apportés par l'input visuel sont relatifs au mode d'articulation et au voisement (voir aussi Munhall & Vatikiotis-Bateson, 2004). Comme il a été décrit plus haut, la complémentarité entre les deux inputs peut s'avérer particulièrement utile

pour la perception de la parole dans le cas où la différenciation perceptive d'un élément de la parole est difficile dans une modalité, mais pas dans l'autre.

Une personne adulte, experte de la perception audiovisuelle de la parole, fait usage des indices visuels aussi bien redondants que complémentaires pour identifier le message parlé. Les indices redondants semblent être importants dans la gestion du traitement attentionnel, permettant, à terme, une meilleure détection et reconnaissance du signal acoustique. En revanche, les indices complémentaires agissent comme des contraintes dans l'identification des sons (phonèmes) et par-là dans la reconnaissance des mots. Il semble ainsi évident que la perception de la parole est facilitée par un input bimodal, audio-visuel.

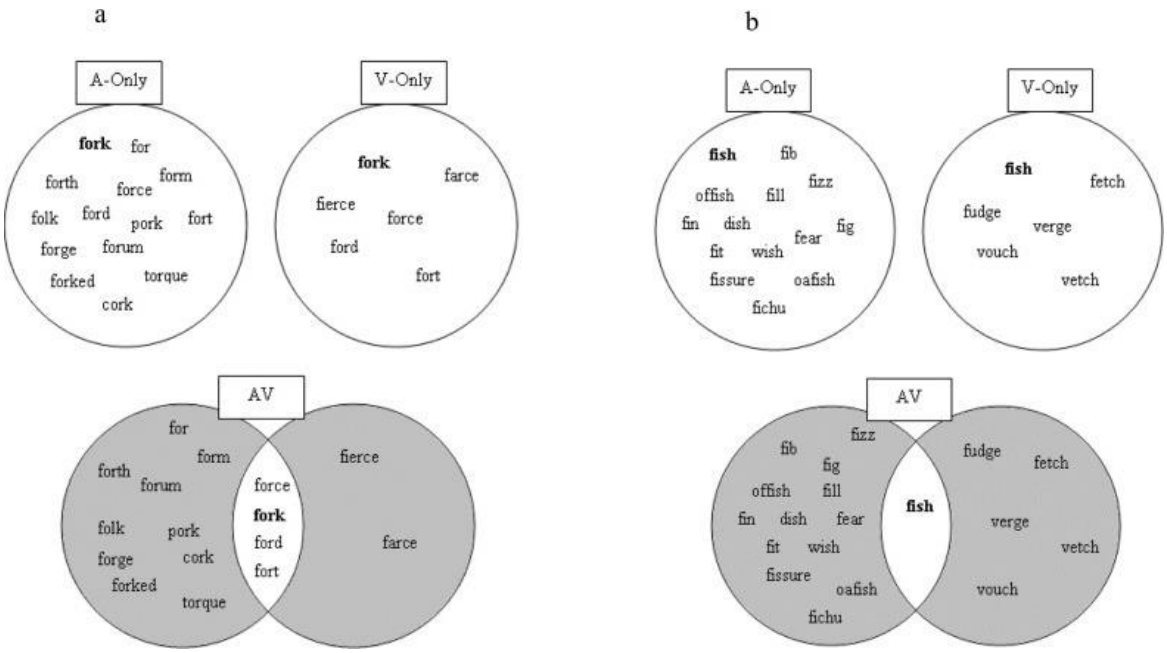

Figure 4. Illustration schématique des voisinages phonologique, visuel et audio-visuel pour les mots « fork » et « fish » (Tye-Muray et *al.*, 2007b).

Le voisinage audio-visuel du mot « fish » (b) comporte considérablement moins d'éléments que ses voisinages phonologique et visuel respectivement. Son identification est ainsi plus facile s'il est présenté sous forme bimodale. Cette relation entre la taille du voisinage bimodal et les voisinages unimodaux ne s'applique pas de manière systématique. Par exemple, le voisinage bimodal du mot « fork » (a) est plus restreint que son voisinage phonologique, mais plus grand que son voisinage visuel. (L'image de Tye-Muray et *al.* (2007b).)

1.4 Au-delà de la facilitation : fusion audio-visuelle

L'input bimodal, nous l'avons vu, facilite la reconnaissance des phonèmes et des mots par le biais des informations que la vision apporte aussi bien sur l'occurrence temporelle de l'input auditif que sur son contenu, ses caractéristiques acoustiques. Si la communauté scientifique admet unanimement que l'input visuel est porteur d'indices très utiles à la perception de la parole, la façon dont les inputs visuel et auditif sont traités pour produire un percept cohérent et stable n'est pas encore bien connue. Par ailleurs, le paradigme de la dégradation de l'information auditive par le bruit met en évidence de grandes différences interindividuelles aussi bien au niveau des performances dans des conditions unimodales (reconnaissance correcte d'un input visuel ou auditif seul) que dans le bénéfice tiré d'un input bimodal (gain audio-visuel) (e.g., Grant, Walden, & Seitz, 1998 ; Sommers et *al*., 2005; Tye-Murray, Sommers, & Spehar, 2007a). Un phénomène est largement cité dans ce contexte ; il s'agit notamment de la fusion audio-visuelle qui impliquerait des mécanismes ayant pour fonction l'intégration des indices véhiculés par les deux modalités en question, visuelle et auditive. Dans une vision systémique des choses où le tout dépasse la somme de ses parties, ce processus de fusion audio-visuelle pourrait expliquer les différences interindividuelles précédemment citées. A ce sujet, la recherche sur la perception bimodale de la parole est tenue à répondre à deux types d'exigences : (i) proposer une façon valide d'opérationnalisation du concept de la fusion audio-visuelle, autrement dit, proposer des mesures permettant d'évaluer quantitativement le phénomène ; (ii) théoriser sur la structure des systèmes de traitement impliqués, leurs relations, leur fonctionnement et les différents types de représentations produits lors du traitement cognitif permettant d'intégrer les informations apportées par les deux modalités, visuelle et auditive, dans un percept unique.

1.4.1 Mesures de la fusion audio-visuelle

1.4.1.1 Effet McGurk-MacDonald

Dans le champ de recherche portant sur la perception bimodale de la parole, le phénomène le plus fréquemment cité en lien avec la fusion audio-visuelle est l'effet McGurk-MacDonald, plus largement connu sous le nom de l'effet ou de l'illusion de McGurk (McGurk & MacDonald, 1976). L'expérience qui a permis de mettre en évidence l'effet en question comportait des stimuli (notamment des syllabes) audio-visuels, synchrones et normalement audibles mais non congruents (le signal visuel ne correspondait pas au signal acoustique), qui avaient été présentés aux participants qui devaient rapporter ce qu'ils avaient perçu. En

examinant les réponses des participants, McGurk et MacDonald (1976) se sont aperçus que pour certaines combinaisons audio-visuelles, les participants pouvaient percevoir une syllabe qui ne correspondait ni à la modalité visuelle ni à la modalité auditive du stimulus correspondant. Pour les auteurs, les deux inputs sensoriels auraient été fusionnés, autrement dit, leurs caractéristiques respectives auraient été intégrées dans un percept cohérent et différent de chacun des inputs unimodaux. Par exemple, pour des syllabes possédant une structure consonne-voyelle où la voyelle est une constante (il s'agit de la voyelle /a/), la présentation simultanée d'une consonne vélaire telle que /k/ (pour la syllabe /ka/) dans la modalité visuelle et d'une consonne bilabiale telle que /p/ (pour la syllabe /pa/) dans la modalité auditive amène généralement les personnes entendantes à percevoir une fusion illusoire entre les deux consonnes, plus précisément une consonne alvéo-dentale, notamment /ta/ (pour d'autres illustrations, voir Fowler & Dekle, 1991). Le principe de l'utilisation du paradigme McGurk-MacDonald pour évaluer les capacités de fusion audio-visuelle d'un individu suit le raisonnement suivant : plus la fréquence d'occurrence de fusion audio-visuelle en présence de stimuli de type McGurk-MacDonald est élevée, meilleures sont les capacités de l'individu à fusionner les inputs visuel et auditif lors de la perception bimodale de la parole (voir Grant, 2002).

En dehors de l'anglais, l'effet McGurk-MacDonald a été observé dans de nombreuses langues mondiales (e.g., Hayashi & Sekiyama, 1998 ; Sekiyama, 1997 ; Sekiyama & Tohkura, 1991 ; Tiippana, Tiainen, Vainio, & Vainio, 2013), ce qui montre ainsi la nature relativement générale du processus de fusion audio-visuelle lors de la perception bimodale de la parole. [5] Aussi, l'illusion de McGurk-MacDonald, résultant d'une incongruence inter-sensorielle, est aujourd'hui une illustration classique d'interactions audio-visuelles sur lesquelles reposerait la perception de la parole. Elle est largement utilisée pour étudier les mécanismes d'intégration audio-visuelle impliqués dans la fusion entre les deux modalités sensorielles lors de la perception d'un message parlé (e.g., Alsius, Navarra, & Soto-Faraco, 2007 ; Campbell, Dodd, & Burnham, 1998; de Gelder & Bertelson, 2003 ; Green, Kuhl, Meltzoff, & Stevens, 1991 ; Munhall, Gribble, Sacco, & Ward, 1996).

[5] Dans certaines langues telles que le japonais, l'effet McGurk-MacDonald est pourtant plus faible que dans les langues européennes (Sekiyama, 1997 ; voir aussi Tiippana, 2014). Un tel résultat suggère que l'effet McGurk-MacDonald n'est pas entièrement indépendant des facteurs linguistiques.

L'un des aspects critiques dans l'utilisation du paradigme de McGurk-MacDonald est l'interprétation des résultats. En effet, l'effet de McGurk-MacDonald est souvent présenté, comme le soulignent Soto-Faraco et Alsius (2009), en tant que preuve indirecte de la nature relativement automatique et involontaire de la fusion entre les deux inputs unimodaux lors de la perception bimodale de la parole impliquant des mécanismes d'intégration des deux sources de l'information pratiquement pré-attentionnels et donc non conscients. Une telle supposition nécessite que la fusion survienne très tôt dans le traitement de la parole dans sa forme bimodale. En effet, notre système cognitif résoudrait le conflit entre les deux signaux sensoriels par la mise en place d'un percept unifié, combinant les caractéristiques de l'input visuel et auditif, sans que l'on se rende compte de l'incongruence entre les deux types d'input (e.g., Kislyuk, Möttönen, & Sams, 2008; McGurk & MacDonald, 1976 ; voir aussi Summerfield & McGrath, 1984). Toutefois, les résultats de l'expérience conduite par Soto-Faraco et Alsius (2009) semblent aller à l'encontre d'une telle supposition (pour d'autres exemples d'études similaires, voir Bernstein, 2005). En effet, ayant utilisé le paradigme de McGurk-MacDonald avec des stimuli dans lesquels les signaux visuel et acoustique n'étaient pas synchronisés, les auteurs ont noté que malgré le fait que l'asynchronie des stimuli expérimentaux ait bien été perçue, l'illusion de fusion telle que décrite plus haut persistait. Aussi, il semblerait que, lors du processus de fusion audio-visuelle, au moins certaines caractéristiques des deux inputs unimodaux puissent être traitées de manière consciente. Par ailleurs, toute combinaison d'incongruence entre les inputs visuel et auditif ne se solde pas obligatoirement par un percept résultant d'une fusion audio-visuelle. En effet, dans certains cas, l'input dans la modalité visuelle peut influencer le percept final en l'emportant simplement sur l'input auditif. Par exemple, en présence de l'input auditif /epe/ et de l'input visuel /eke/, la voyelle la plus fréquemment perçue est /eke/ (Tiippana, Tiainen, Vainio, & Vainio, 2013). De ce fait, les mécanismes de traitement intermodal (audio-visuel) pourraient s'avérer bien plus complexes que ce qui a été initialement supposé par les auteurs ayant eu recours au paradigme de McGurk-MacDonald.

1.4.1.2 D'autres approches de l'évaluation de la fusion audio-visuelle : détection d'asynchronie audio-visuelle et modélisation

Une des caractéristiques importantes dont dépendrait la fusion d'inputs provenant de différentes modalités sensorielles en un percept unifié est la synchronie des inputs concernés (Bishop & Miller, 2009 ; Meredith, 2002 ; Miller & D'Esposito, 2005 ; Stevenson,

VanDerKolk, Pisoni, & James, 2010)[6]. La synchronie entre différents inputs serait en relation avec leur fusion, car c'est une des caractéristiques qui permettraient à l'organisme d'identifier les signaux qui proviennent d'un même évènement et peuvent ainsi être fusionnés. En effet, la recherche a établi que l'asynchronie entre deux signaux environnementaux est négativement corrélée avec la probabilité qu'ils soient fusionnés par la personne les percevant (Corney & Pisoni, 2006 ; van Wassenhowe, Grant, & Poeppel, 2007). Aussi, une autre mesure comportementale communément utilisée pour évaluer la fusion audio-visuelle lors de la perception bimodale de la parole est celle proposée dans le cadre du paradigme de détection d'asynchronie entre les inputs visuel et auditif (Grant & Seitz, 1998). La tâche utilisée ici consiste en présentation des stimuli audo-visuels dans lesquels les deux types d'inputs présentent différents degrés d'asynchronie ; l'objectif des participants étant de détecter la non correspondance temporelle des deux inputs, ce qui permet d'évaluer le seuil individuel de détection d'asynchronie audio-visuelle. Selon le raisonnement de Grant et Seitz (1998), les personnes ayant de meilleures capacités de fusion audio-visuelle devraient présenter des seuils de détection d'asynchronie audio-visuelle plus bas. Toutefois, Grant et Seitz (1998) ont trouvé que les seuils individuels de détection d'asynchronie n'étaient pas corrélés avec l'ampleur du gain audio-visuel, tel qu'évalué dans par le paradigme de la dégradation de l'information auditive par le bruit, et ceci que les stimuli soient des syllabes ou des phrases. Certes, le gain audio-visuel ne peut pas être clairement assimilé aux capacités de fusion audio-visuelle de l'individu. Toutefois, d'autres études, telle que celle de Soto-Faraco et Alsius (2009), citée dans la section précédente, dans laquelle les participants procédaient à une fusion audio-visuelle avec des stimuli de type McGurk-MacDonald dans lesquels ils ont pu identifier une asynchronie entre les inputs visuel et auditif. Des données similaires obtenues récemment par Vroomen et Stekelenburg (2011) mettent à mal la validité de la détection d'asynchronie audio-visuelle en

[6] Notons que la synchronie entre les inputs auditif et visuel est un des aspects fondamentaux de la perception bimodale. Cet aspect concerne la règle de temporalité, qui est une des trois règles majeures dans le domaine en question. Selon la règle de temporalité, la perception multimodale est plus probable et/ou plus efficace en présence de synchronie entre les différents inputs unimodaux (Meredith & Stein, 1986). Les deux autres règles sont : (i) la règle spatiale selon laquelle la perception multimodale est plus probable et/ou plus efficace quand les différents inputs sensoriels proviennent de la même direction/d'une même source spatiale (Meredith, Nemitz, & Stein, 1987), et (ii) le principe d'efficacité inversée (pour plus de détails, revoir la section 1.2).

tant que mesure des capacités de fusion audio-visuelle lors de la perception bimodale de la parole.[7]

A côté des tâches comportementales, un autre type d'approches de l'évaluation des capacités individuelles de fusion audio-visuelle ou encore de l'efficience de l'intégration audio-visuelle (voir Grant, 2002) est la modélisation. Dans cette veine de recherche on classe les modèles dont l'objectif est d'expliquer et de prédire le gain audio-visuel, c'est-à-dire les capacités de l'individu à reconnaître correctement un message parlé bimodal comparé à la reconnaissance de l'input auditif seul. Dans ces modèles, le gain audio-visuel d'un individu est prédit à partir de ses capacités de reconnaissance unimodale, relative à l'input auditif seul et à l'input visuel seul. Par exemple, le modèle PROB de Blamey, et *al*. (1989) est un modèle probabiliste simple qui explique les erreurs de reconnaissance dans la condition bimodale par l'incapacité à reconnaître l'input verbal aussi bien dans sa modalité visuelle que dans sa modalité auditive. Toutefois, les gains audio-visuels réels sont globalement plus élevés que ceux prédits par le modèle PROB. Inversement, dans le modèle de Braida (1991) (PRE (*Pre-Labeling Integration*)) l'individu est supposé traiter les inputs visuel et auditif de manière optimale. Le modèle PRE est basé sur les notions de la théorie de la détection multimodale du signal. Ses prédictions du gain audio-visuel sont généralement trop élevées par rapports aux gains audio-visuels réels. Finalement, dans le modèle de Massaro (1998) (FLMP (*Fuzzy Logical Model of Perception*)), l'individu est également supposé faire un usage optimal des inputs visuel et auditif. Toutefois, les prédictions de ce modèle, qui intègre des algorithmes conformes à la théorie de la logique floue, sont globalement trop basses par rapport aux observations réelles (pour plus de détails, voir Altieri, 2010).

La précision avec laquelle les trois modèles prédisent les différences interindividuelles dans le gain audio-visuel a été évaluée par Grant et *al.*, (1998) sur des sujets malentendants. En prenant en compte les gains audio-visuels prédits par chaque modèle et les gains audio-visuels réels, les auteurs ont trouvé une corrélation positive pour les trois modèles. Les gains prédits étaient explicatifs d'une partie modérée à élevée de la variabilité totale des gains réels. De tels

[7] Dans ce contexte, il convient de noter que la notion de fenêtre temporelle d'intégration (*temporal window of intgration*) a été proposée en relation avec le phénomène de fusion audio-visuelle en absence de synchronie entre les deux inputs sensoriels (e.g., Lewkowicz, 1996, 2000 ; van Wassenhove, Grant, & Poeppel, 2007). En ce qui concerne la perception bimodale de la parole, les inputs auditif et visuel tendent à être fusionnés s'ils sont présentés à l'intérieur de la fenêtre d'environ 200ms, située entre 30ms d'avance du signal auditif sur le signal visuel et 170ms de retard du signal auditif sur le signal visuel (van Wassenhove et *al.*, 2007).

résultats pourraient accorder un certain crédit aux modèles de la fusion audio-visuelle. Toutefois, le degré d'exactitude des prédictions d'un modèle donné n'est pas forcément révélateur de la validité des suppositions faites au sujet des principes/règles sur lesquels/lesquelles reposent les mécanismes de l'intégration audio-visuelle. Ce point semble être assez problématique dans les modèles PROB, PRE et FLMP, car ils ont été conçus avec l'objectif de simuler les capacités individuelles de perception bimodale de la parole de manière statique, c'est-à-dire à partir des valeurs correspondantes aux capacités individuelles à reconnaître un message parlé dans sa forme auditive et visuelle seule, sans vraiment rendre compte de la complexité et de l'aspect dynamique du traitement cognitif sous-jacent à ces mesures comportementales. Aussi, les prédictions des modèles en question ne peuvent pas être considérées comme étant relatives aux caractéristiques structurelles et fonctionnelles d'un système cognitif qui traite la parole bimodale et produit une réponse comportementale, ce qui limite évidemment leur intérêt.

1.4.2 Modèles de la fusion audio-visuelle

En parlant de la fusion audio-visuelle, l'objectif ultime des sciences cognitives est de rendre compte, d'une part, des représentations impliquées dans le processus donnant lieu au phénomène en question et, d'autre part, des mécanismes qui constituent le traitement cognitif ayant pour résultat un percept cohérent. Pour ce faire, des données comportementales, neurophysiologiques, celles provenant de l'utilisation de l'imagerie cérébrale, essentiellement fonctionnelle, ainsi que des observations cliniques sont communément prises en compte. Finalement, des modèles de la structure et du fonctionnement cognitifs sont proposés. Ils devraient pouvoir rendre compte des différents types d'observations faites au sujet du phénomène à théoriser. Cette procédure a également été suivie dans la mise en place des modèles de la fusion audio-visuelle.

Les premières explications de la perception audio-visuelle de la parole ont été élaborées par Summerfield (1987). En effet, cet auteur a proposé quatre alternatives conceptuelles au phénomène, centrées essentiellement sur les différents types de représentations mises en place lors du traitement bimodal de la parole. Selon Summerfield (1987), le processus de la perception bimodale de la parole pourrait donner lieu à deux types de représentations audio-visuelles et motrices. Pour cet auteur, le processus de la mise en place des représentations audio-visuelles pourrait être basé sur (i) l'intégration des composantes acoustiques discrètes dans lesquelles l'information sur le point d'articulation est apportée par la modalité visuelle et l'information

sur le mode d'articulation par la modalité auditive (voir la section 1.3.2 pour plus de détails) ; (ii) l'intégration des composantes visuelle et auditive pouvant être représentées mathématiquement sous forme de vecteurs décrivant les paramètres visuels et auditifs en tant que valeurs indépendantes. Contrairement aux représentations audio-visuelles qui prennent en compte les caractéristiques des inputs visuel et auditif, les représentations motrices seraient mises en place à partir d'un input somesthésique issu de la réalisation motrice des sons à percevoir dont on garde les traces mnésiques. Plus précisément, Summerfield (1987) parle (i) des représentations abstraites en relation avec les organes du tract vocal jouant un rôle de filtre lors de la production verbale orale ; (ii) des représentations sur les aspects dynamiques de la production articulatoire.

Les propositions de Summerfield (1987) ont donné lieu à deux catégories majeures de théorisations sur la perception audio-visuelle de la parole. La première catégorie consiste ainsi en modèles expliquant la perception de la parole par la mise en place de représentations relativement abstraites, non dépendantes des modalités visuelle et auditive. En l'occurrence, il s'agirait des représentations sur les mouvements articulatoires impliqués dans la production des sons d'une langue. Le principe d'encodage de la parole en « monnaie commune », modalité non-dépendante, lors de la perception de la parole est l'idée centrale dans la théorie du codage commun ou la théorie de la perception amodale (Massaro, 2004 ; Rosenblum, 2005 ; voir aussi Rosenblum, 2008) et la théorie motrice de la perception de la parole (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967 ; Liberman & Mattingly, 1985 ; pour une revue, voir Galantucci, Fowler, & Turvey, 2006). Les modèles de la seconde catégorie supposent au contraire que la parole est perçue à partir des informations provenant des deux canaux sensoriels séparés, auditif et visuel. Ces deux inputs seraient encodés en représentations dépendantes des deux modalités en question. La première étape du traitement bimodal de la parole correspondrait à l'encodage sensoriel du message parlé et son traitement dans des régions du cortex sensoriel primaire, auditif et visuel. A ce niveau, le traitement des deux types d'information se déroulerait en parallèle. Suit la seconde grande étape dans la perception bimodale de la parole qui correspondrait à la reconnaissance d'un élément au niveau du lexique mental. A un certain moment entre les deux étapes, le traitement perceptif consisterait à intégrer les deux sources d'informations en représentations plus complexes, audio-visuelles. Finalement, une boucle de feed-back partirait de l'étape de la reconnaissance lexicale vers les étapes ultérieures du traitement, phonologique et visuelle. Ainsi, l'information lexicale restituée du lexique mental représenterait en elle-même une contrainte pour le traitement phonologique

et visuel de la suite du message. Les différents types de données recueillies au sujet de la perception audio-visuelle de la parole concordent davantage avec cette deuxième catégorie de modèles qui sont généralement considérés comme dominants (pour plus de détails, voir Peelle & Sommers, *in press*).

Les différents modèles expliquant la perception bimodale de la parole par la mise en place des représentations phonologiques et visuelles concordent dans les points conceptuels exposés ci-dessus. En revanche, ils diffèrent quant au moment de l'intégration entre les informations auditive et visuelle. En effet, les modèles de l'intégration tardive (Grant & Seitz, 1998) stipulent que lors de la première étape, les traitements phonologique et visuel se déroulent de manière entièrement indépendante. L'intégration entre les représentations phonologiques et visuelles se ferait une fois que celles-ci sont mises en place à l'issue de la première étape du traitement. Inversement, les modèles de l'intégration précoce (Schroeder & Foxe, 2005) stipulent l'existence d'interactions à la première étape du processus, au niveau des traitements parallèles, phonologique et visuel. L'intégration entre les informations apportées par les deux inputs donnerait ainsi directement lieu à une seule représentation complexe, audio-visuelle (pour une présentation schématique des modèles, voir la Figure 5).[8]

Les modèles de l'intégration audio-visuelle précoce sont concordants avec la découverte de connexions fonctionnelles entre le cortex auditif et visuel primaire, ainsi que celle de la modulation de l'activité du A1 par le traitement au niveau du cortex visuel primaire (V1) lors de la perception bimodale de la parole (ces données sont exposées dans la section 1.5.1). Par ailleurs, certaines données suggérant des interactions audio-visuelles précoces ont également été recueillies dans le cadre de la perception des stimuli non verbaux (Barutchu, Crewther, & Crewther, 2009 ; Barutchu, Danaher, Crewther, Innes-Brown, Shivdasani, & Paolini, 2010 ; Berryhill, Kveraga, Webb, & Hughes, 2007 ; Miller, 1986 ; etc.). En revanche, le phénomène de persistance de l'effet McGurk-MacDonald avec des stimuli audio-visuels présentant une asynchronie détectable pour les sujets (e.g., Soto-Faraco & Alsius, 2009) ne peut s'expliquer que dans le cadre des modèles de l'intégration tardive. Ce sont également les modèles de l'intégration tardive qui semblent mieux rendre compte d'une certaine indépendance constatée entre les performances en perception bimodale, audio-visuelle, et les performances en perception unimodale (visuelle et auditive) telles que mesurées par le paradigme de la

[8] Cette partie décrit le cadre conceptuel général des modèles de l'intégration audio-visuelle. Pour plus de détails sur d'autres variantes conceptuelles au sujet du processus de mise en place des représentations impliquées dans la perception de la parole, voir Altieri, Pisoni et Townsend (2011), et Dohen (2009).

dégradation de l'information auditive par le bruit (Grant & Seitz, 1998). Aussi, certains auteurs (Altieri et *al.*, 2011 ; Peelle & Sommers, 2015) ont été amenés à proposer des modèles hybrides qui combinent aussi bien une première étape d'intégration précoce qu'une seconde étape d'intégration tardive. La recherche future dans le domaine de la perception bimodale de la parole sera probablement centrée sur les caractéristiques dynamiques du traitement. Pour de telles problématiques, les données neurophysiologiques, offrant une résolution temporelle élevée de la mesure, sont particulièrement prometteuses. Les propositions de nouvelles tâches expérimentales, élaborées pour évaluer les aspects dynamiques du traitement au lieu de la précision perceptive seule, à partir d'indices comportementaux (Altieri et *al.*, 2011 ; voir aussi Altieri & Townsend, 2011), peuvent également porter leurs fruits.

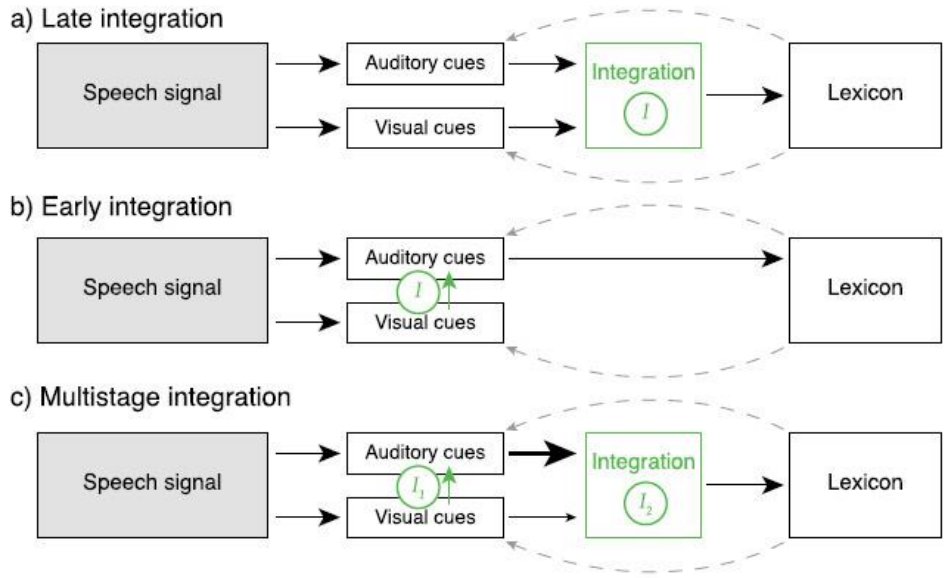

Figure 5. Représentation schématique des trois principales classes des modèles expliquant l'intégration audio-visuelle (Peelle & Sommers, *in press*).

Selon la classe de modèle, l'intégration audio-visuelle (*I*) est située à des endroits différents du traitement du message parlé (*Speech signal*) qui implique deux étapes clefs : (i) le traitement du message dans ses modalités sensorielles, auditive (*Auditory cues*) et visuelle (*Visual cues*), (ii) l'accès au lexique mental (*Lexicon*). En (a), le modèle de l'intégration tardive situe l'intégration audio-visuelle après un traitement parallèle des composantes sensorielles élémentaires de la parole, alors que le modèle de l'intégration précoce en (b) la situe durant ce même traitement. Finalement, le modèle hybride en (c) prévoit une possibilité d'intégration

audio-visuelle durant et après cette première étape du traitement des inputs auditif et visuel. (L'image de Peelle et Sommers (*in press*).)

1.5 Corrélats neuronaux de la perception audio-visuelle de la parole

La recherche en neuroscience et en neurophysiologie dans le domaine de la perception bimodale de la parole a pour objectif d'identifier les corrélats et les mécanismes neuronaux qui y sont impliqués, avec un intérêt particulier pour le mécanisme de l'intégration audio-visuelle. L'étude des corrélats neuronaux est réalisée par la méthode de l'imagerie cérébrale fonctionnelle. Elle permet d'évaluer l'activité métabolique du cerveau à partir de la mesure du flux de sang apportant de l'oxygène et du glucose au cerveau. De manière générale, un flux sanguin accru dans une région donnée signifierait une consommation élevée en oxygène par cette région et, à termes, une implication de la région en question dans le traitement en cours[9]. L'étude des corrélats neuronaux qui sous-tendent le traitement de la parole dans sa forme bimodale consiste à comparer le flux sanguin dans certaines régions cérébrales entre les conditions de perception unimodale (auditive seule et visuelle seule) et la condition de perception bimodale. Dans ce cadre, un effet supra additif, c'est-à-dire un taux de flux sanguin plus important dans la condition de perception bimodale que la somme des taux de flux sanguin mesurés dans les conditions de perception unimodale, est toujours interprété comme signifiant l'implication de la région en question dans l'intégration audio-visuelle de la parole. La région en question est censée répondre préférentiellement à la parole bimodale. En revanche, l'interprétation d'un effet sous-additif (un taux de flux sanguin moins important dans la condition de perception bimodale que la somme des taux de flux sanguin mesurés dans les conditions de perception unimodale) est plus problématique. Cependant, la plupart des auteurs considèrent actuellement qu'il signifie également l'implication de la région en question dans le traitement bimodal de la parole.

[9] Dans la technique d'imagerie cérébrale fonctionnelle la plus utilisée, celle d'imagerie par résonance magnétique nucléaire (IRMf), la quantité du flux sanguin dans une région cérébrale est induite à partir des indices mesurant le taux d'oxygène qui augmente avec l'augmentation du flux sanguin. On parle de la réponse BOLD (*blood oxygen-level dependent response*). Une autre technique de l'imagerie cérébrale fonctionnelle, la tomographie par émission de positions (TEP), permet de mesurer l'activité métabolique du cerveau à partir du traçage du glucose dans le sang par le biais d'un marqueur radioactif.

Si les données de l'imagerie cérébrale fonctionnelle permettent d'identifier les régions potentiellement impliquées dans l'intégration audio-visuelle dans la perception de la parole, ou au moins qui participent au traitement audio-visuel de la parole, grâce à leur résolution spatiale élevée, les données neurophysiologiques sont essentiellement utilisées pour étudier les aspects dynamiques du fonctionnement cérébral. En effet, reflétant les changements dans les potentiels post-synaptiques d'une région cérébrale, ces données présentent une résolution temporelle plus élevée que celles de l'imagerie cérébrale. Il existe deux types d'approches de recueil de données neurophysiologiques. La première résulte de la mesure de l'activité globale d'une région cérébrale (potentiels des champs locaux et électroencéphalogramme (EEG)) lors d'une activité donnée se caractérisant par sa fréquence, son amplitude et son emplacement sur le scalpe où elle est enregistrée. La seconde approche résulte de la mesure de l'activité induite par un stimulus précis (potentiels évoqués) se caractérisant par sa latence d'occurrence, son amplitude et son emplacement sur le scalpe. Dans l'étude de la perception audio-visuelle de la parole, les deux approches sont utilisées.

Les études utilisant l'approche de l'imagerie cérébrale fonctionnelle et celles se basant sur les données neurophysiologiques ont mis en évidence l'implication de deux régions cérébrales majeures dans le traitement bimodal de la parole, le A1 et le sillon temporal supérieur. De manière moins concordante, certaines régions motrices sont également citées comme étant impliquées dans la perception audio-visuelle de la parole. Les études en question et les conclusions pouvant être tirées des résultats produits sont présentées ci-dessous.

1.5.1 Cortex auditif primaire (A1)

Le A1 est une région corticale qui assure le traitement perceptif de l'input auditif véhiculé jusqu'au cerveau par le nerf auditif. Le A1, ainsi que les autres régions sensorielles primaires du cerveau, était initialement conçu comme recevant et traitant uniquement l'input sensoriel correspondant, en l'occurrence acoustique. Néanmoins, les études anatomiques menées sur des animaux ont établi que les régions sensorielles primaires sont en réalité connectées entre elles (Campi, Bales, Grunewald, & Krubitzer, 2010 ; Cappe & Barone, 2005; Clavagnier; Falchier, & Kennedy, 2004 ; Falchier, Clavagnier, Barone, & Kennedy, 2002 ; Rockland & Ojima, 2003; Rockland & Van Hoesen, 1994). Un fait important pour la perception bimodale de la parole dans ce cadre est la connexion entre les cortex primaires auditif et visuel (Eckert, Kamdar, Chang, Beckmann, Greicius, & Menon, 2008 ; Falchier et *al.*, 2002; Rockland & Ojima, 2003). Sur le plan anatomique, tout semble donc prêt pour une communication

intermodale précoce lors de la perception audio-visuelle de la parole. Les mécanismes neurologiques d'un tel traitement pourraient s'approcher de ceux mis en évidence par des études neurophysiologiques s'intéressant à l'activité globale du A1, lors du traitement d'un message parlé bimodal.

L'activité neurophysiologique globale des différentes régions cérébrales est profondément rythmique, reflétant les changements entre les états de haute et faible excitabilité des groupes neuronaux (Buzsaki & Draguhn, 2004 ; Raichle, 2010 ; Schroeder & Lakatos, 2009 ; Sporns, 2011 ; Stone & Hughes, 2013 ; Thut, Miniussi, & Gross, 2012). Les études neurophysiologiques montrent que l'activité rythmique dans l'excitabilité neuronale dans différentes régions du cerveau joue un rôle important dans l'efficacité du traitement des stimuli environnementaux. En effet, le traitement le plus efficace est observé dans le cas où l'arrivée du stimulus concorde avec l'état de haute excitabilité neuronale. C'est notamment dans les conditions de haute excitabilité que la probabilité qu'un stimulus déclenche une réponse neuronale est la plus élevée. En revanche, le traitement est le moins efficace est observé dans le cas où le stimulus survient lors de l'état d'une excitabilité neuronale basse (e.g., Henry & Obleser, 2012 ; Schroeder, Wilson, Radman, Scharfman, & Lakatos, 2010; VanRullen, Busch, Drewes, & Dubois, 2011 ; voir aussi la Figue 6). Une caractéristique importante des oscillations dans l'activité neuronale est l'adaptation de leur phase à la fréquence de l'occurrence des stimuli environnementaux si cette dernière est prévisible, car rythmique (e.g., Cravo, Rohenkohl, Wyart, & Nobre, 2013 ; Henry & Herrmann, 2014 ; Henry, Herman, & Obleser, 2014 ; Schroeder et *al.*, 2010). Les modulations dans la phase de l'activité rythmique de groupes neuronaux tendent à établir un pattern d'activation neuronale tel que les moments d'occurrence des stimuli à traiter correspondent aux moments de haute excitabilité (pour une revue, voir Calderone, Lakatos, Butler, & Castellanos, 2014 ; Schroeder & Lakatos, 2009)[10]. C'est ainsi que ce mécanisme neuronal d'ajustement de phase entre la structure des stimuli environnementaux et l'activité neuronale est actuellement considéré comme relevant de l'attention sélective qui nous permet de traiter de manière optimale certains stimuli

[10] Dans la littérature anglo-saxonne, ce phénomène d'ajustement de phase entre les oscillations dans l'excitabilité d'une région cérébrale et la structure rythmique des stimuli environnementaux est connu sous le nom de *entrainment.*

environnementaux tout en ignorant ou atténuant les autres (Lakatos, Karmos, Mehta, Ulbert, & Schroeder, 2008).[11]

Le phénomène d'ajustement de phase des oscillations dans l'excitabilité neuronale en fonction de la structure rythmique des stimuli environnementaux a été également observé pour la perception de la parole au niveau du A1. En effet, lors de la perception auditive de la parole, la phase de l'activité oscillatoire du A1 est ajustée en fonction des pics dans l'amplitude du signal auditif dont la structure rythmique correspond grossièrement à la structure syllabique des unités produites (Ghitza, 2011 ; Giraud & Poeppel, 2012 ; voir aussi Ghitza, Giraud, & Poeppel, 2013). Selon Schroeder et *al*. (2008), l'input visuel dans le cas de la perception bimodale de la parole pourrait également affecter l'activité rythmique du A1 grâce à sa valeur prédictive quant à l'arrivée de l'input auditif. C'est ce qui expliquerait, à terme, l'effet facilitateur de l'input bimodal lors de la perception de la parole dans des conditions de bruit acoustique (voir la Figure 6 pour une illustration schématique du phénomène présenté).[12]

Outre la connexion anatomique entre le A1 et le V1 mentionnée plus haut, il existe en effet plusieurs arguments en faveur de cette hypothèse. Premièrement, des études sur des animaux, des primates, montrent que l'activité oscillatoire du A1 peut être modulée par des stimuli visuels (Kayser et *al.*, 2008 ; Perrodin et *al.*, 2015). Deuxièmement, dans les conditions où la perception de la parole est soumise à une charge attentionnelle élevée, comme par exemple dans des situations qui intègrent plusieurs personnes parlant simultanément, alors que notre objectif est de n'en écouter qu'une, un input bimodal, audio-visuel, permet un meilleur ajustement entre l'activité oscillatoire du A1 et la structure rythmique du message cible qu'un input auditif seul (Lou, Liu, & Poeppel, 2010).

[11] Notons également que les oscillations dans l'excitabilité neuronale sont catégorisées en fonction de leur fréquence, chaque catégorie étant associée à un ensemble de traitements ou fonctions cognitives. Ce sujet dépasse cependant l'objectif du présent texte et ne sera pas développé davantage (pour plus de détails, voir Schroeder & Lakatos, 2009).

[12] Toutefois, voir Schwartz et Savariaux (2014) (section 1.3.1) au sujet des relations temporelles entre les inputs visuel et auditif dans un message parlé.

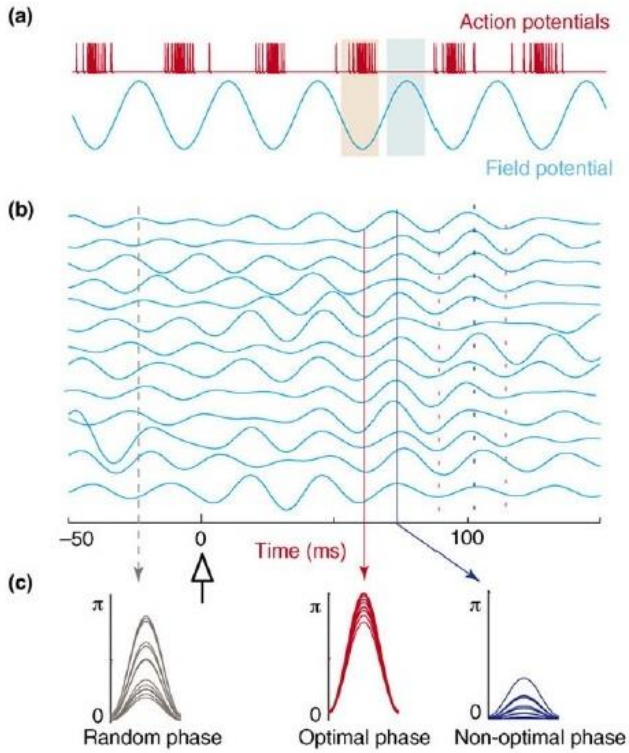

Figure 6. Représentation schématique du mécanisme d'ajustement de phase des oscillations dans l'excitabilité neuronale du A1 et l'occurrence des stimuli environnementaux induit par un stimulus visuel (Schroeder et *al*., 2008).

En (a), la relation entre les variations dans l'excitabilité d'un ensemble neuronal (en rouge) correspondant aux potentiels d'action (*Action potentials*) et les variations dans le champ électrique d'une région cérébrale (*Field potential*) telles qu'enregistrées par l'EEG sur le scalpe. (Notons toutefois que l'EEG enregistre les variations dans le champ électrique engendrées par les potentiels post-synaptiques qui sont une conséquence des potentiels d'action.) En (b), une série simulée d'enregistrements EEG au niveau du A1 pour des essais à un seul stimulus de nature visuelle avant et après la présentation du stimulus au temps 0 (indiqué par la flèche). Avant la présentation du stimulus, les variations dans le champ électriques sont aléatoires (ligne à traits en gris correspondant à la représentation graphique de *Random phase* en (c)). L'arrivée du stimulus visuel entraîne un ajustement de phase dans l'activité neuronale – les moments de haute excitabilité neuronale (ligne pleine en rouge correspondant à *Optimal phase* en (c)) et de faible excitabilité neuronale (ligne pleine en bleu correspondant à *Non-optimal phase* en (c)) s'alignent sur l'ensemble des essais et restent alignés (lignes rouges et bleues à traits) durant un certain délai succédant l'arrivée du stimulus. Si d'autres inputs sensoriels sont présentés, la réponse du système diffère en fonction de la phase de l'excitabilité neuronale. Un stimulus arrivant durant la phase aléatoire (*Random phase*) engendre des réponses hautement variées, un

stimulus arrivant durant la phase d'excitabilité optimale (*Optimal phase*) est amplifié, alors que celui qui arrive durant la phase d'excitabilité non optimale (*Non-optimal phase*) est supprimé/n'est pas traité par le système. Cet exemple est illustratif de la position de Schroeder et *al.*, 2008 selon laquelle l'input visuel pourrait affecter le traitement du signal acoustique lors de la perception bimodale de la parole. (L'image de Schroeder et *al.* (2008).)

Ces données laissent ainsi supposer que le A1 est un lieu des interactions audio-visuelles de bas niveau, apparaissant très tôt dans le traitement bimodal de la parole. Ceci est soutenu davantage par les résultats des études utilisant l'approche des potentiels évoqués afin d'évaluer un éventuel effet de l'input visuel sur le traitement de l'input auditif lors de la perception bimodale de la parole (revoir la section 1.3.1). De telles études ont en effet montré qu'un input bimodal modulait l'amplitude et la latence des ondes N1 et P2 qui apparaissent à 100 ms et 200 ms respectivement après l'occurrence du signal acoustique, au niveau du A1. En effet, contrairement à l'input auditif seul, ou encore comparé à la somme des effets observés dans les conditions de présentation unimodale, auditive et visuelle, l'input bimodal entraîne une diminution des deux paramètres des ondes en question (Alsius et *al.*, 2014 ; Besle et *al.*, 2004 ; van Wassenhove et *al.*, 2005). La réduction dans la latence des composantes est considérée comme un marqueur d'accélération du traitement, alors que la réduction de l'amplitude marque l'efficacité du traitement bimodal car moins de ressources neuronales sont recrutées pour obtenir une performance meilleure. Plus précisément, van Wassenhove et *al.* (2005) ont noté que la réduction dans l'amplitude de la N1 était corrélée avec la valeur prédictive de l'input visuel quant à l'arrivée du signal acoustique (voir aussi Baart, Stekelenburg, & Vroomen 2014 ; Brandwein, Foxe, Russo, Altschuler, Gomes, & Molholm, 2011 ; Stekelenburg & Vroomen, 2007). En revanche, ces mêmes auteurs rapportent qu'une diminution dans la latence de la N1 est relative à la valeur informative du signal visuel quant aux caractéristiques acoustiques d'un son. Finalement, des changements au niveau de la P2 (latence et/ou amplitude), refléteraient, quant à eux, l'intégration audio-visuelle (Baart et *al.*, 2014). Aussi, l'input visuel semble accélérer le traitement du signal acoustique par le A1 lors de la perception bimodale de la parole et possiblement faciliter l'intégration audio-visuelle à ce niveau. Ceci se produit dans une fenêtre temporelle très courte, dans les premières 200ms du traitement. Les latences extrêmement courtes des effets décrits ci-dessus laissent supposer que l'input visuel affecte l'activité du A1 par le biais d'une connexion directe entre cette région et le cortex visuel

primaire, plutôt que par une connexion indirecte, passant par des régions associatives, qui nécessiterait probablement des délais plus longs (voir Arnal, Morillon, Kell, & Giraud, 2009).

Finalement, une étude utilisant la méthode de l'imagerie cérébrale par résonance magnétique nucléaire fonctionnelle (IRMf) a mis en évidence que le cortex auditif pouvait être activé par les seuls indices visuels lors de la tâche de la lecture de la parole sur les lèvres de l'orateur (Calvert, Bullmore, Brammer, Campbell, Iversen, Woodruff, … David, 1997). Malgré le fait que la région cérébrale activée ici ne soit pas le A1, qui ne s'active qu'en présence d'inputs acoustiques, ce résultat confirme davantage le rôle du cortex auditif dans le traitement bimodal de la parole. Toutefois, la résolution temporelle de la mesure de l'IRMf étant faible, les aspects dynamiques de cette activation – essentiellement le délai de son occurrence – ne peuvent pas être établis. Il est néanmoins probable que le mécanisme impliqué dans le phénomène observé par Calvert et *al.* (1997) ait été induit par une connexion corticale de type feed-back partant des régions corticales associatives censées jouer un rôle important dans l'intégration des inputs visuel et auditif dans la perception bimodale de la parole, telle que le sillon temporal supérieur. De manière globale, les études présentées ci-dessus mettent en évidence l'existence d'interactions intermodales au niveau du cortex auditif, essentiellement primaire, lors de la perception bimodale de la parole, suggérant ainsi son implication dans le traitement bimodal de la parole.

1.5.2 Sillon temporal supérieur (STS)

Une des premières régions corticales ayant été identifiée comme étant impliquée dans le traitement multimodal est le sillon temporal supérieur (STS) et notamment de sa partie postérieure. En effet, les études menées sur des primates non humains ont mis en évidence la présence de neurones répondant à des stimuli audio-visuels ou encore somato-visuels dans cette région cérébrale (e.g., Benevento, Fallon, Davis, & Rezak, 1977 ; Bruce, Desimone, & Gross, 1981 ; Hikosaka, Iwai, Saito, & Tanaka, 1988). Chez les animaux, les études anatomiques ont également établi que le STS est connecté aussi bien avec le A1 qu'avec les aires visuelles extrastriées (Saleem, Suzuki, Tanaka, & Hashikawa, 2000; Seltzer & Pandya, 1994). Finalement, chez les humains, la région en question semble être impliquée dans le traitement des stimuli aussi bien visuels qu'auditifs (Beauchamps, Argall, Bodurka, Duyn, & Martin, 2004 ; Beauchamp, Lee, Argall, & Martin, 2004) et a été fréquemment identifiée comme présentant des réponses supra additives à la parole audio-visuelle, qu'il s'agisse de syllabes isolées (Baum, Martin, Hamilton, & Beauchamp, 2012 ; Calvert, Campbell, & Brammer, 2000 ;

Sekiyama, Kanno, Miura, & Sugita, 2003 ; Vander Wyk, Ramsay, Hudac, Jones, Lin, Klin, … Pephly, 2010), de mots (Stevenson & James, 2009) ou encore de phrases (Yi, Smiljanic, & Chandrasekaran, 2014).

Les effets supra aditifs observés au niveau du STS suggèrent que cette région possède des neurones qui répondent de manière maximale à des stimuli multimodaux, notamment, lors de la perception de la parole, à des stimuli audio-visuels. Ceci accorde au STS le statut d'une des régions majeures du traitement bimodal de la parole, assurant très probablement l'intégration des inputs auditif et visuel. Toutefois, cette conclusion n'est pas acceptée de manière unanime pour trois raisons principales. Tout d'abord, l'IRMf possédant une faible résolution temporelle de mesure, les aspects dynamiques exacts dans le flux sanguin vers le STS restent inconnus. Ensuite, l'IRMf permet de mesurer le flux sanguin d'une région cérébrale donnée ce qui n'est qu'une mesure très indirecte du traitement assuré par les neurones de cette région. Enfin, les effets supra additifs aux messages parlés audio-visuels pourraient être dus à la présence de neurones unimodaux, visuels et auditifs, entremêlés (Bernstein, Auer, & Moore, 2004 ; Meredith, 2002). Hormis ces trois objections quant à une interprétation possiblement trop hâtive des données précédemment exposées, une autre théorisation possible du rôle du STS dans la perception bimodale de la parole a été proposée récemment par Nath et Beauchamp (2011). Ces auteurs se sont intéressés à la relation entre la fiabilité unimodale, auditive ou visuelle, du signal bimodal lors de la perception de la parole et la connectivité fonctionnelle[13] du STS avec le cortex visuel et le cortex auditif. Ayant dégradé respectivement l'input visuel et l'input auditif dans des mots présentés de manière bimodale, Nath et Beauchmap (2011) ont observé une plus grande connectivité du STS avec le cortex traitant le signal le plus informatif/le plus fiable, notamment avec le cortex visuel en cas de dégradation de l'input auditif seul et, inversement, avec le cortex auditif en cas de dégradation de l'input visuel seul. (Voire la Figure 7 pour une présentation schématique de la connectivité fonctionnelle des différentes régions cérébrales en fonction des caractéristiques du message parlé telle qu'étudiée par Nath et Beauchamp (2011).)

Les résultats de Nath et Beauchamp (2011) concordent avec ceux des études comportementales montrant également une pondération d'inputs sensoriels en cas de

[13] La connectivité fonctionnelle est relative à la force des connexions anatomique entre différentes régions cérébrales qui peut varier en fonction de facteurs tels que la nature de la tâche à réaliser, les caractéristiques des stimuli à traiter, le type de réponse à produire ; etc. (voir Buchel & Friston, 2001 ; de Marco, Vrignaud, Destrieux, de Marco, Testelin, Devauchelle, & Berquin 2009 ; McIntosh & Gonzalez-Lima, 1994).

différences de fiabilité dans la perception multimodale (e.g., Alais & Burr, 2004 ; Ernst & Banks, 2002 ; Ma et *al.*, 2009) et offrent ainsi une explication alternative quant au rôle du STS dans la perception bimodale de la parole. En effet, au lieu d'être considéré comme une structure assurant l'intégration audio-visuelle, le STS pourrait en réalité être impliqué dans la gestion de l'allocation des ressources attentionnelles au traitement des inputs visuel et auditif. Aussi, si l'implication du STS dans la perception bimodale de la parole fait objet d'un consensus général, son rôle exact n'est pas certain à l'état actuel de la recherche. Des études supplémentaires seront nécessaires pour l'élucider.

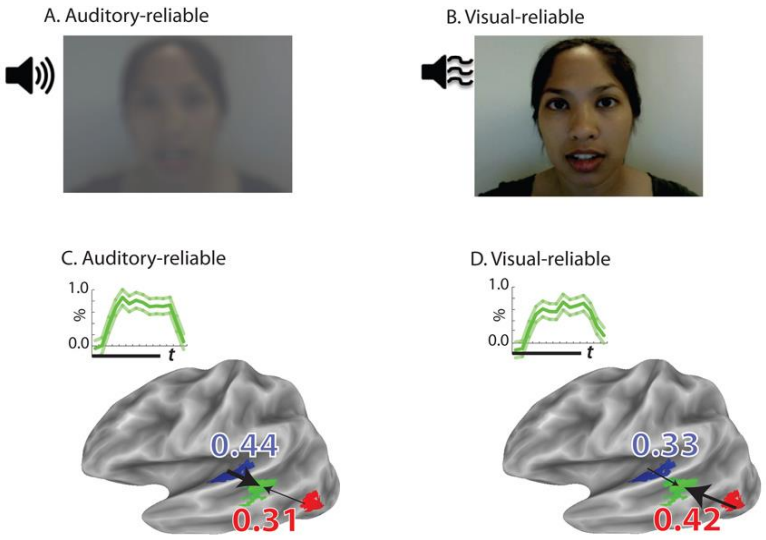

Figure 7. Représentation schématique de la connectivité fonctionnelle entre différentes régions cérébrales lors de la perception bimodale de la parole en fonction de la fiabilité des inputs auditif et visuel (Nath & Beauchamp, 2011).

La section supérieure de l'image représente les stimuli dégradés dans leur aspect visuel, étant ainsi fiable sur le plan acoustique (*Auditory-reliable*) (en A), et dans leur aspect auditif, étant ainsi fiable sur le plan visuel (*Visual-reliable*) (en B). La section inférieure de l'image représente les coefficients de connectivité fonctionnelle entre le STS (en vert) et le A1 (en bleu ; le coefficient en bleu) et le V1 (en rouge ; le coefficient en rouge) respectivement. On note que la connectivité fonctionnelle change en fonction des caractéristiques audio-visuelles du stimulus. (L'image de Nath et Beauchamp (2011).)

1.5.3 Régions motrices

Outre le A1 et la partie postérieure du STS, considérés aujourd'hui comme les régions cérébrales majeures du traitement bimodal de la parole, de nombreuses études utilisant

l'imagerie cérébrale fonctionnelle ont également montré que la perception bimodale de la parole s'accompagne de l'activation de certaines régions motrices du cerveau, connues pour leur implication dans la production motrice de la parole[14]. Ces régions sont le cortex prémoteur avec l'aire de Broca et, dans une moindre mesure, le cervelet (Fridrikson et *al.*, 2008 ; Hall, Fussell, & Summerfield, 2005 ; Skipper, Nusbaum, & Small, 2005 ; Skipper, Wassenhove, Nasbaum, & Small, 2007 ; voir aussi la Figure 8). Le fait que les régions motrices du langage soient sollicitées lors de la perception audio-visuelle d'un message parlé n'implique cependant pas qu'elles participent également au traitement bimodal de ce dernier. Pour explorer une telle éventualité, Skipper et *al.* (2007) se sont intéressés à la forme des patterns d'activation dans les régions motrices en les associant avec les syllabes, audio-visuellement congruentes ou incongruentes (stimuli de type McGurk-MacDonald). Le stimulus de type McGurk-MacDonald consistait en /pa/ dans la modalité auditive et /ka/ dans la modalité visuelle, généralement perçue comme /ta/. Les auteurs ont comparé le pattern d'activation dans les régions motrices frontales entre, d'une part, la condition comportant le stimulus de type McGurk-MacDonald et, d'autre part, trois autres stimuli audio-visuellement congruents. Ces stimuli correspondaient (i) au stimulus présenté dans la modalité auditive dans le stimulus de type McGurk-MacDonald (en l'occurrence /pa/), (ii) au stimulus présenté dans la modalité visuelle dans le stimulus de type McGurk-MacDonald (en l'occurrence /ka/), (iii) au stimulus perçu par fusion à partir de l'input audio-visuellement incongruent de type McGurk-MacDonald (en l'occurrence /ta/). Skipper et *al.* (2007) ont observé que les patterns d'activation cérébrale présentaient des similitudes plus importantes entre la condition McGurk-MacDonald et la condition comportant le stimulus audio-visuellement cogruent /ta/, qu'entre la condition McGurk-MacDonald et les stimuli audio-visuellement congruents correspondants aux composantes sensorielles du stimulus McGurk-MacDonald, /pa/ et /ka/ respectivement.

Les résultats de Skipper et *al.* (2007) ont été interprétés comme indiquant l'implication des régions motrices du langage dans le traitement bimodal de la parole. Toutefois, une étude récente ayant utilisé l'approche comportementale ainsi que les données de l'IRMf afin d'explorer la même problématique, est arrivée à une conclusion différente (Matchin, Groulx, & Hickok, 2014). En effet, Matchin et *al.* (2014) ont tenté de répondre à la question du rôle des

[14] Dans ce contexte, les réseaux de neurones miroirs ont le statut de corrélat neuronal qui permet d'établir le lien entre l'action et la perception (Di Pellegrino, Fadiga, Fogassi, Gallese, & Rizzolatti, 1992 ; Gallese, Fadiga, Fogassi, & Rizzolatti, 1996 ; Rizzolatti, Fadiga, Gallese, & Fogassi, 1996). En effet, la particularité de ces neurones est qu'ils s'activent aussi bien lors de l'observation d'une action que lors de la production de cette même action.

régions motrices dans la perception bimodale de la parole en testant deux hypothèses en appliquant le raisonnement suivant : (i) Si les régions motrices sont impliquées dans le traitemet bimodal de la parole, alors le fait d'interférer avec leur activité lors d'une tâche de la perception bimodale telle que McGurk-MacDonald, en imposant une tâche articulatoire simultanée, devrait diminuer la performance individuelle dans la tâche principale de perception. (ii) L'implication des régions motrices dans le traitement bimodal de la parole devrait être reflétée par le fait que ces régions soient porteuses des effets supra additif lors de la perception audio-visuelle de la parole comparativement aux conditions de perception unimodale. Partant du raisonnement des auteurs, les résultats de Matchin et *al.* (2014) rejettent l'implication des régions motrices dans le traitement bimodal de la parole. En effet, la tâche articulatoire n'a pas eu d'effet sur les performances individuelles dans la tâche de McGurk-MacDonald (la fréquence de percepts par fusion des stimuli McGurk-MacDonald n'a pas été affectée). Par ailleurs, l'activité des régions motrices du langage ne présentait pas le profil type de traitement bimodal. Aussi, les auteurs en concluent que si les régions motrices peuvent bel et bien être considérées comme étant impliquées dans la perception bimodale de la parole, elles n'assurent pas l'aspect bimodal du traitement de celle-ci.

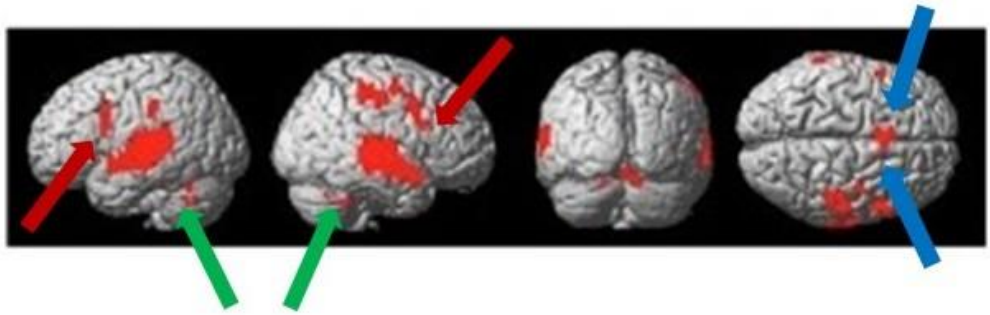

Figure 8. Représentation de l'activation des régions corticales lors de la perception bimodale de la parole (Callan, Callan, & Jones, 2014).

L'image représente l'activité significative (le seuil de *pFDR*[15] étant fixé à p<0,05) dans différentes régions cérébrales moyennée sur quatre conditions expérimentales de perception auditive de la parole : (i) parole en langue maternelle sans accent, (ii) parle en langue maternelle

[15] Le *positive false discovery rate* (*pFDR*) est une procédure statistique permettant de contrôler l'accumulation de la probabilité de commettre des erreurs de type 1 (rejet incorrect de l'hypothèse nulle) dans des conditions de comparaisons multiples (pour plus de détails, voir Benjamin & Hochberg, 1995).

avec accent, (iii) parole en langue étrangère sans accent, (iv) parole en langue étrangère avec accent. L'analyse par conjonction des patterns d'activité pour les quatre conditions révèle une activité bilatérale dans différentes régions motrices, telles que le cortex prémoteur, notamment l'aire de Broca (indiquée par les flèches rouges), l'aire prémotrice supplémentaire (flèches bleues), le cervelet (flèches vertes), ainsi que dans le sillon/gyrus temporal supérieur et le lobe pariétal inférieur. (L'image adaptée (ajout de flèches) de Callan et *al.* (2014).)

En somme, la sollicitation des régions motrices du langage lors de la perception de la parole est une donnée courante dans les études sur ce domaine. Outre les conditions de perception bimodale, leur implication semble essentiellement importante dans des conditions de perception auditive seule où le message est difficile à comprendre, car dégradé par du bruit ou encore par un accent étranger (Adank, Rueschemeyer, & Bekkering, 2013 ; Callan, Callan, Gamez, Sato, & Kawato, 2010 ; Callan et *al.*, 2014 ; Moulin-Frier & Arbib, 2013 Schwartz, Basirat, Menard, & Sato, 2012). De tels résultats suggèrent que la perception audio-visuelle de la parole pourrait également reposer sur des représentations articulatoires des gestes ayant produit les sons à traiter. A terme, ces représentations pourraient avoir le même rôle de contrainte dans l'identification des unités produites que celui en lien avec les inputs acoustique et auditif (Poeppel, Idsardi, & van Wassenhove, 2008 ; Rauschecker, 2011 ; Schwartz et *al.*, 2012 ; Wilson & Iacobini, 2006). Ainsi, même si l'idée proposée par les auteurs de la théorie de perception amodale ou encore de la théorie motrice de la perception de la parole (voir la section 1.4.2) selon laquelle la parole n'est perçue que par le biais des représentations articulatoires ne semble pas valide (pour une revue, voir Galantucci et *al.*, 2006 ; Hickok, 2010), les représentations motrices devraient probablement trouver leur place dans les modèles dominants qui opèrent exclusivement avec les représentations visuelles, phonologiques et audio-visuelles.

Résumé du chapitre

Nos expériences du monde sont très souvent multimodales – notre système cognitif reçoit des informations de différentes modalités sensorielles en lien avec un même objet et/ou évènement. La tâche du système cognitif est de traiter ces informations et en former des représentations unifiées et stables. Il en est de même pour la perception de la parole dans les conditions où notre interlocuteur peut être à la fois vu et entendu.

La parole représente un élément clef pour la régulation des relations humaines ce qui lui accorde une place importante dans l'évolution de l'homme et de son cerveau. En tant qu'êtres humains, nous avons une capacité remarquable à percevoir la parole, même quand celle-ci est émise dans des conditions bruyantes ou encore présente des caractéristiques phonologiques plus ou moins éloignées de celles qui spécifient notre langue maternelle. Ce chapitre a été consacré à la nature bimodale de la perception de la parole qui repose sur l'intégration des inputs auditif (relatif au signal acoustique) et visuel (relatif aux mouvements articulatoires) et rend la perception de la parole particulièrement efficace. En effet, les mouvements articulatoires étant à l'origine des sons produits, ils sont porteurs d'indices aussi bien sur les changements dans les aspects temporel et énergétique des évènements acoustiques (ils les annoncent ou encore leur sont cooccurrents) que sur leurs caractéristiques phonémiques (différents phonèmes sont produits par différents mouvements et/ou positions d'organes articulateurs dont certains sont visibles). Le système cognitif humain utilise respectivement ces indices visuels pour réguler les activités attentionnelles sous-jacentes au traitement de la parole et en tant que système de contraintes qui, combinées avec des contraintes phonologiques et lexicales du système linguistique qu'il connaît, facilitent l'identification du message parlé.

Outre la facilitation de la perception de la parole qui est observée dans les conditions où un input bimodal audio-visuel est disponible, la recherche s'est longuement intéressée à la façon dont les inputs auditif et visuel sont traités et potentiellement intégrés dans une représentation commune. D'une part, les approches neurophysiologique et celle de potentiels évoqués suggèrent que l'input visuel pourrait influencer le traitement de l'input auditif très tôt dans le processus perceptif, notamment déjà lors de son traitement au niveau du cortex auditif primaire. D'autre part, les études utilisant l'imagerie cérébrale fonctionnelle ont unanimement identifié le sillon temporal supérieur comme le site majeur de l'intégration audio-visuelle présentant des connexions avec les deux régions de traitement perceptif primaire des deux inputs sensoriels, le cortex auditif primaire et le cortex visuel primaire. Aussi, le sillon temporal supérieur pourrait correspondre à une seconde étape dans l'intégration audio-visuelle, se produisant plus tardivement dans le traitement bimodal de la parole. Finalement, la perception uni ou bimodale de la parole semble impliquer également certaines régions motrices du langage. Ceci pourrait suggérer que, dans une certaine mesure, les représentations dans lesquelles est encodée la parole par notre système cognitif sont également de nature motrice. Néanmoins, le rôle de ces régions en tant que sites d'intégration audio-visuelle est encore controversé.

Pour conclure, il convient de signaler que la perception bimodale de la parole est un champ de recherche important et qui bénéficie de beaucoup d'attention. Toutefois, il reste un champ de recherche relativement jeune et complexe où le traitement cognitif est potentiellement influencé par de nombreux facteurs relativement interdépendants qui impliquent aussi bien les caractéristiques visuelles et acoustiques des stimuli à traiter que celles de la personne émettrice et réceptrice du message[16]. Par ailleurs, le cerveau humain étant sujet aux processus de maturation, de plasticité neuronale et de mort cellulaire, tout traitement cognitif est susceptible d'évoluer au cours du développement personnel. Le reste du document sera consacré à l'émergence et au développement de la perception audio-visuelle de la parole, ainsi qu'aux problématiques concernant les caractéristiques de l'information visuelle et leur impact sur le traitement bimodal de la parole.

2 Développement de la perception audio-visuelle de la parole

2.1 Introduction

Nous avons tous la conscience qu'un nouveau-né est très différent sur le plan sensoriel, moteur et cognitif d'une personne adulte. Son cerveau est marqué par une certaine immaturité. Au cours de son développement, le cerveau gagnera en volume, il s'organisera morphologiquement et fonctionnellement, au niveau local et global, en fonction de l'expérience que l'enfant fera du monde dans lequel il évolue. Ce processus de maturation cérébrale connaît une durée longue et ne s'arrête entièrement qu'à l'âge adulte où le fonctionnement cognitif de l'individu est censé atteindre l'efficacité optimale (pour une revue sur la question, voir Schneider, 2014). Dans la deuxième moitié de la vie, la courbe des performances sensorielles, motrices et cognitives s'inverse ; on parle d'un déclin dans ces domaines qui est caractéristique du processus de vieillissement, particulièrement marqué par la perte de neurones. Toutefois, tout le long de notre vie, notre cerveau reste plastique et génère de nouvelles cellules souches.

[16] Dans ce contexte, il convient de noter que le domaine d'une langue va bien au-delà des aspects phonologiques et lexicaux qui ont été mentionnés. Le traitement de la parole ayant pour objectif primaire l'identification du message, la perception de la parole est également influencée par des facteurs syntaxiques ou encore pragmatiques. Les informations relatives aux aspects suprasegmentaux du langage, telle que la prosodie, la gestuelle et les expressions faciales qui accompagnent une production verbale orale jouent un rôle crucial dans ce contexte. Toutes ces dimensions ne font cependant pas objet de ce texte (pour plus de détails sur ces sujets, voir Dohen, 2009).

Il peut ainsi se réorganiser en vue d'atteindre, en fonction des conditions, une performance optimale (pour une revue sur la question, voir Hof & Mobbs, 2009).

Parallèlement aux changements morphologiques et fonctionnels dans le cerveau, la capacité de perception multimodale évolue aussi. Dans la mesure où la plupart des objets et évènements du monde offrent une expérience multimodale, l'enfant, dans son développement post-natal, est exposé d'emblée à ce type d'expériences. Par ailleurs, la perception multimodale étant globalement plus efficace et ainsi plus adaptative que la perception unimodale, on peut supposer que la capacité d'intégrer les signaux en lien avec un même objet/évènement provenant des différentes modalités sensorielles pour en former un percept uni émerge rapidement. Pour certains auteurs, avançant l'hypothèse de différenciation inter-sensorielle, de telles capacités seraient présentes dès la naissance et s'amélioreraient sous l'influence de l'expérience perceptive (Bower, 1974 ; Gibson, 1969 ; 1984). D'autres auteurs, en revanche, stipulent que les capacités perceptives d'un nouveau-né sont strictement unimodales (Birch & Lefford, 1963 ; Piaget, 1952). En effet, dans cette vision dite d'intégration sensorielle, les connexions entre les différents systèmes sensoriels se formeraient, sous l'effet de l'expérience perceptive, dans les premiers mois ou même les premières années de la vie. En l'état actuel de la recherche, il semblerait que les deux processus, la différentiation sensorielle ainsi que l'intégration inter-sensorielle, sont impliqués dans le développement des capacités de perception multimodale (pour une revue sur la question, voir Bahrick & Lickliter, 2009 ; Lewkowicz, 2002).

Les deux visions concurrentes du développement de la perception multimodale, et donc également de la perception audio-visuelle de la parole, mettent en avant deux questions de recherche centrales : (i) A quel moment du développement individuel, la perception multimodale émerge-t-elle ? (ii) Quel est le rôle de l'expérience dans le développement de la perception multimodale ? D'autres questions communément abordées dans une approche développementale de la perception multimodale concernent (i) les indices sensoriels/aspects de l'expérience sensorielle utilisés par le système cognitif pour mettre en place la perception multimodale, en l'occurrence la perception audio-visuelle de la parole ; (ii) les changements/l'évolution dans la perception audio-visuelle de la parole dans les différentes périodes de la vie ; (iii) le niveau auquel ces changements se produisent en termes de types de mécanismes cognitifs qui y sont sujets.

Ce chapitre est consacré au développement de la perception audio-visuelle de la parole de la naissance à la vieillesse. Il n'abordera que la dimension du développement normal et

monolingue. (Pour plus de détails sur la dimension inter-linguistique ou encore sur l'effet de déficits sensoriels sur le développement de la perception audio-visuelle de la parole, voir Soto-Faraco, Calabresi, Navarra, Werker, Lewkowicz, 2012.)

2.2 Petite enfance

Durant la petite enfance, notre système cognitif est caractérisé par son immaturité et un manque certain d'expérience perceptive. Les capacités cognitives d'un nouveau-né ou d'un enfant en bas âge, y compris ses capacités perceptives, sont ainsi bien différentes de celles d'une personne adulte. Malgré cette immaturité, on observe chez l'enfant, dans les premiers mois de la vie, l'émergence des capacités perceptives de la parole non négligeables. Elles ont été documentées largement pour la perception unimodale de la parole. Par exemple, dès l'âge de 4 mois, les enfants ont la capacité de distinguer entre deux langues à partir de l'information visuelle seule (Weikum, Vouloumanos, Navarra, Soto-Faraco, Sebastián-Gallés, & Werker, 2007 ; voir la Figure 9). Vers la même période (entre l'âge de 4 à 5 mois), on observe également la capacité de l'enfant à distinguer entre les langues des différentes catégories rythmiques (Bosch & Sebastian-Galles, 2001 ; Nazzi & Ramus, 2003), vers 6 mois celle de segmenter le flux continu de la parole en mots à partir des informations contenues dans le spectre sonore du message acoustique (Maye, Werker, & Gerken, 2002).

Notons également que le développement de la perception de la parole ne suit pas une courbe de progression linéaire. Certaines capacités perceptives sont même perdues au cours de la première année de la vie. En effet, avant l'âge de 6 à 9 mois, les enfants apparaissent comme des « auditeurs universels », pouvant percevoir des contrastes phonémiques non typiques de leur langue maternelle. Au fur et à mesure qu'ils acquièrent de l'expérience avec leur langue maternelle, cette capacité est perdue pour les langues non présentes dans leur environnement social au profit de la langue maternelle (Kuhl, Williams, Lacerda, Stevens, & Lindblom, 1992 ; Kuhl, Stevens, Hayashi, Deguchi, Kiritani, & Iverson, 2006 ; Narayan, Werker, & Beddor 2010 ; Werker & Tees, 1984). Ce phénomène, connu sous le nom de rétrécissement perceptif (*perceptual narrwing*), apparaît également dans d'autres domaines de la perception (pour plus de détails, voir Lewkowicz & Ghazanfar, 2009). Il est révélateur de la réorganisation de notre système cognitif en fonction de la nature de l'expérience qu'un enfant peut avoir/vivre au vu de son environnement.

Les expériences linguistiques et les situations de communication face à face sont fréquentes dans la vie d'un enfant. Il est ainsi légitime de supposer l'existence de nombreux changements dans les capacités linguistiques, y compris dans les capacités de la perception audio-visuelle de la parole, durant l'enfance et essentiellement la petite enfance où le manque d'expérience perceptive est le plus prononcé. La nature de l'expérience linguistique est ici un facteur crucial en lien non seulement avec les caractéristiques de l'environnement linguistique de l'enfant, mais également avec les caractéristiques des systèmes sensoriels (d'éventuels déficits sensoriels) ainsi que celles du système nerveux central (possibilité de troubles neurologiques). Tous ces éléments entrent en jeu en impactant le processus de maturation du système cognitif, et ainsi, à terme, l'expertise de la personne adulte que l'enfant deviendra. Comme il a été annoncé plus haut, cette section du chapitre 2 est consacrée essentiellement au développement typique de la perception bimodale de la parole durant la petite enfance ; elle négligera également l'aspect inter-linguistique.

Figure 9. Représentation schématique du paradigme de l'habituation utilisé par Weikum et *al*. (2007).

Durant la première phase, la phase de l'habituation (*Habituation phase*), l'enfant est exposé à des phrases en langue anglaise présentées de façon visuelle. Au début de la phase (à gauche), l'enfant montre un intérêt pour le stimulus, son attention visuelle est orientée vers la vidéo. A force d'y être exposé, l'enfant s'habitue au stimulus en question ce qui est marqué par une baisse de l'intérêt pour la vidéo et des changements dans l'orientation de son attention visuelle (l'enfant regarde la vidéo moins qu'auparavant ; l'image à droite de la partie *Habituation*

phase). Durant la deuxième phase, la phase de test (*Test phase*), deux types de stimulus sont présentés de façon alternée, des phrases en anglais et des phrases en français, toujours dans leur modalité visuelle seule. Le temps de regard que l'enfant accorde à chaque type de stimulus nous informe s'il le considère comme nouveau ou comme celui auquel il est déjà habitué. Dans l'expérience de Weikum et *al*. (2007), un temps de regard plus long accordé à la vidéo en langue française et une baisse d'intérêt pour la vidéo des phrases en langue anglaise est un indicateur de la capacité de l'enfant à distinguer entre les deux langues à la base des indices visuels seuls. Notons également que l'adulte accompagnant l'enfant est équipé de lunettes teintées pour ne pas influencer l'enfant. (L'image de Weikum et *al*. (2007).)

2.2.1 Détection des correspondances inter-sensorielles

Dans les conditions où la perception de la parole peut être bimodale, le message parlé consiste en une chaîne d'indices auditifs et visuels qui sont aussi bien temporellement couplés et redondants que complémentaires (revoir la section 1.3 du chapitre 1 pour plus de détails). Les informations provenant de deux canaux sensoriels différents, le traitement bimodal de la parole nécessite la capacité de lier l'input auditif, relatif à un évènement de production de la parole, à l'input visuel correspondant. Aussi, la détection de ces correspondances inter-sensorielles chez un enfant préverbal est considérée comme un des marqueurs des capacités de traitement bimodal de la parole.

En l'état actuel de la recherche, les éléments empiriques montrent l'apparition d'une préférence pour la congruence et la synchronie audio-visuelle dans un message parlé, marquée par des fixations plus longues de la source audio-visuelle, vers l'âge de 2,5 mois (Burnham & Dodd, 1998 ; Dodd, 1979). Toutefois, la tâche classiquement utilisée pour étudier la capacité à détecter les correspondances inter-sensorielles dans un cadre de perception bimodale de la parole est celle d'association inter-sensorielle (*intersensory matching*). Cette tâche implique une présentation simultanée de deux vidéos comportant chacune une personne articulant un item différent. Le son correspondant à l'un des items est administré par une enceinte à position centrale entre les deux vidéos. La préférence pour une image, marquée par une exploration visuelle prolongée, marquerait le fait que l'enfant associe le son de l'item à l'image en question. Globalement, il convient de souligner que l'utilisation de la tâche d'association inter-sensorielle a bien plus de poids dans l'étude sur la détection des correspondances audio-visuelles dans la parole qu'une mesure simple de préférence, car elle montre la capacité des enfants à percevoir la cohérence dans l'input bimodal.

Les études ayant utilisé la tâche d'association inter-sensorielle ont montré que les enfants étaient capables de détecter les correspondances audio-visuelles dans une situation de perception bimodale de la parole à partir de l'âge de 4 mois environ. A cet âge, les enfants sont capables d'associer un visage articulant une voyelle au signal acoustique correspondant, aussi bien pour les voyelles de leur langue maternelle (Kuhl & Meltzoff, 1982 ; 1984 ; 1988) que pour les langues étrangères (Walton & Bower, 1993). Patterson et Werker (2003) rapportent des manifestations d'une telle capacité de couplage inter-sensoriel dès l'âge de 2 mois. Plus tard, vers l'âge de 6 mois, les enfants acquièrent la capacité à procéder aux associations audio-visuelles pour les unités plus complexes, notamment syllabes (MacKain, Studdert-Kennedy, Spieker, & Stern, 1983 ; Pons, Lewkowicz, Soto-Faraco, & Sebastián-Gallés, 2009). En ce qui concerne les segments plus importants tels que les phrases, les résultats des études sont plus divergents. Alors que certains auteurs ont trouvé que la capacité à percevoir les correspondances audio-visuelles pour un flux continu de parole émergeait également vers 4,5 à 5 mois (e.g., Dodd & Burnham, 1988 ; Kubicek, de Boisferon, Dupierrix, Pascalis, Loevenbruck, Gervain, & Schwarzer, 2014), d'autres la situent bien plus tard dans le développement vers l'âge de 12 mois (Lewkowicz, Minar, Tift, & Brandon, 2015). Quant à son organisation développementale, les résultats des études sont bien plus concordants. En effet, l'ensemble des études ont établi que c'est une capacité d'abord générale, s'appliquant aussi bien à la langue maternelle de l'enfant qu'aux langues qui lui sont étrangères. Plus loin dans le développement (2 à 3 mois plus tard), la capacité à associer correctement l'input visuel à l'input auditif est maintenue uniquement pour la langue maternelle. Une telle évolution montre que cette capacité est sujette au phénomène de rétrécissement perceptif (voir cependant Kubicek et *al.*, 2014 pour des résultats allant à l'encontre du pattern habituel).

La recherche sur l'émergence de la capacité à détecter les relations inter-sensorielles lors de la perception bimodale de la parole s'est également intéressée au type d'indices qui sont utilisés pour lier les inputs auditif et visuel. Étonnamment, les résultats montrent que l'absence de synchronie audio-visuelle ne perturbe pas les enfants qui maintiennent la capacité à associer un signal acoustique au signal visuel correspondant (e.g., Kubicek et *al.*, 2014 ; Pons et *al.*, 2009 ; voir aussi la Figure 10 pour une présentation schématique des conditions expérimentales de l'étude de Pons et *al.*, 2009)[17]. Une telle constatation suggère que les relations inter-

[17] Un tel résultat peut être qualifié d'étonnant car la synchronie audio-visuelle est souvent considérée comme un des indices les plus informatifs pour la mise en place d'un couplage audio-visuel par le système cognitif (King, 2005 ; Spence & Squire, 2003).

sensorielles dans ces études ont été établies à partir d'indices de haut niveau, consistant en correspondances entre les patterns dynamiques du tract vocal durant la production motrice et les patterns dynamiques acoustiques correspondants. Ce type d'indices est bien plus abstrait que les indices relatifs aux correspondances dans les propriétés amodales et redondantes de la parole audio-visuelle qui sont basés sur une synchronie temporelle entre les deux inputs sensoriels et généralement considérés comme des indices de bas niveau (revoir la section 1.3 du chapitre 1 pour plus de détails).

Il semblerait ainsi que les enfants très jeunes soient déjà capables de tenir compte d'un seul type d'indices pour établir une correspondance entre les inputs auditif et visuel d'une production verbale orale. L'étude de Lewkowicz (2010), qui met en évidence la capacité des enfants, dès l'âge de 4 mois, à détecter une correspondance audio-visuelle en présence d'indices de bas niveau (synchronie audio-visuelle) en l'absence d'indices de haut niveau (corrélations entre les patterns dynamiques, visuels et acoustiques), semble apporter une confirmation supplémentaire à cette hypothèse. Néanmoins, l'étude de Kubicek et *al.* (2014) montre également que, lors du déclin de la capacité à détecter les correspondances audio-visuelles dans une langue étrangère (vers l'âge de 6 mois), la seule présence d'indices de haut niveau s'avère rapidement inefficace. Dans ce cas, la détection de la cohérence audio-visuelle est facilitée par l'ajout des indices de bas niveau, notamment par la synchronie entre les inputs auditif et visuel. L'étude de Kubicek et *al.* (2014) n'a cependant pas étudié la question de savoir si, dans tel cas, la seule présence d'indices de bas niveau est suffisante pour détecter les correspondances audio-visuelles. Ceci pourrait apporter un éclairage supplémentaire à la question d'une éventuelle hiérarchie des indices de bas et de haut niveau quant à leur rôle dans le développement des capacités à détecter les correspondances inter-sensorielles dans la perception audio-visuelle de la parole.

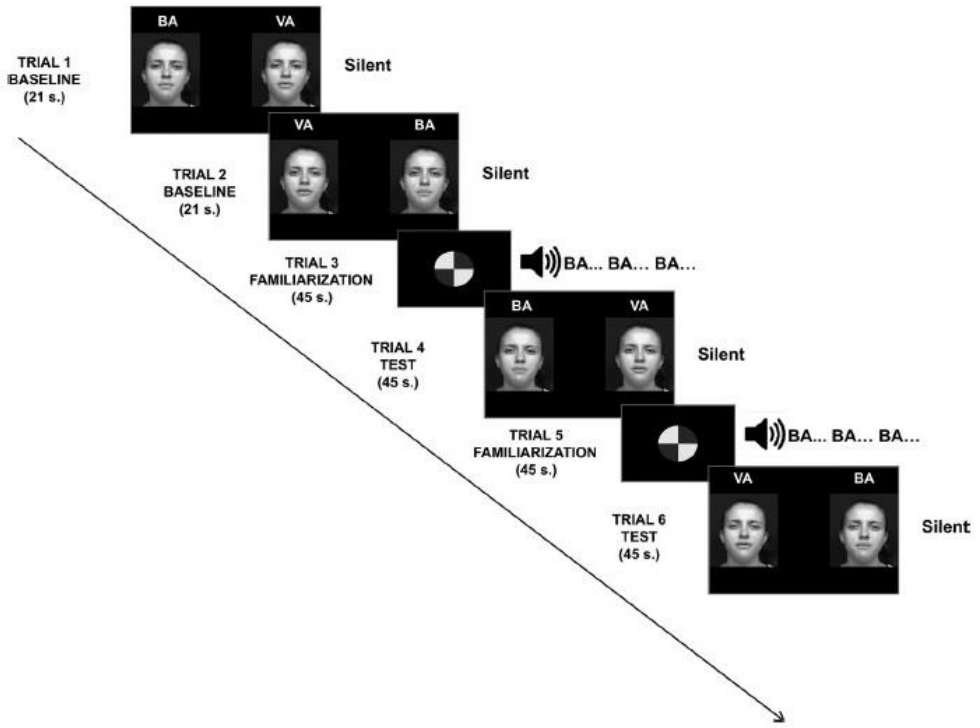

Figure 10. Représentation schématique des conditions expérimentales de l'étude de Pons et *al*. (2009).

L'image représente la succession chronologique des conditions expérimentales comportant des deux types d'items, vidéo et son. La partie vidéo consistait en présentation simultanée d'un modèle articulant la syllabe /ba/ (pour une partie de la vidéo) et la syllabe /va/ (pour l'autre partie de la vidéo) de manière continue. La partie son consistait en une des deux syllabes, /ba/ pour la moitié des participants, /va/ pour l'autre moitié des participants. L'étude a commencé par la mesure du comportement oculaire pour contrôler l'intérêt/la préférence éventuelle de l'enfant pour une des vidéos (*trial baseline*). Le reste de l'expérience consistait en deux types d'essais, les essais de familiarisation qui comportaient l'item son, et les essais tests qui comportaient l'item vidéo. Durant la phase de familiarisation, l'enfant a ainsi été exposé à l'aspect auditif d'une des deux syllabes. Les indices de son comportement oculaire, notamment la différence dans la durée de fixation entre les deux vidéos indique si l'enfant a développé une préférence pour la vidéo correspondant à l'item son présenté précédemment ou pas. Notons que l'étude de Pons et *al*. (2009) a été réalisée avec des enfants espagnols dont l'environnement linguistique ne comporte pas de contraste phonémique /b/-/v/. Conformément au phénomène de rétrécissement perceptif, Pons et *al*. (2009) ont observé que la capacité à associer correctement une syllabe à la vidéo correspondante dans le contexte de l'étude déclinait vers l'âge de 9 mois. (L'image de Pons et *al*. (2010).)

2.2.2 Facilitation audio-visuelle

Un autre phénomène associé au traitement bimodal de la parole est l'effet facilitateur de la perception de la parole en présence de l'input bimodal, comparé à l'input auditif seul. Cet effet peut être associé à une capacité de l'individu à extraire et à convenablement traiter les indices sur différents paramètres du signal acoustique à partir du signal visuel (revoir les sections 1.2 et 1.3 du chapitre 1).

L'émergence et les changements développementaux dans le phénomène de facilitation audio-visuelle de la perception de la parole ont été très peu étudiés pour la période de la petite enfance. En l'état actuel de la recherche, l'étude clef pouvant être classée dans ce domaine est celle de Teinonen, Aslin, Alku, & Csibra (2008). Ces auteurs se sont intéressés à la capacité des enfants de 6 mois à utiliser les indices visuels pour résoudre les ambiguïtés perceptives pouvant apparaître sur le plan auditif. Les participants de l'étude ont été exposés à des stimuli acoustiques variant sur le continuum de la distribution de la fréquence F1 entre les syllabes /ba/ et /da/. Les enfants ont été répartis en deux groupes, différant selon le type de support visuel qui était présenté de manière synchrone au signal acoustique. La moitié des participants (groupe à deux catégories) a été exposée à une présentation des items acoustiques plus proches de la syllabe /ba/ avec la vidéo d'une personne prononçant cette même syllabe, alors que les items acoustiquement plus proches de la syllabe /da/ ont été couplés avec une vidéo de la syllabe /da/. En revanche, l'autre moitié des participants (groupe à une catégorie) a été exposé à un couplage systématique entre l'intégrité des stimuli acoustiques avec une présentation visuelle soit de la syllabe /ba/, soit de la syllabe /da/. Ayant utilisé une variante du paradigme d'inspection préférentielle (*preferential looking paradigm*) (voir Jusczyk & Aslin, 1995 ; Maye et *al*., 2002), Teinonen et *al.* (2008) ont constaté que les enfants du groupe à deux catégories différencient le contraste /ba/ *vs* /da/ dans les stimuli expérimentaux, alors que ce n'était pas le cas des enfants du groupe à une catégorie.

Selon Tienonen et *al.* (2008), ces résultats montrent que les enfants de 6 mois ont la capacité d'extraire les indices sur les paramètres acoustiques à partir du signal visuel et que ceux-ci sont utilisés avec succès dans la perception des contrastes phonémiques pouvant être ambigus dans la seule modalité auditive. Les auteurs suggèrent que les indices visuels relatifs aux mouvements articulatoires pourraient jouer un rôle dans l'apprentissage des catégories phonémiques. Toutefois, une étude ultérieure de Yeung et Werker (2009) a mis en évidence que le simple fait d'associer systématiquement la présentation des stimuli acoustiques avec celle de deux stimuli visuels quelconques mais différents (selon le principe du groupe à deux

catégorie de Tienonen et *al.* (2008)) facilitait également la différenciation phonémique. Le phénomène mis en évidence par Tienonen et *al.* (2008) peut ainsi ne pas être spécifiquement lié à la nature linguistique des stimuli visuels, mais plutôt à la nature systématique du couplage audio-visuel qui pourrait faciliter la perception de la différence dans les items sur le plan acoustique seul. D'autres études sont nécessaires pour élucider le rôle précis de l'information visuelle relative aux mouvements articulatoires dans le développement de la perception de la parole.

2.2.3 Fusion audio-visuelle

Les résultats des études précédemment présentées, qui portent en grande majorité sur la capacité à détecter les correspondances inter-sensorielles dans des conditions de perception bimodale de la parole, mettent en évidence l'existence d'une capacité à extraire les correspondances structurelles de la parole audio-visuelle chez les enfants de 4 mois (voire possiblement plus tôt (Burnham & Dodd, 1998 ; Dodd, 1979)). Cependant, comme le soulignent Sekiyama et Burnham (2004), ces études ne peuvent trancher entre deux types de mécanismes possiblement impliqués dans la perception de la parole. Le premier consisterait à associer simplement les caractéristiques d'un type d'input avec l'autre, le deuxième en revanche, consisterait à intégrer les différents types d'indices dans un percept unifié et cohérent. Les études sur l'émergence de l'effet McGurk-MacDonald durant la petite enfance sont d'une importance cruciale pour trancher entre les deux alternatives.

Chez les enfants préverbaux, l'exploration de l'effet McGurk-MacDonald est classiquement réalisée avec l'utilisation du paradigme de l'habituation. Ce paradigme comporte deux phases, la première étant celle d'habituation de l'enfant à un ou plusieurs items, visuels ou auditifs. Chaque item est présenté jusqu'à ce que l'enfant s'y habitue et perde l'attrait pour l'item en question, ce qui est marqué par un changement dans l'orientation du regard. En effet, l'enfant détourne son attention visuelle de l'item en question et la porte vers d'autres objets[18]. Durant la deuxième phase du paradigme, on observe la réaction de l'enfant face à la présentation d'un item, pouvant être soit nouveau soit connu de l'enfant. Un item que l'enfant considère

[18] Notons que les stimuli auditifs sont également localisés par une orientation du regard vers l'emplacement de la source des stimuli. Cette capacité est présente dès la naissance (e.g. Field, Muir, Pilon, Sinclair, & Dodwell, 1980), disparaît rapidement pour réapparaître vers l'âge de 4 mois (e.g., Field et *al*., 1980 ; Muir, Clifton, & Clarkson, 1989).

comme nouveau attirera son attention visuelle (c'est l'effet de la nouveauté), alors que l'item déjà connu de l'enfant (l'item auquel il a été habitué) n'attirera plus son attention. Par exemple, dans l'expérience de Burnham et Dodd (2004), des enfants de 4 mois ont été répartis en deux groupes. Les enfants du premier groupe ont été habitués à la présentation d'un stimulus de type McGurk-MacDonald consistant en /ba/ auditif et /ga/ visuel, perçu en tant que /da/ ou /ða/ par les adultes (McGurk & MacDonald, 1976). En revanche, les enfants du deuxième groupe ont été habitués à un stimulus audio-visuellement congruent, la syllabe /ba/. Durant la deuxième phase de l'expérience, trois items auditifs uniquement, /ba/, /da/ et /ða/, ont été présentés aux enfants, couplés avec une image statique d'un visage à bouche fermée. Les auteurs ont constaté que les enfants du premier groupe traitaient les stimuli /da/ et /ða/, qui sont les percepts issus de la fusion entre les items /ba/ auditif et /ga/ visuel, comme connus, et le stimulus /ba/ comme nouveau, alors que le pattern a été inversé pour les enfants du deuxième groupe. Un tel résultat montre que les enfants de 4 mois sont déjà susceptibles à l'illusion de McGurk-MacDonald, ce qui implique qu'un traitement par intégration des indices auditifs et visuels dans un percept unifié lors de la perception de la parole est déjà opérationnel à cet âge.

En l'état actuel de recherche, de nombreuses études ont montré l'existence de l'effet McGurk-MacDonald chez les enfants âgés de 4 à 5 mois (e.g., Burnham & Dodd, 1998 ; Desjardin & Werker, 1996, 2004 ; Rosenblum, Schmuckler, & Johnson, 1997). De plus, Desjardin et Werker (1996) et Rosenblum et *al*. (1997) ont établi que le phénomène de domination d'une modalité sensorielle sur l'autre dans certaines combinaisons audio-visuelles dans les stimuli de type McGurk-MacDonald, est également présent chez les enfants préverbaux et suit le même pattern que celui trouvé avec les adultes. En revanche, ces mêmes auteurs ont aussi noté que certaines combinaisons audio-visuelles ne provoquaient pas l'effet McGurk-MacDonald chez les enfants préverbaux, observé chez les adultes. Il semblerait ainsi que, si les mécanismes d'intégration audio-visuelle commencent à se mettre en place très tôt dans le développement post-natal, ils ne sont pas encore pleinement fonctionnels. Leur développement dépendrait vraisemblablement de l'expérience (pour une discussion plus détaillée, voir Desjardin & Werker, 1996).

2.2.4 Corrélats neuronaux

Si le cerveau d'un nouveau-né est marqué d'une relativement grande immaturité, le développement post-natal du cerveau, notamment des régions corticales et de leurs connexions, se caractérise, lui, par une hétérogénéité certaine (e.g., Dubois, Dehaene-Lambertz, Perrin,

Mangin, Cointepas, Duchesnay, …, Hertz-Pannier 2008; Yakovlev & Lecours, 1967). De ce fait, il est, d'un côté, possible que les structures nerveuses impliquées dans le traitement bimodal de la parole nécessitent un certain temps pour s'organiser sur le plan structurel et fonctionnel. D'un autre côté toutefois, les résultats des études comportementales suggèrent que les corrélats nerveux de la perception audio-visuelle de la parole sont assez bien fonctionnels très tôt dans le développement post-natal, même si leur fonctionnement n'est vraisemblablement pas celui qui peut être observé dans le cerveau adulte. Les études portant sur les corrélats neuronaux de la perception de la parole et leur développement durant la petite enfance ont ainsi pour objectif de comparer les phénomènes observés au niveau d'un cerveau adultes à ceux qui sont présents dans un cerveau d'enfant. Elles sont relativement peu nombreuses et sont focalisées sur les aspects dynamiques du traitement bimodal lors de la perception bimodale de la parole. La méthode utilisée dans ces études est celle des potentiels évoqués (pour plus de détails sur cette méthode, revoir la section 1.5 du chapitre 1).

Tout comme les études comportementales, essentiellement celles utilisant le paradigme de McGurk-MacDonald, les études portant sur le traitement cérébral de la parole audio-visuelle durant la petite enfance ont révélé aussi bien des similitudes que des différences entre les enfants et les adultes. La première similitude concerne la topographie des potentiels évoqués par un input linguistique audio-visuel observés essentiellement au niveau du A1 (Bristow, Dehaene-Lambertz, Matut, Soares, Gliga, Baillet, Mangin, 2009 ; Hyde, Jones, Flom, & Porter, 2011 ; Kushnerenko, Teinonen, Volein, & Csibra, 2008 ; Reynolds, Bahrick, Likliter, & Guy, 2014), mais également au niveau des régions temporales postérieures, vraisemblablement en lien avec une activité du sillon temporal postérieur (Bristow et *al.*, 2009), et au niveau des régions frontales (Reynolds et *al.*, 2014), possiblement en lien avec l'activité des régions motrices du langage et/ou des réseaux de neurones miroirs. Ce pattern suggère que les corrélats neuronaux du traitement bimodal de la parole sont les mêmes chez l'enfant comme chez l'adulte, ce qui n'est pas particulièrement surprenant. D'autres constatations d'un plus grand intérêt ont été inférées à partir des changements dans la réponse cérébrale, notamment dans l'amplitude et la latence d'occurrence des différentes composantes des potentiels évoqués, en fonction des conditions expérimentales et particulièrement en fonction des caractéristiques des stimuli. Plus précisément, les ondes prises en compte dans les études développementales sont : (i) les ondes N1 et P2 qui sont relatives à un traitement auditif du signal, (ii) l'onde Nc en lien avec le traitement attentionnel d'un stimulus et (iii) l'onde PSW qui est associée à un traitement mnésique (pour plus de détails, voir Hyde et *al.*, 2011).

Les quatre études citées rapportent des marqueurs (changements dans l'amplitude et/ou la latence des ondes N1 et P1) suggérant l'existence d'une intégration audio-visuelle précoce dans le traitement bimodal de la parole (revoir la section 1.5 du chapitre 1 pour le type et la signification des changements en question). Un tel mécanisme, qui est également observé chez les adultes, semble être en place vers l'âge de 4 à 5 mois (Hyde et *al.*, 2011 ; Kushnerenko et *al.*, 2008 ; Reynolds et *al.*, 2014), et pourrait émerger même avant l'âge de 2,5 mois (Bristow et *al.*, 2009). Utilisant le paradigme de McGurk-MacDonald, Kushnerenko et *al.* (2008), ont montré que, tout comme les adultes, les enfants de 5 mois avaient la capacité à distinguer entre les stimuli audio-visuellement incongruents pouvant être fusionnés en un percept cohérent et ceux ne pouvant pas l'être. Les différences dans la réponse cérébrale à chaque type de stimuli ont été observées aussi bien au niveau des régions temporales qu'au niveau des régions frontales et ont révélé l'implication des mécanismes de fusion précoce dans la construction perceptive de l'illusion de McGurk-MacDonald. L'étude de Hyde et *al.* (2011) et celle de Reynolds et *al.* (2014) se sont intéressées au traitement de la synchronie audio-visuelle. Cette dernière semble être détectée précocement lors du traitement bimodal. Contrairement à celui des adultes, le système nerveux de l'enfant pourrait présenter un biais vers la synchronie inter-sensorielle dans la perception bimodale de la parole (Hyde et *al.*, 2011)[19]. Finalement, prenant en compte les différences dans la réponse de discordance (la négativité de discordance)[20] chez les enfants de 2,5 mois, suite à l'exposition à des stimuli audio-visuels congruents et incongruents, Bristow et *al*. (2009) ont conclu que les correspondances inter-sensorielles dans la parole pouvaient être détectées dès l'âge en question. De nouveau, les mécanismes impliqués dans cette capacité présentaient les caractéristiques de traitement précoce.

[19] Ce résultat est conforme aux prédictions de l'hypothèse de redondance inter-sensorielle (Bahrick & Lickliter, 2000 ; 2002 ; 2012 ; Bremner, Lewkowicz, & Spence, 2012) qui accorde un rôle centrale aux indices amodaux (dont la synchronie inter-sensorielle) dans le développement perceptif et cognitif en général. L'importance des indices redondants résiderait dans le fait qu'ils attirent l'attention sélective du sujet, ce qui serait particulièrement facilitateur pour la mise en place du couplage inter-sensoriel dans la perception multimodale. La recherche a en effet montré que l'allocation des ressources attentionnelles aux indices redondants est une dimension primordiale du développement cognitif (e.g., Bahrick & Lickliter, 2000 ; 2012 ; Lewkowicz, 2000 ; pour une revue, voir Bahrick & Lickliter, 2014).

[20] La négativité de discordance (*mismatch negativity* (*MMN*)) est une composante des potentiels évoqués qui marque un changement dans le stimulus (pour plus de détails, voir Näätänen, Paavilainen, Titinen, Jiang, & Alho, 1993).

Globalement, les résultats des études utilisant la méthode des potentiels évoqués sont concordants avec ceux des études comportementales. Ils apportent également des éclairages précis sur la dimension dynamique du traitement bimodal de la parole durant la petite enfance. Ils accentuent le caractère vraisemblable de l'existence d'une intégration précoce des deux inputs sensoriels lors de la perception audio-visuelle de la parole, ou au moins le fait que les mécanismes de fusion audio-visuelle précoce sont relativement bien fonctionnels très tôt dans le développement. (Le développement des mécanismes d'intégration tardive durant la petite enfance, probablement sous-tendus par le sillon temporal supérieur, dont l'étude mériterait une approche d'imagerie fonctionnelle, reste un champ obscure de la recherche.) Tout comme les études comportementales, les études des potentiels évoqués mettent en évidence l'importance de l'information visuelle pour le développement de la perception de la parole.

2.3 De l'enfance à l'âge adulte

Les capacités relatives au traitement bimodal de la parole émergent, nous l'avons vu, très tôt dans le développement post-natal. Durant la première année de la vie, les capacités de l'enfant dans ce domaine perceptif présentent des similitudes importantes avec celles observées chez l'adulte. On pourrait ainsi s'attendre à ce que, dans la suite du développement, la perception audio-visuelle de la parole atteigne assez rapidement le niveau adulte. En effet, la période de l'enfance et l'intégration de l'école offrent un cadre riche en échanges verbaux dans des situations de communication face à face. Toutefois, les études comportementales et celles qui s'intéressent aux corrélats neurologique du traitement bimodal de la parole apportent des éléments empiriques qui vont à l'encontre de cette supposition. En effet, les mécanismes de la perception bimodale de la parole connaissent des changements qui se prolongent jusqu'en milieu de l'adolescence. La présente section est consacrée au développement de la perception audio-visuelle de la parole durant l'enfance, la puberté et l'adolescence.

2.3.1 Facilitation audio-visuelle

Lun des domaines de l'étude sur la perception bimodale de la parole qui a fourni des éléments empiriques montrant que le développement de la perception bimodale de la parole dure jusqu'en adolescence est celui qui concerne la capacité de l'individu à restituer l'input auditif, quand ce dernier est dégradé, à l'aide d'un input bimodal. En effet, les résultats des rares études développementales (Barutchu et *al*., 2010 ; Ross, Molholm, Blanco, Gomez-

Ramirez, Saint-Amour, & Foxe, 2011) dans ce domaine montrent que la capacité des enfants à extraire les indices relatifs au signal acoustique à partir de l'input visuel pour résoudre l'ambiguïté perceptive n'est pas optimale et qu'elle connaît des changements possiblement jusqu'à l'adolescence (Ross et *al*., 2011). Malheureusement, un manque important d'études en l'état actuel de recherche ne permet pas d'aboutir à des conclusions d'un degré satisfaisant de validité quant à l'évolution de cette capacité durant l'enfance et, éventuellement, l'adolescence.

L'étude la plus complète dans le domaine en question est celle de Ross et *al*. (2011). En appliquant le paradigme de la dégradation de l'information auditive par le bruit, les auteurs ont exploré d'éventuels changements développementaux dans l'effet du degré de dégradation de l'input auditif sur le gain audio-visuel dans un intervalle d'âge important (5-7 ans, 8-9 ans, 10-11 ans, 12-14 ans, 16-65). Les résultats ont révélé une augmentation constante de la performance d'identifications correctes des items verbaux sur les 5 groupes d'âge. Quant au gain audio-visuel, les résultats sont plus complexes. Conformément à ce qui avait été observé précédemment (Ross et *al*., 2007 ; Ma et *al*., 2009), Ross et *al*. (2011) ont trouvé, chez les adultes, le pattern caractéristique dans la relation entre le gain audio-visuel et l'ampleur du rapport signal-bruit (*signal-to-noise ration* (*SNR*)), avec un gain maximal au SNR-12 (pour plus de détails sur le pattern en question, revoir la section 1.2 du chapitre 1). Les gains audio-visuels des enfants et des préadolescents/adolescents (5-7 ans, 8-9 ans, 10-11 ans, 12-14 ans) ont été globalement moins élevés que ceux des adultes. Cette différence devenait plus importante avec l'augmentation de l'ampleur du SNR. Une caractéristique importante rapportée par Ross et *al*. (2011) est l'absence du pattern caractéristique dans la relation entre le gain audio-visuel et le degré de dégradation de l'information auditive chez les plus jeunes enfants qui présentaient le gain audio-visuel le plus important à un SNR plus élevé (-9dB). (Pour cet aspect, les résultats des enfants plus âgés et des préadolescents/adolescents présentaient un pattern proche de celui des adultes.) Les deux dernières dimensions des résultats suggèrent que, comparativement aux adultes, le traitement bimodal de la parole est, chez les enfants, davantage dépendant de l'input auditif. Finalement, l'augmentation dans le gain audio-visuel global en fonction de l'âge n'était pas uniforme, suggérant l'existence d'une période sensible pour le développement de la perception bimodale de la parole située possiblement entre l'âge de 8 à 9 ans et 10 à 11 ans. Toutefois, l'observation des variations du gain audio-visuel parmi les différents groupes d'âge en fonction du SNR montre une évolution constante et uniforme pour les SNR les plus bas (-9dB, -12dB, -15dB, -18dB), alors que pour les SNR plus élevés (-3dB, -6dB) le gain audio-visuel présente des variations importantes en fonction de l'âge. Un tel pattern de résultats ne

permet pas de tirer des conclusions aussi directes et simples quant au développement de la capacité à extraire des informations acoustiques du signal visuel pour faciliter la perception du message parlé. (Pour une présentation graphique de certains résultats de l'étude, voir la Figure 11). En effet, ce pattern suggère que le développement de cette capacité est possiblement complexe, mais il est également possible qu'il soit lié à certaines caractéristiques de l'échantillon qui n'ont pas été prises en compte/contrôlées dans l'étude. Des constatations similaires à celles de Ross et *al*. (2011) ont été rapportées également par Barutchu et *al*. (2010). Toutefois, les différences méthodologiques entre les deux études empêchent d'aboutir à des conclusions définitives sur les points les plus importants. Plus précisément, le choix des groupes d'âge, bien plus restreint chez Barutchu et *al*. (2010) (10 ans, 11 ans et adultes) ne permet pas de situer la fin du développement de la capacité à extraire des indices acoustiques à partir de l'input auditif pour restituer un message parlé acoustiquement dégradé. Par ailleurs, le choix des SNR (-9dB, -12dB, -22dB et -30dB pour Barutchu et *al*., 2010) ne permet pas d'apporter plus d'éclairage sur le développement de cette capacité à des SNR les plus élevés (-3dB, -6dB), associés à une variabilité inter-groupe importante dans le gain audio-visuel chez Ross et *al*. (2011). Barutchu et *al*. (2010) avancent l'hypothèse selon laquelle les changements développementaux dans le gain audio-visuel en fonction du degré de dégradation de l'information auditive refléteraient les différences dans le traitement attentionnel. Cette hypothèse semble en lien avec d'autres études ayant trouvé une diminution de la distractibilité par le bruit environnant chez les enfants plus âgés (Hetu, Truchon-Gagnon, & Bilodeau, 1990) ou encore une augmentation, avec l'âge, de la capacité à correctement identifié le dernier mot d'une phrase produite dans une situation de *cocktail party* (Elliot, 1979). D'autres études sont néanmoins nécessaires pour apporter plus d'éclairage sur le sujet.

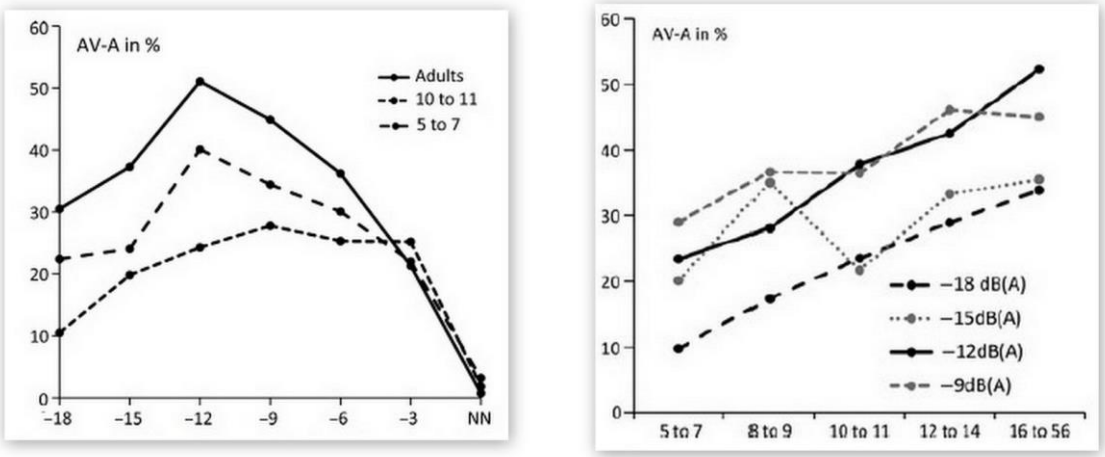

Figure 11. Représentations graphiques des variations du gain audio-visuel (AV-A), exprimé en termes de pourcentage, en fonction de l'âge et du SNR, telles que rapportées par Ross et *al*. (2011).

A gauche, la présentation graphique des variations du gain audio-visuel sur l'ensemble des SNRs telles qu'elles ont été observées dans le groupe d'adultes, d'enfants de 10 à 11 ans et d'enfants de 5 à 7 ans. A droite, les variations du gain audio-visuel sur les SNRs les plus bas et pour l'ensemble des groupes d'âge. (Les images de Ross et *al*. (2011).)

2.3.2 Fusion audio-visuelle

Malgré une quasi non existence d'études s'intéressant au développement de la facilitation de la perception de la parole par un input bimodal durant l'enfance et l'adolescence, les résultats des études utilisant le paradigme de McGurk-MacDonald sont relativement concordants avec les observations de Ross et *al.* (2011). Globalement, ils montrent que l'impact de l'information visuelle sur la perception de la parole augmente avec l'âge jusqu'à l'adolescence.

La première étude développementale s'intéressant à l'évolution du traitement bimodal de la parole durant l'enfance au moyen de l'effet McGurk-MacDonald est l'étude princeps de McGurk et MacDonald (1976). En effet, il s'agit d'une étude transversale, réalisée avec des participants de trois groupes d'âge, 3 à 5 ans, 7 à 8 ans et 18 à 40 ans. Les auteurs ont établi que les enfants des deux groupes d'âge ont été moins sensibles à l'illusion McGurk-MacDonald que les adultes. Ce même pattern de résultats a été largement répliqué depuis (Desjardin, Rogers, & Werker, 1997 ; Dupont, Aubin, & Ménard, 2005 ; Hockley & Polka, 1994 ; Massaro,

Thompson, Barron, & Lauren, 1986 ; Narinesingh, Goltz, Raashid, & Wong, 2015 ; Sekiyama & Burnham, 2008 ; Trembay, Champoux, Voss, Bacon, Lepore, Théoret, 2007 ; Wightman, Kistler, & Brungart, 2006 ; etc.). Toutefois, l'intervalle d'âge pris en compte dans la plupart des études a été relativement restreint, ne permettant pas de situer précisément le moment à partir duquel la sensibilité à l'effet McGurk-MacDonald se stabilise et atteint le niveau adulte. L'étude de Tremblay et *al*. (2007) avait pour vocation de remédier à ce problème en prenant en compte un intervalle large, allant de 5 à 19 ans. Les auteurs ont trouvé que le point de rupture de la trajectoire développementale se situait vers l'âge de 10 ans. En effet, les enfants du groupe 5 à 9 ans différaient significativement dans leur susceptibilité à l'illusion McGurk-MacDonald, alors que ce n'était plus le cas pour le groupe 10 à 14 ans. Toutefois, l'étendue assez importante des intervalles d'âges pris en compte dans l'étude de Tremblay et *al*. (2007) ne permet pas de situer avec précision l'âge auquel les mécanismes de perception audio-visuelle atteignent leur maturité. Notons finalement que Tremblay et *al*. (2007) ont également testé les participants pour la susceptibilité à des illusions audio-visuelles non caractéristiques de la perception de la parole. Aucun effet d'âge n'a été trouvé à ce niveau ce qui suggère un processus de maturation différent entre les mécanismes de traitement bimodal spécifiques de la perception de la parole et ceux impliqués dans la perception bimodale des stimuli non linguistiques.

De manière générale, les résultats des études développementales ayant eu recours au paradigme McGurk-MacDonald suggèrent que l'intégration audio-visuelle n'est pas entièrement opérationnelle chez les jeunes enfants. Par ailleurs, les caractéristiques des réponses des participants dans lesquelles les stimuli de type McGurk-MacDonald n'ont pas provoqué la fusion illusoire révèlent un moindre effet de l'information visuelle sur la perception de la parole chez les enfants que chez les adultes (e.g., Hockley & Polka, 1994 ; Massaro, 1984 ; Wightman et *al*., 2006 ; etc.). Le développement des mécanismes de la perception audio-visuelle semblent ainsi sujet au processus de maturation qui dure au moins jusqu'à l'âge de 6 ans et possiblement même jusqu'en adolescence. (Par exemple, Hockley et Polka (1994) ont trouvé que l'influence de l'input visuel sur la perception de la parole augmentait graduellement entre l'âge de 6 et 11 ans. Sekiyama et Burnham (2008) ont rapporté un développement rapide du processus de l'intégration audio-visuelle entre l'âge de 6 et 8 ans.)[21]

[21] Au sujet de la maturation des mécanismes de l'intégration audio-visuelle, il convient de noter que des différences importantes dans l'étendue de la fenêtre d'intégration audio-visuelle (revoir la section 1.4.1 du chapitre 1) ont été trouvées entre les enfants et les adultes pour des stimuli non verbaux. La trajectoire développementale présenterait un pattern de rétrécissement de la fenêtre en question qui se terminerait probablement en adolescence (Hillock-

Quant à la nature des changements dans les mécanismes de la perception bimodale de la parole, il peut s'agir soit d'une optimisation de l'extraction des indices acoustiques à partir de l'input visuel qui se ferait sous l'effet de l'expérience perceptive (hypothèse émise par Massaro et *al.*, 1986), soit d'une optimisation de l'intégration des indices visuels et auditifs. Des changements dans les deux domaines sont également possibles. Cette question reste encore ouverte, car si Massaro et *al.* (1986) ont rapporté une capacité moindre des enfants à identifier un message parlé présenté dans sa modalité visuelle seule, d'autres auteurs ont trouvé une capacité des enfants à lire sur les lèvres comparable à celle des adultes (Dupont et *al.*, 2005 ; Tremblay et *al.*, 2007). Par ailleurs, dans la mesure où les enfants, dans les premiers mois de vie post-natale, développent une susceptibilité à l'illusion McGurk-MacDonald et qu'ils sont même capables de distinguer entre deux langues différentes en se basant sur l'input visuel seul (Weikum et *al.*, 2007), tout le développement dans le traitement bimodal de la parole ne s'explique probablement pas exclusivement par l'acquisition de l'expérience visuelle dans ce domaine. Toutefois, la nature de l'expérience perceptive est certainement un facteur important dans le développement du traitement bimodal de la parole même en dehors de la petite enfance où elle est le plus souvent étudiée en lien avec le phénomène de rétrécissement perceptif. En effet, la recherche montre que différents aspects de l'expérience perceptive impactent les changements développementaux dans la susceptibilité de l'individu à développer l'illusion McGurk-MacDonald également durant l'enfance. Un de ces aspects est relatif aux caractéristiques de l'environnement linguistique de l'enfant (Sekiyama & Burnham, 2008 ; voir aussi la Figure 12)[22], d'autres sont en lien avec ses capacités sensorielles (e.g., Narinesingh et

Dunn & Wallace, 2012). Si des résultats similaires étaient trouvés pour le matériel verbal, ils suggéraient que les mécanismes de l'intégration audio-visuelle connaissent une maturation encore plus longue que ce qui peut être supposé à partir des résultats des études ayant eu recours au paradigme de McGurk-MacDonald. Le processus de maturation pourrait aussi se révéler être hétéroclite.

[22] L'étude de Sekiyama et Burnham (2008) a comparé l'évolution développementale de la susceptibilité à l'effet McGurk-MacDonald entre les enfants de l'environnement linguistique japonais et les enfants de l'environnement anglophone. Les résultats montrent une susceptibilité comparable à l'illusion en question à l'âge de 6 ans, alors que chez les enfants plus âgés, une augmentation de l'impact de l'information visuelle est notée chez les enfants anglophones, mais pas chez les Japonais (voir la Figure 12). Notons que l'explication la plus plausible de cette différence inter-linguistique pointe vers les différences dans la taille du répertoire phonémique entre les deux langues. En effet, le répertoire phonémique du japonais étant plus restreint, les Japonais auraient un moindre besoin que les personnes anglophones à s'appuyer sur les indices visuels dans la perception de la parole. Toutefois, une explication concurrente à ce phénomène met un accent plus important sur les différences culturelles. Plus précisément, le contact visuel entre deux interlocuteurs étant beaucoup moins fréquent dans l'espace culturel

al., 2015 ; Rouger, Fraysse, Deguine, & Barone, 2008) ou encore ses particularités cognitives/neurologiques telles qu'elles sont observées dans l'autisme (e.g., Bebko, Schroeder, & Weiss, 2014) et dans la dyslexie (e.g., Cavé, Stroumza, & Bastein-Tonazio, 2007).

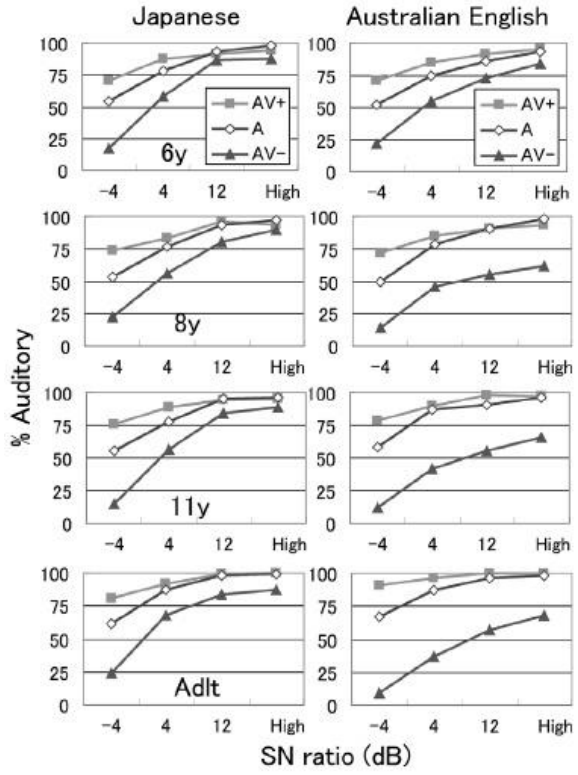

Figure 12. Présentation graphique des variations du taux des réponses correspondant à la modalité visuelle pour les syllabes de l'étude de Sekiyama et Burnham (2008) en fonction des facteurs expérimentaux.

Le taux des réponses correspondant à la modalité visuelle des syllabes est exprimé en pourcentage. Il varie en fonction du mode de présentation des syllabes, pouvant être audio-visuel congruent (*AV+*), audio-visuel incongruent (*AV-*) et auditif (*A*), du SNR, de l'âge des participants (6 ans, 8 ans, 11 ans, adultes) et de leur langue maternelle. On note des trajectoires développementales différentes entre les participants Japonais (graphiques de gauche) et

japonais, les mécanismes d'extraction d'indices acoustiques à partir de l'input visuel seraient moins performants dans cette population à cause d'un manque d'expérience perceptive (Seiyama & Tohkura, 1993). Néanmoins, cette explication semble peu plausible car un moindre poids de l'information visuelle dans l'effet McGurk-MacDonald a été observé dans d'autres milieux culturels ne possédant pas les mêmes règles d'interactions sociales que l'espace culturel japonais (pour la langue hébreu, voir Aloufy, Lapidot, & Myslobodsky, 1996).

Australiens (graphiques de droite). Une baisse importante dans le taux en question apparaît à partir de 6 ans chez les participants Australiens, marquant ainsi une plus grande influence de l'information visuelle sur le traitement bimodal de la parole. Ce changement n'est pas observé chez les Japonais.

2.3.3 Corrélats neuronaux

Les résultats des études comportementales s'intéressant au développement du traitement bimodal de la parole de l'enfance à l'âge adulte suggèrent que ce dernier connaît une maturation longue, se prolongeant loin dans la période de l'enfance et possiblement jusqu'en adolescence. Ceci est en lien avec le développement d'autres types de capacités relevant de la fonction linguistique. En effet, les capacités de production, ainsi que de réception du langage connaissent un développement important durant l'enfance (voir Saffran, Werker, & Werner, 2006). Ce dernier est dépendant de la maturation des régions corticales périsylviennes qui est plus longue que la maturation des régions sous-tendant les capacités sensorielles et perceptives plus basiques (Shaw, Kabani, Lerch, Eckstrand, Lenroot, Gogtay, …, Wise 2008). Les études s'intéressant spécifiquement au développement neurologique et celui des mécanismes neurophysiologiques impliqués dans la perception bimodale de la parole ont abordé la question à l'aide de l'imagerie cérébrale fonctionnelle et de la méthode des potentiels évoqués. Notons qu'à l'état actuel de la recherche, de telles études sont peu nombreuses.

La première étude qui s'est intéressée aux trajectoires développementales des mécanismes neurophysiologiques de la perception audio-visuelle de la parole de l'enfance à l'âge adulte est celle de Knowland, Mercure, Karmiloff-Smith, Dick, et Thomas (2014). Ayant opté pour des groupes d'âge de 6-7 ans, 8-9 ans, 10-11 ans et 20-34 ans, les auteurs ont analysé les changements dans le délai et l'amplitude des ondes N1 et P2 dans la condition de présentation audio-visuelle des items (mots) comparativement à la condition de présentation auditive. Les auteurs ont observé des différences développementales dans la réduction de l'amplitude des ondes en question significative pour les stimuli audio-visuels. Dans les groupes d'enfants, la réduction de l'amplitude a été observée pour la composante P2 uniquement, alors que le groupe adulte présentait le pattern classique, c'est-à-dire une réduction de l'amplitude des deux ondes. Plus précisément, Knowland et *al*. (2014) ont noté que la réduction de l'amplitude de la N1 dans la condition de présentation audio-visuelle apparaissait vers l'âge de 10 ans, alors que la réduction de l'amplitude de la P2 émergeait vers l'âge de 7 ans. Quant à la réduction du délai des ondes N1 et P2, également significative pour un matériel verbal audio-

visuel, les auteurs n'ont trouvé aucune différence entre les différents groupes d'âge. La réduction de l'amplitude de la N1 étant relative à la valeur informative de l'input visuel quant au phonème qui est produit, les résultats de Knowland et *al*. (2014) pourraient suggérer que les mécanismes d'extraction d'indices phonémiques du signal visuel atteignent leur maturité vers l'âge de 10 ans. En revanche, il semblerait que les mécanismes d'intégration audio-visuelle, classiquement associés à une réduction de l'amplitude et de la latence de la P2 (revoir la section 1.5.1 du chapitre 1 pour plus de détails sur la signification des changements dans les paramètres des ondes N1 et P2 par un input linguistique audio-visuel), soient au point déjà avant cet âge. Toutefois, ayant conduit une étude similaire, Kaganovich et Schumaker (2014) n'ont pas reproduit les résultats de Knowland et *al*. (2014). En effet, Kaganovich et Schumaker (2014) ont trouvé que la morphologie et l'amplitude des composantes différaient globalement entre les enfants (7-8 ans, 10-11 ans) et les adultes (pour une présentation schématiques des potentiels évoqués enregistrés par Kaganovich & Schumaker, 2014, voir la Figure 13)[23]. En revanche, les patterns de réduction de la latence et de l'amplitude étaient stables chez les différents groupes d'âge. Les auteurs ont interprété ces résultats comme indiquant que les mécanismes impliqués dans la perception bimodale de la parole étaient opérationnels avant l'âge de 7 ans. Toutefois, la recherche future devra répondre à la question de savoir si, en dehors de la réduction des paramètres des composantes de potentiels évoqués, la morphologie globale des ondes peut également avoir une valeur informative quant aux mécanismes neurophysiologiques étudiés.

[23] Notons que Kaganovich et Schumaker (2014) ont également observé une différence dans la topographie de la réduction de l'amplitude entre les ondes N1 et P2. En effet, la réduction de l'amplitude de la N1 par un matériel audio-visuel a été le plus importante sur la partie droite du scalpe, alors que, pour la P2, la réduction était plus importante au niveau du scalpe gauche et médian. Ces résultats sont conformes aux résultats des études précédentes pointant vers le fait que les changements dans chaque composante sont relatifs à des mécanismes de perception audio-visuelle différents.

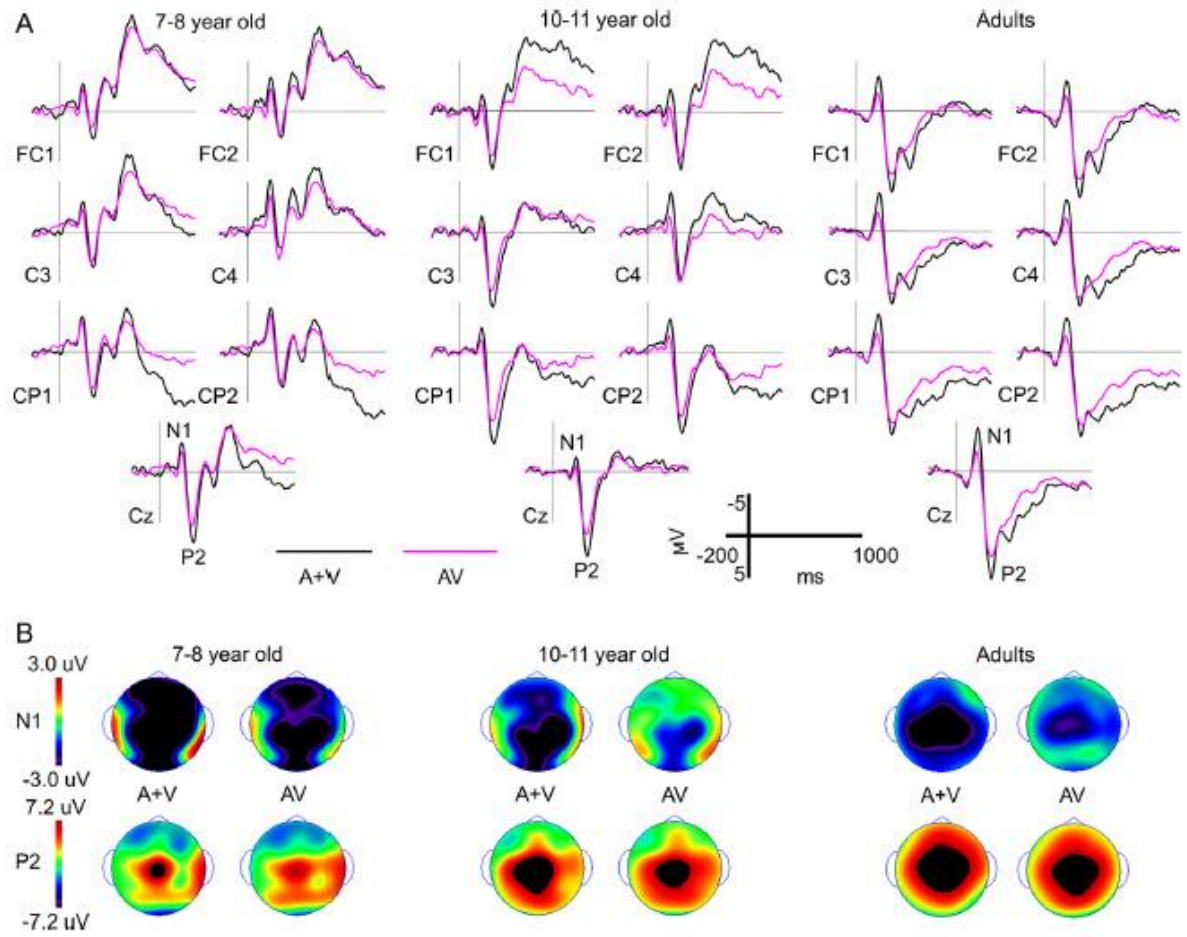

Figure 13. Représentation graphique (latence et amplitude) et schématique (topographie) des composantes N1 et P2 telles que mesurées par Kaganovich et *al*. (2014).

Les potentiels évoqués enregistrés suite à l'exposition à des syllabes pour les conditions de présentation audio-visuelle (*AV*), visuelle (*V*) et auditive (*A*). En A, la moyenne des ondes N1 et P2 pour la condition AV (en noir) et la moyenne de la somme des conditions A et V, A+V, (en rose), pour les trois groupes d'âge (7-8 ans, 10-11 ans, adultes). Six sites médio-latéraux et un site central (Cz) sont présentés pour chaque groupe. Les valeurs négatives sont présentées sur la partie supérieure de l'axe des ordonnées. En B, les topographies du voltage sur le scalpe au moment du pic de la N1 (la ligne supérieure) et de laP2 (la ligne inférieure), présentées pour les conditions AV et A+V ainsi que pour chaque groupe d'âge. On note une réduction dans l'ampleur des deux composantes dans la condition AV comparativement à la condition A+V. (L'image de Kaganovich et *al*. (2014).)

Outre le cortex auditif primaire, la structure corticale censée jouer un rôle majeur dans l'intégration audio-visuelle chez les adultes est le sillon temporal supérieur (revoir la section

1.5.2 du chapitre 1). Aussi, deux études se sont intéressées aux trajectoires développementales neurologiques fonctionnelles au niveau de cette région durant l'enfance. En effet, Dick, Solodkin et Small (2010) ont eu recours à l'IRM fonctionnelle et à la modélisation en équations structurales pour étudier les changements développementaux dans la connectivité du réseau fronto-temporo-pariétal impliqué dans la perception audio-visuelle de la parole. L'étude a été réalisée avec deux groupes d'enfants (8 et 11 ans) et un groupe d'adultes. Les participants ont été exposés à des phrases présentées soit de façon auditive, soit de façon audio-visuelle. Les résultats de Dick et *al*. (2010) ont montré que les régions cérébrales activées dans chaque condition de présentation des stimuli ont été les mêmes chez les adultes et chez les enfants. Par ailleurs, les auteurs ont trouvé une réponse BOLD comparable dans son intensité[24] dans les différentes régions du réseau fronto-temporo-pariétal pour les trois groupes et les deux conditions de présentation des stimuli (pour une représentation schématique des résultats, voir la Figure 14). En revanche, les résultats de la modélisation en équations structurales ont révélé une différence dans la connectivité fonctionnelle dans les régions du réseau étudié qui est apparue pour les trois comparaisons inter-groupes uniquement dans la condition audio-visuelle. En effet, l'influence du gyrus frontal inférieur/cortex ventral prémoteur sur le gyrus supramarginal a été plus importante chez les adultes comparativement aux enfants. Le rôle du réseau en question consisterait à établir des liens entre l'information motrice (articulation) et les informations sensorielles, auditive et somesthésiques (Callan, Jones, Callan, & Akahane-Yamada, 2004 ; Skipper, Nusbaum, & Small, 2006 ; van Wassenhove et *al.*, 2005 ; Wilson & Iacobini, 2006 ; etc.). Selon les auteurs, le fait que les différences inter-groupes dans la connectivité du réseau frontal inférieur/cortex ventral prémoteur – gyrus supramarginal soient observées uniquement dans la condition audio-visuelle pourrait suggérer que le développement de ce réseau est en lien avec les changements dans les mécanismes qui relient l'information visuelle de la parole à l'information auditive. Ceci se ferait avec l'acquisition de l'expérience aussi bien dans la production et que dans la perception de la parole. Notons finalement que Dick et *al*. (2010) ont également trouvé une connectivité modérément positive entre d'un côté, le gyrus frontal postérieur inférieur/cortex prémoteur ventral et, d'autre côté, le gyrus supramarginal et le sillon temporal supérieur chez les adultes, alors que cette connectivité était modérément négative chez les enfants. Toutefois, dans ce réseau, la différence dans la connectivité fonctionnelle entre les adultes et les enfants n'a pas atteint le seuil de signification.

[24] L'intensité de la réponse BOLD a été mesurée par rapport à la valeur de base relevée au moment du repos du sujet.

Aussi, le sillon temporal supérieur ne semble pas impliqué dans le processus de maturation durant l'enfance tardive de manière directe, mais plutôt indirecte. Toutefois, d'autres études sont nécessaires pour répliquer les résultats de Dick et *al*. (2010) et explorer plus précisément le rôle du sillon temporal supérieur dans le développement des réseaux fonctionnels impliqués dans la perception bimodale de la parole.

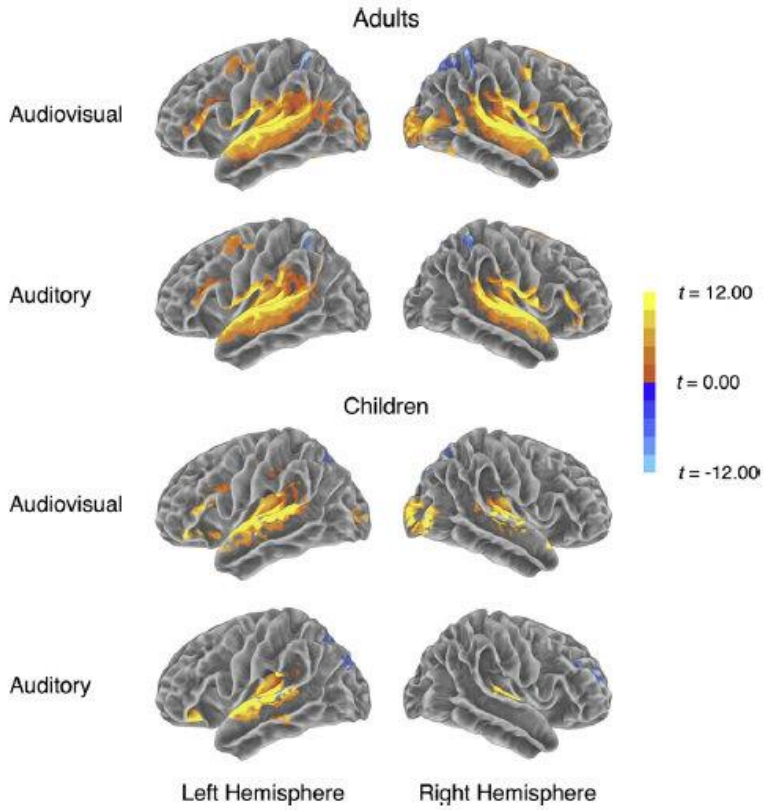

Figure 14. Représentation schématique des scans de l'IRMf tels que présentés par Dick et *al*. (2010).

L'image schématique représente l'intensité de la réponse BOLD, comparativement à la réponse de base enregistrée lors du repos du sujet, chez les enfants (*Children*) et les adultes (*Adults*) pour les deux conditions de présentation des stimuli, auditive (*Auditory*) et audio-visuelle (*Auditoryvisual*). (L'image de Dick et *al*. (2010).)

Pour finir, notons qu'une autre étude, réalisée par Nath, Fava et Beauchamp (2011), s'est également intéressée au processus de maturation du sillon temporal supérieur lors de l'enfance. Plus précisément, l'objectif primaire de Nath et *al*. (2011) était de rendre compte de

la variabilité interindividuelle, qui est très importante, dans la sensibilité des enfants à l'effet McGurk-MacDonald, par les différences dans la réponse du sillon temporal supérieur face aux stimuli de type McGurk-MacDonad. Les auteurs ont en effet trouvé une corrélation positive entre la tendance d'un enfant à présenter l'effet McGurk-MacDonald et l'intensité de la réponse dans la région étudiée. Les participants de l'étude étant âgés de 6 à 12 ans, une corrélation entre l'âge et la susceptibilité à l'effet McGurk-MacDonald a également été réalisée. Toutefois, elle s'est révélée non significative. Les résultats de l'étude de Nath et *al*. (2011) pourraient suggérer que les mécanismes de l'intégration audio-visuelle sont déjà entièrement au point à l'âge de 6 ans. Toutefois, une taille très réduite de l'échantillon (17 participants), ainsi qu'une absence d'un groupe d'adultes ne permettent pas de tirer de telles conclusions.

Globalement, les résultats des études s'étant intéressées aux trajectoires développementales des corrélats neurologiques et neurophysiologiques des mécanismes de la perception audio-visuelle de la parole ne sont pas très concluant. D'une part, les études ayant abordé l'aspect neurophysiologique du traitement bimodal de la parole (Kaganovich & Schumaker, 2014 ; Knowland et *al*., 2014) ont produit des résultats divergents quant à la durée du processus de maturation, d'autre part, les études ayant exploré les aspects neurologiques fonctionnels des mécanismes en question présentent des faiblesses méthodologiques (Nath et *al*., 2011) ou ont produit des résultats qui n'ont pas encore été répliqués (Dick et *al*., 2010). D'autres études sont ainsi nécessaires pour apporter un éclairage sur les aspects qui ont été étudiés jusqu'à présent.

2.4 Vieillissement

Le vieillissement normal est un processus associé à un déclin des fonctions sensorielles et cognitives. En outre, les personnes âgées perdent de l'acuité visuelle (Spear, 1993), leur seuil de détection auditive augmente (Liu & Yan, 2007 ; Mills, Schmiedt, & Dubno, 2006), leurs fonctions cognitives, essentiellement celles de haut niveau, se dégradent (Glisky, 2007). Tout type de perception multimodale étant dépendant aussi bien des caractéristiques (notamment de la qualité) des inputs sensoriels entrant que du fonctionnement cognitif du système, le processus de vieillissement peut affecter les mécanismes de la perception multimodale en général et ceux de la perception bimodale de la parole en particulier (pour une revue, voir Freiherr, Lundström, Habel, & Reetz, 2013). Toutefois, un cerveau humain vieillissant est aussi sujet à des mécanismes de plasticité neuronale. Aussi, il peut se réorganiser, sur le plan structurel et

fonctionnel, et développer des modes de fonctionnement compensateurs/des stratégies compensatrices en vue d'atteindre un niveau optimal de fonctionnement dans les conditions qui sont les siennes (e.g., Burke & Barnes, 2006 ; Greenwood & Parasuraman, 2010). Les études qui ont exploré le développement de la perception bimodale de la parole ont tenté avant tout de répondre à la question si, comment, à quel degré et à quel niveau le traitement bimodal de la parole est affecté par le processus de vieillissement.

2.4.1 Facilitation audio-visuelle

La grande majorité des études qui se sont intéressées aux changements dans le traitement bimodal de la parole induit par le processus de vieillissement normal ont pris en compte l'effet facilitateur qu'un input bimodal, audio-visuel, induit au niveau de la perception de la parole. Ce choix méthodologique est ainsi le plus fréquent dans l'étude du phénomène d'efficacité inversée qui est une question centrale dans la théorisation sur l'évolution des capacités individuelles de la perception bimodale de la parole au cours du développement âgées.

Le principe d'efficacité inversée désigne un mode de fonctionnement cognitif compensateur (revoir la section 1.2 du chapitre 1). Plus précisément, il stipule qu'avec le déclin de la performance du système cognitif dans le traitement perceptif unimodal, l'efficacité du traitement multimodal augmente (Stein & Meredith, 1993). Dans le domaine de la perception de la parole, le principe d'efficacité inversée implique que toute dégradation de l'information auditive ou visuelle, qui induit un déclin de la perception unimodale, s'accompagne d'une augmentation de l'efficacité de la perception bimodale[25]. En appliquant ce principe à la perception bimodale de la parole chez les personnes âgées, dont les capacités sensorielles se dégradent, ces dernières devraient connaître une augmentation de l'efficacité de la perception audio-visuelle de la parole.

Les résultats de certaines études sont en effet compatibles avec la prédiction issue du principe d'efficacité inversée. Par exemple, Laurienti, Burdette, Maldjian et Wallace (2006) ont mesuré les temps de réponse des jeunes adultes et des adultes âgés dans une tâche de reconnaissance des stimuli présentés soit de façon unimodale, auditive et visuelle, soit de manière bimodale, audio-visuelle. Les temps de réponses ont été globalement plus bas chez les

[25] Toutefois revoir la section 1.2 du chapitre 1 au sujet de la relation entre le degré de la dégradation de l'information auditive par le bruit et le gain audio-visuel.

adultes jeunes comparés aux adultes âgés. Par ailleurs, les auteurs ont noté une diminution des temps de réponse dans la condition audio-visuelle qui variait en fonction du groupe d'âge. En effet, la facilitation de la réponse induite par un input bimodal a été plus importante chez les adultes âgés. Les auteurs concluent ainsi que le traitement bimodal de la parole représente un mécanisme compensatoire efficace pour le déclin des performances dans la perception unimodale. Behne, Wang, Alm, Arnsten, Eg et Valsø (2007) ont abouti à la même conclusion en ayant étudié l'évolution du traitement bimodal de la parole lors du vieillissement *via* le gain audio-visuel tel qu'il est mesuré dans le cadre du paradigme de dégradation de l'information auditive par le bruit. Ces auteurs interprètent leurs résultats comme un indicateur du changement dans le traitement audio-visuel de la parole qui se prolonge au-delà de l'âge adulte. Pour Behne et *al*. (2007), ces changements impliqueraient une plus grande dépendance de l'information visuelle dans la perception bimodale de la parole, facilitée par l'expérience acquise avec ce type d'input.

Toutefois, de nombreuses autres études qui ont exploré le traitement bimodal de la parole en s'intéressant à l'effet facilitateur des performances perceptives induit par un input audio-visuel n'ont pas répliqué les observations de Laurienti et *al*. (2006) et Behne et *al*. (2007). En effet, si certains auteurs n'ont trouvé aucune différence entre les adultes jeunes et âgés quant au gain de la performance perceptive tiré d'un input bimodal (Behne, Wang, Alm, Arntsen, Eg, & Vals, 2007 ; Cienkowki & Carney, 2002 ; Sommers et *al.*, 2005 ; Tye-Murray, Sommers, Spehar, Myerson, & Hale, 2010 ; Tye-Murray, Spehar, Myerson, Sommers, & Hale, 2011; etc.), d'autres ont même rapporté que les adultes âgés tiraient un profit perceptif considérablement moindre d'un input bimodal que les adultes jeunes (Ross et *al*., 2007). Par exemple, si Sommers et *al*. (2005) n'ont observé aucune différence dans le gain audio-visuel entre les adultes jeunes et âgés, Ross et *al*. (2007) ont rapporté une corrélation négative entre l'âge et les performances perceptives audio-visuelles chez un échantillon d'adultes âgés de 18 à 59 ans. Notons toutefois que les performances des participants dans la condition de présentation auditive d'items ont été contrôlées dans l'étude de Sommers et *al*. (2005), mais pas dans celle de Ross et *al*. (2007). Une autre faiblesse méthodologique affectant la validité des résultats de Ross et *al*. (2007) est un échantillon assez réduit, comportant uniquement 9 participants.

En somme, les résultats de la plupart des études qui ont pris en compte l'effet facilitateur de la perception de la parole induit par un input bimodal pour étudier les capacités de traitement audio-visuel de la parole des personnes âgées n'ont ainsi pas confirmé la prédiction du principe de l'efficacité inversée. Une des explications possibles à de tels résultats pourrait se trouver

dans le déclin des capacités à identifier les caractéristiques acoustiques des sons produits à partir de l'input visuel seul avec l'âge. En effet, plusieurs auteurs ont rapporté que la lecture sur les lèvres était moins exacte chez les adultes âgés que chez les adultes jeunes (e.g., Arlinger, 1991 ; Gordon & Allen, 2009 ; Sommers et *al*., 2005 ; Tye-Murray et *al*., 2010). Par ailleurs, Tye-Murray et *al*. (2011) ont trouvé une meilleure résistance aux altérations dans l'acuité l'input visuel lors de la perception bimodale de la parole chez les adultes jeunes comparé aux adultes âgés. En effet, ces auteurs ont observé que toute altération sous forme de floutage de l'input visuel, se soldait par une baisse du gain audio-visuel chez les adultes âgés mais pas chez les adultes jeunes. Néanmoins, Sekiyama et *al*. (2014) ont rapporté que la capacité à lire sur les lèvres des adultes âgés était comparable à celle des adultes jeunes.

Les résultats des études portant sur le développement du traitement bimodal de la parole lors du vieillissement en prenant en compte l'effet facilitateur de la perception induit par un input bimodal sont ainsi assez divergents. Ces divergences sont probablement liées, au moins en partie, aux différences méthodologiques des études en lien avec les conditions expérimentales. Une telle dimension méthodologique a été récemment étudiée par Stevenson, Neums, Baum, Zurkovsky, Barense, Newhouse et Wallace (2015). En effet, ces auteurs se sont intéressés à l'effet facilitateur de l'input bimodal dans la perception des phonèmes et des mots chez les adultes jeunes et âgés. Ayant utilisé le paradigme de la dégradation de l'information auditive par le bruit avec différents niveaux de dégradation, les auteurs n'ont trouvé aucune différence dans les variations du gain audio-visuel en fonction du SNR entre les deux groupes pour la reconnaissance des phonèmes. En revanche, le gain audio-visuel était considérablement moindre chez les adultes âgés pour la reconnaissance des mots. Stevenson et *al*. (2015) en concluent ainsi que le traitement bimodal de la parole n'est affecté par le vieillissement qu'à un niveau relativement complexe, celui des mots, alors qu'il reste intact au niveau plus élémentaire. Aussi, le type d'items semble être un facteur à prendre en compte dans l'interprétation des résultats des études précédentes. C'est un résultat intéressant pouvant potentiellement expliquer une partie des divergences des résultats des études précédentes qui attend cependant d'être répliqué.

2.4.2 Fusion audio-visuelle

Le paradigme de McGurk-MacDonald, nous l'avons vu, est celui qui permet la mesure la plus directe du traitement bimodal de la parole. En effet, d'une part, le degré de susceptibilité à l'effet McGurk-MacDonald refléterait les capacités d'intégration audio-visuelle d'un

individu. D'autre part, le type de réponse face à un stimulus audio-visuellement incogruent permet d'évaluer le poids relatif de chaque composante sensorielle, auditive et visuelle, dans l'intégration audio-visuelle. Tout comme dans les études ayant exploré le développement du mécanisme de l'intégration audio-visuelle durant l'enfance, les études s'étant intéressées à l'impact du processus de vieillissement sur le mécanisme en question ont eu recours au paradigme de McGurk-MacDonald pour étudier les deux dimensions mentionnées.

Les études ayant eu recours au paradigme de McGurk-MacDonad afin d'explorer l'évolution de la perception audio-visuelle de la parole lors du vieillissement sont peu nombreuses. En revanche, leurs résultats sont assez consistants. Ainsi, l'étude de Cienkowski et Carney (2002) a eu pour objectif de répondre à deux questions. Premièrement, les adultes âgés, sont-ils aussi performants que les adultes jeunes dans l'intégration audio-visuelle lors de la perception de la parole ? Deuxièmement, les performances dans l'intégration audio-visuelle, sont-elles en relation avec les performances en lecture de la parole sur les lèvres. Cienkowski et Carney (2002) ont réalisé leur étude sur 3 groupes de participants. Le premier groupe consistait en adultes jeunes avec des capacités sensorielles normales, le deuxième groupe consistait en adultes âgés et le troisième groupe (le groupe contrôle) en adultes jeunes chez qui les seuils de reconnaissance auditive ont été augmentés par l'ajout du bruit de façon à ce qu'ils soient équivalents à ceux des adultes âgés. Les participants ont été testés sur une tâche de reconnaissance auditive de syllabe (évaluation des seuils de reconnaissance auditive), une tâche de lecture sur les lèvres et sur une tâche de type McGurk-MacDonald. Les résultats ont montré que la tendance à procéder à l'intégration audio-visuelle a été comparable parmi les trois groupes. Conformément à ce qui a été trouvé par Stevenson et *al*. (2015) (voir la section précédente), ceci suggère que l'intégration audio-visuelle n'est pas affectée par le vieillissement au moins pour un niveau relativement élémentaire, celui des syllabes. En revanche, l'analyse des réponses ne relevant pas de fusion audio-visuelle a révélé des différences entre, d'une part, les adultes âgés et le groupe contrôle et d'autre part, les adultes jeunes. En effet, si les adultes jeunes optaient plus souvent pour la réponse conforme à la modalité auditive du stimulus McGurk-MacDonald, le pattern a été inversé pour les adultes âgés et le groupe contrôle. Finalement, la capacité à identifier les unités syllabiques à partir de l'information visuelle a été inférieure dans le groupe des adultes âgés comparativement aux deux autres groupes. Toutefois, les performances en lecture de la parole sur les lèvres n'ont pas été corrélées avec la tendance des sujets à percevoir la fusion illusoire de type McGurk-MacDonald. Globalement, Cienkowski et Carney (2002) concluent que, dans les situations où l'intégration audio-visuelle

ne peut pas être réalisée, les sujets choisissent la réponse qui correspond à la modalité apportant l'input le moins ambigu. Suivant ce raisonnement, l'information visuelle aurait plus de poids dans la perception de la parole chez les adultes âgés que chez les adultes jeunes.

Les résultats de Cienkowski et Carney (2002) sont concordants avec ceux obtenus par Sekiyama et *al*. (2014). La conclusion finale quant à l'importance de l'information visuelle dans la perception de la parole pour les adultes âgés a été ainsi la même dans les deux études. Toutefois, Sekiyama et *al*. (2014) avancent que la raison pour laquelle l'information visuelle semble exploiter davantage chez les adultes âgés que chez les adultes jeunes réside dans les différences dans la rapidité du traitement de l'input auditif. En effet, des études ayant utilisé la méthode de potentiels évoqués, ont montré que le délai du traitement auditif était plus long chez les personnes âgées que chez les adultes jeunes. Ceci a été observé aussi bien pour la perception de la parole (Tremblay & Ross, 2007) que pour du matériel non linguistique (Schroeder, Lipton, Ritter, Giesser, & Vaughan, 1995). L'objectif de l'étude Sekiyama et *al*. (2014) a été ainsi double. Premièrement, les auteurs ont cherché à savoir si les adultes âgés utilisaient l'information visuelle dans la perception de la parole davantage que les adultes jeunes. Deuxièmement, les auteurs ont testé l'hypothèse d'amorçage visuel (Sekiyama & Burnham, 2008). Selon cette hypothèse, la contribution de l'information visuelle à la perception bimodale de la parole sera plus importante pour les individus qui traitent l'input visuel plus rapidement que l'input auditif, comparé à ceux qui traitent les deux types d'input avec une rapidité comparable. Ainsi, Sekiyama et *al*. (2014) ont mesuré les temps de réponse, reflétant le temps nécessaire à reconnaître une syllabe (/ba/, /da/ ou /ga/) dans les conditions de présentation auditive et visuelle seules. La différence entre les deux mesures de temps de réponse a été utilisée pour évaluer l'effet d'amorçage visuel pour chaque participant. Les stimuli dans la condition auditive de présentation ainsi que dans la condition bimodale de type McGurk-MacDonald ont été dégradés par le bruit à des degrés différents. Ces derniers ont été calibrés en fonction des performances dans la reconnaissance d'items présentés dans la modalité auditive seule. Les résultats de Sekiyama et *al*. (2014) ont révélé un effet d'amorçage visuel plus important dans le groupe des adultes âgés que chez les adultes jeunes. Par ailleurs, les résultats du paradigme McGurk-MacDonald ont montré une plus grande tendance des adultes âgés à opter pour la réponse visuelle dans les cas où la fusion illusoire audio-visuelle n'était pas perçue, et ceci pour l'ensemble de niveaux de dégradation audio-visuelle. Combinant les deux dimensions des résultats, Sekiyama et *al*. (2014) concluent que l'augmentation de l'impact de

l'information visuelle dans la perception de la parole qui semble apparaître avec l'âge est due à un ralentissement général du traitement de l'input auditif.

Dans la mesure où l'information visuelle affectait davantage la perception de la parole chez les adultes âgés que chez les adultes jeunes, les résultats de Sekiyama et *al*. (2014) sont concordants avec ceux de Cienkowski et Carney (2002). Les explications causales de ce résultat proposées par les auteurs respectifs ne sont cependant pas les mêmes. Par ailleurs, contrairement à Cienkowski et Carney (2002), Sekiyama et *al*. (2014) ont trouvé que la susceptibilité des adultes plus âgés était plus importante que celle des adultes jeunes. Ceci pourrait suggérer que le mécanisme d'intégration audio-visuelle devient en effet plus performant avec l'âge, comme le prédit le principe d'efficacité inversé. Aussi, il s'agit ici d'une différence fondamentale entre les résultats des deux seules études dans le domaine en question. Elle pourrait s'expliquer par les différences méthodologiques, notamment par une dégradation supplémentaire de l'information auditive appliquée par Sekiyama et *al*. (2014). Cette hypothèse reste ainsi à vérifier.

2.4.3 Corrélats neuronaux

Au vu des divergences au niveau des résultats des études comportementales, les études explorant la dimension neuro-fonctionnelle et neurophysiologique de la perception bimodale de la parole pourraient apporter un éclairage important quant à l'impact du processus de vieillissement sur le phénomène en question. Néanmoins, ces études étant presque inexistantes, leur impact reste assez limité.

Le domaine en question comporte une seule étude qui a eu recours à l'approche de potentiels évoqués (Winneke & Phillips, 2011). Elle s'est intéressée aux différences développementales, relatives au processus de vieillissement, dans l'impact d'un input bimodal, audio-visuel, sur la perception de la parole telle qu'elle est traitée au niveau du A1. Il s'agit de l'approche classique présentée à de nombreux endroits dans le document présent

L'étude de Winneke et Phillips (2011) a exploré l'impact du vieillissement sur les mécanismes neurophysiologiques sous-tendant l'effet facilitateur de l'information bimodale dans la perception de la parole. Ayant utilisé le paradigme de la dégradation de l'information

auditive par le bruit couplée avec la tâche de catégorisation d'objets[26], en l'occurrence des mots, les auteurs ont mesuré le taux de reconnaissance correcte des adultes jeunes et des adultes âgés dans les conditions de présentation auditive, visuelle et audio-visuelle. Le niveau de dégradation auditive a été ajusté de manière individuelle de façon à ce que la reconnaissance correcte des mots atteigne un taux environnant 55%. L'analyse des données comportementales a révélé le pattern classique quant aux différences inter-conditions (les performances étaient les meilleures dans la condition de présentation audio-visuelle et les moins bonnes dans la condition de présentation visuelle), ainsi qu'une fluctuation de la performance des adultes âgés comparativement aux adultes jeunes en fonction de la condition de présentation. En effet, les performances des adultes âgés étaient moins bonnes que celles des adultes jeunes dans la condition de présentation visuelle, suggérant ainsi une moins bonne capacité des adultes âgés à lire la parole sur les lèvres. Quant aux données des potentiels évoqués, Winneke et Phillips (2011) ont trouvé une diminution dans la latence de la composante N1. Chose importante, cet effet présentait une variation en fonction du groupe d'âge. En effet, la diminution a été plus importante chez les adultes âgés (pour une présentation graphique de ces résultats, voir la Figure 15).

A la base de ces résultats, notamment ceux basés sur les données neurophysiologiques, les auteurs concluent que le traitement audio-visuel de la parole semble être non seulement intact, mais amplifié chez les adultes âgés. Ceci serait conforme à la prédiction basée sur le principe d'efficacité inversée. Toutefois, le fait que l'effet de l'input bimodal sur les composantes classiquement marquées par le traitement bimodal de la parole, la N1 et la P2, était réduit à la seule diminution de la latence de la N1, couplé avec les observations qui lient un tel changement à l'anticipation du signal auditif induit par l'occurrence du signal visuel (revoir la section 1.5.1 du chapitre 1), pourrait tout simplement indiquer une accélération du traitement de l'information auditive par un input bimodal. En effet, dans la mesure où le traitement auditif semble connaître un ralentissement avec l'âge (e.g., Tremblay & Ross, 2007), une telle stratégie compensatrice répondrait bien aux besoins du système. Il n'est cependant pas sûr qu'on puisse interpréter les résultats de Winneke et Phillips (2011) comme révélant une amélioration globale du traitement bimodal de la parole avec l'âge, d'autant plus que les données comportementales de l'étude ne semblent pas être conformes avec une telle

[26] Dans sa version la plus classique, la tâche de catégorisation d'objets demande des participants de ranger les items de l'expérience dans des catégories proposées. Dans l'étude de Winneke et Phillips (2011), cette tâche a été utilisée pour mesurer le taux de reconnaissance d'items.

interprétation. En effet, le gain audio-visuel était comparable dans les deux groupes d'âge. Comme pour de nombreux champs de recherche présentés dans le chapitre 2, l'étude de l'impact de vieillissement sur la perception bimodale de la parole par la prise en compte des aspects neurophysiologiques et/ou neuro-fonctionnels du phénomène est en attente de nouveaux éléments empiriques pour pouvoir aboutir à un degré raisonnable de validité des conclusions.

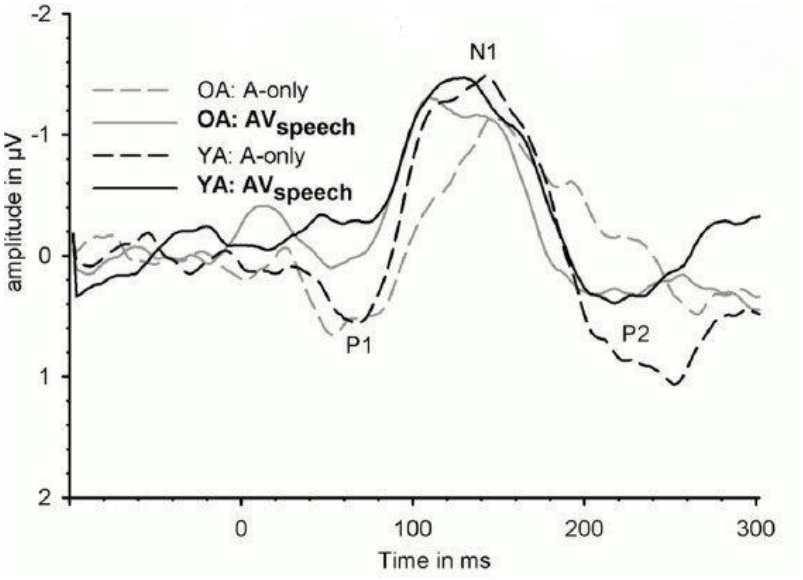

Figure 15. Représentation graphique des variations des composantes P1, N1 et P2 en fonction de l'âge et de condition expérimentale telles que rapportées par Winneke et Phillips (2011).

Le graphique représente la moyenne de l'amplitude (en µV) et de la latence (en ms) des ondes P1, N1 et P2 pour les adultes jeunes (*YA*) et les adultes âgés (*OA*), enregistrés suite à des stimuli auditifs (*A*) ou audio-visuels (*AV*). La réduction de la latence de la N1, consécutive à l'exposition à des stimuli audio-visuels, s'est révélée être plus importante chez les adultes âgés que chez les adultes jeunes. Notons que les valeurs négatives sont représentées sur la partie supérieure de l'axe des ordonnées. (L'image de Winneke et Phillips (2011).)

Résumé du chapitre

Ayant survolé les principales problématiques abordées dans les études portant sur le développement de la perception bimodale de la parole de la naissance à la vieillesse, nous pouvons constater que les éléments empiriques les plus abondants sur le sujet ont été produits pour la période de la petite enfance, plus précisément pour la première année de la vie. Malgré

des difficultés méthodologiques non négligeables pour les études comportementales, relatives à l'absence du langage chez l'enfant durant cette période, il est actuellement relativement bien établi que de nombreuses capacités de perception bimodale de la parole apparaissent dans les premiers mois de la vie, dans une période allant de 2 à 6 mois. En effet, la capacité à reconnaître les correspondances audio-visuelles dans la parole émerge vers l'âge de 4 mois (possiblement même avant). C'est également vers cet âge que les enfants commencent à présenter une susceptibilité à l'effet McGurk-MacDonald. C'est une donnée qui suggère que les mécanismes d'intégration audio-visuelle dans la perception bimodale de la parole atteignent une maturité suffisante pour être opérationnels dans les 4 premiers mois de la vie. L'effet facilitateur de l'input bimodal dans la perception de la parole est également observé. Plus précisément, l'input bimodal pourrait faciliter l'apprentissage des contrastes phonémiques dès l'âge de 6 mois. Les résultats des études comportementales sont validés, au moins partiellement, par des études portant sur les mécanismes neurophysiologiques impliqués dans la perception audio-visuelle de la parole. L'approche neurophysiologique à la perception bimodale de la parole étant restreinte aux mécanismes de l'intégration précoce, ces derniers semblent en effet opérationnels dès l'âge de 4 mois.

Si les capacités de traitement bimodal de la parole émergent très tôt dans le développement post-natal, leur développement est relativement long. En effet, la maturation des mécanismes impliqués dans la perception audio-visuelle de la parole semble durer pendant toute l'enfance, voire même jusqu'à l'adolescence. Durant cette période, la sensibilité des enfants à l'effet McGurk-MacDonald reste moindre comparée à celle des adultes, ce qui suggère une moindre efficacité des mécanismes d'intégration audio-visuelle. Les enfants sont également moins influencés par l'information visuelle relative à la parole et tirent un moindre bénéfice pour la perception de la parole de la présentation d'un input bimodal comparativement à l'input auditif seul. Les études s'étant intéressées aux corrélats neurophysiologiques des mécanismes de la perception bimodale de la parole étant très peu nombreuses et ayant produit des résultats non concordants, elles ne peuvent apporter d'éclairage supplémentaire dans ce domaine de recherche. Finalement, l'étude des aspects neuro-fonctionnels du traitement audio-visuel de la parole semble apporter des éléments suggérant l'implication du sillon temporal supérieur dans le développement des circuits de la perception de la parole. Son implication serait cependant indirecte ; son rôle précis n'est pas connu. En conclusion, il semblerait que le développement des mécanismes de perception bimodale de la parole durant l'enfance pourrait s'expliquer soit par une optimisation de l'extraction des indices acoustiques à partir de l'input visuel qui se

ferait sous l'effet de l'expérience perceptive, soit d'une optimisation de l'intégration des indices visuels et auditifs, soit par des changements sur les deux plans. L'éclairage à de nombreuses questions ouvertes sera apporté par la recherche future.

Les mécanismes de la perception bimodale de la parole connaissent également des changements qui accompagnent le processus de vieillissement. Ces derniers pourraient être consécutifs à un déclin des fonctions sensorielles et possiblement aussi à celui des fonctions cognitives. L'hypothèse du principe de l'efficacité inversée prévoit que la perception audio-visuelle de la parole sera optimisée avec l'âge, composant ainsi le déclin dans l'acuité des inputs sensoriels respectifs. La plupart des études ayant exploré les variations de l'effet facilitateur de l'input bimodal dans la perception de la parole ou encore le paradigme de McGurk-MacDonald n'ont cependant pas confirmé la prédiction du principe d'efficacité inversé. En effet, les résultats de ces études suggèrent une efficacité comparable du traitement bimodal de la parole entre les adultes jeunes et les adultes âgés quand les seuils auditifs sont contrôlés. La seule étude ayant exploré les possibles changements dans les mécanismes neurophysiologiques au cours du vieillissement a apporté des éléments qui suggèrent que la présence d'un input bimodal pourrait essentiellement accélérer le traitement de l'input auditif et ainsi compenser le déclin de la rapidité de traitement qui touche cette modalité sensorielle. Tout comme pour le développement de la perception bimodale de la parole durant la période de l'enfance, les effets du vieillissement dans ce domaine restent encore peu connus. Un nombre peu important d'études et des résultats relativement divergents sont un obstacle évident à la mise en place de conclusions valides en l'état actuel de recherche.

3 Information faciale et son traitement dans la perception audio-visuelle de la parole

3.1 Introduction

Comme nous l'avons exposé dès le début du présent document, la perception audio-visuelle de la parole repose sur deux types d'input sensoriel, auditif et visuel. Aussi, une veine d'études dans ce domaine s'est intéressée aux différentes dimensions/caractéristiques de l'input visuel et leur impact sur la perception audio-visuelle de la parole, ainsi qu'à la façon dont l'information visuelle est traitée dans ce type de perception. Les problématiques de cette veine d'études sont multiples et relativement hétéroclites. Ce chapitre a ainsi pour objectif d'en

présenter celles qui ont été les plus fréquemment traitées sans se vouloir être exhaustif sur le sujet.

3.2 Quantité, qualité et type de l'information faciale impliquée dans la perception audio-visuelle de la parole

Dès les débuts de la recherche en perception audio-visuelle de la parole, trois catégories de problématiques ont émergé au sujet de l'input visuel dans ce cadre. La première concerne la quantité de l'information visuelle, provenant des différents articulateurs, et son impact sur la perception audio-visuelle de la parole. Il s'agit ici d'identifier la quantité minimale provoquant un certain degré d'intégration audio-visuelle, ainsi que la quantité optimale de l'input visuel pour ce type de perception. La deuxième catégorie de problématiques concerne le rôle/l'apport des différents types d'information faciale dans/à la perception bimodale de la parole. Notons qu'avec le terme « type d'information faciale », nous désignons essentiellement l'information des différents articulateurs et des différentes régions faciales. Finalement, la troisième catégorie de problématique est en lien avec l'effet des manipulations dans la qualité de l'input visuel sur la perception bimodale de la parole. Les trois catégories de problématiques ne sont cependant pas indépendantes. En effet, dans la plupart des cas, les manipulations expérimentales dans une des trois dimensions des stimuli visuels s'accompagnent de changement dans l'une ou les deux autres.

3.2.1 Information minimale sur les mouvements articulatoires

La recherche sur l'effet des différentes dimensions de l'information faciale sur la perception audio-visuelle de la parole est un champ large et relativement hétéroclite. Toutefois, les premières études du domaine ont été assez consistantes quant à la problématique traitée. Cette dernière concernait les variations dans l'effet de l'information visuelle sur la perception audio-visuelle de la parole en fonction de la quantité, de l'information faciale disponible. La quantité de l'information faciale impliquant différentes composantes faciale ainsi que différents types de paramètres spatiaux (mouvement, forme, position), ce type de problématique est nécessairement lié à ces deux dimensions.

La première étude ayant porté sur la problématique en question est celle de Sumerfield (1979). Ayant utilisé le paradigme de la dégradation de l'information auditive par le bruit, l'auteur a proposé trois conditions de présentation audio-visuelle des items expérimentaux (des

syllabes) différant dans la quantité de l'information faciale qu'elles comportaient. La première condition comportait le visage entier de la personne prononçant les items, la deuxième condition comportait uniquement les lèvres de l'orateur. Finalement, dans la troisième condition, l'information faciale relative à la production de la parole a été réduite à quatre points lumineux, placés aux extrémités verticales et horizontales des lèvres de l'orateur. Avec un SNR de -12dB, Summerfield (1979) a montré que l'effet facilitateur de l'information visuelle dépendait de la quantité de l'information faciale présentée. Plus précisément, les gains audio-visuels pour les trois conditions de présentation audio-visuelle respectives étaient de 43, 31 et 8%. Toutefois, le gain de 8%, associé au format comportant les points lumineux posés sur les lèvres de l'orateur, n'était pas significatif.

Se basant sur ces résultats, on pourrait supposer que l'information relative au seul mouvement des lèvres (troisième condition de présentation), mais pas sur leur forme et leur position exacte, manque de spécification sur les qualités acoustiques des sons produits. La deuxième partie de l'expérience de Summerfield (1979) a en effet montré que la reconnaissance du phonème /b/ dans la syllabe /aba/ dépendait de l'information relative aux mouvements des lèvres dans leur intégralité, ne comportant aucune ambiguïté quant à la fermeture de la bouche. Sachant que ces indices visuels sont en lien avec le point d'articulation de la consonne /b/ et que différents types de points d'articulation n'impliquent pas uniquement les lèvres, Summerfield, MacLeod, McGrath, & Brooke (1989) ont conduit une seconde étude pour montrer que l'ajout des dents, induites de peinture ultraviolette, aux vidéos des lèvres de l'orateur est associé à un bénéfice supplémentaire de la perception audio-visuelle de la parole.

Se basant sur les résultats de l'étude de Summerfield (1979), plus précisément sur l'absence de l'avantage audio-visuel tiré de l'input visuel dans lequel l'information faciale était réduite à quatre points lumineux circonscrivant les lèvres de l'orateur, Rosenblum, Johnson, & Saldaña (1996) ont exploré l'effet de l'extension de l'information visuelle sous forme de points lumineux à d'autres points des lèvres et à d'autres régions faciales dans un contexte de perception bimodale de la parole. Un second objectif de l'étude consistait à évaluer un éventuel effet de l'apprentissage perceptif, dépendant de l'expérience perceptive avec les indices lumineux proposés dans l'expérience. Aussi, Rosenblum et *al*. (1996) ont construit quatre conditions de présentation audio-visuelle : (i) condition « visage » dans laquelle le visage entier de l'orateur était visible, (ii) condition « lèvres et points lumineux » dans laquelle la forme des lèvres de l'orateur a été précisée au moyen de 14 points lumineux, (iii) « points lumineux sur les lèvres, dents et langue » dans laquelle les points lumineux étaient placés sur les trois

composantes en question, et (iv) « points lumineux sur le visage » dans laquelle des points lumineux étaient placés également sur le front, le nez, le menton et la mâchoire. Les résultats ont montré que la reconnaissance des phrases dégradées par le bruit a été significativement améliorée par tous les types de présentation audio-visuelle des stimuli. La performance perceptive a été la meilleure dans la condition « visage », elle était suivie par les conditions « points lumineux sur les lèvres, dents et langue » et « points lumineux sur le visage », dans lesquelles la performance était comparable. Finalement, la perception audio-visuelle a été la moins bonne dans la condition « lèvres et points lumineux » (pour une présentation graphique des résultats, voir la Figure 16). Les auteurs ont également trouvé un effet global de l'apprentissage perceptif, qui ne variait pas en fonction de la condition de présentation audio-visuelle.

Globalement, les résultats de Rosenblum et *al*. (1996) semblent montrer que les indices sur les mouvements du menton et la mâchoire ne sont pas particulièrement informatifs pour la perception de la parole. Par ailleurs, contrairement à Summerfield (1979), les indices des mouvements des lèvres, réduits aux seuls points lumineux, peuvent faciliter la perception de la parole à condition qu'ils soient suffisamment fidèles aux caractéristiques des articulateurs que sont les lèvres. Les indices visuels portant sur la temporalité et l'énergie acoustique du signal acoustique ne semblent pas facilitateurs de la perception de la parole s'ils ne sont pas suffisamment suggestifs de l'information visuelle rencontrée dans des situations écologiques[27]. Par ailleurs, les indices visuels en lien avec les lèvres, même s'ils sont considérablement réduits, apparaissent comme un élément clef dans la perception audio-visuelle de la parole. Néanmoins, les indices des autres articulateurs et d'autres régions faciales semblent nécessaires pour une perception audio-visuelle optimale.

[27] Bernstein et *al*. (2004) ont toutefois trouvé un effet facilitateur des stimuli visuels qui consistaient en formes géométriques dont les changements dans la taille verticale ou horizontale présentaient une corrélation avec l'enveloppe acoustique de l'input auditif. Toutefois, la taille de l'effet était très réduite par rapport à la vision de la bouche de l'orateur. Par ailleurs, l'effet facilitateur de ces stimuli géométriques était comparable à celui d'une stimulus qui signalait simplement le début et la fin de voisement.

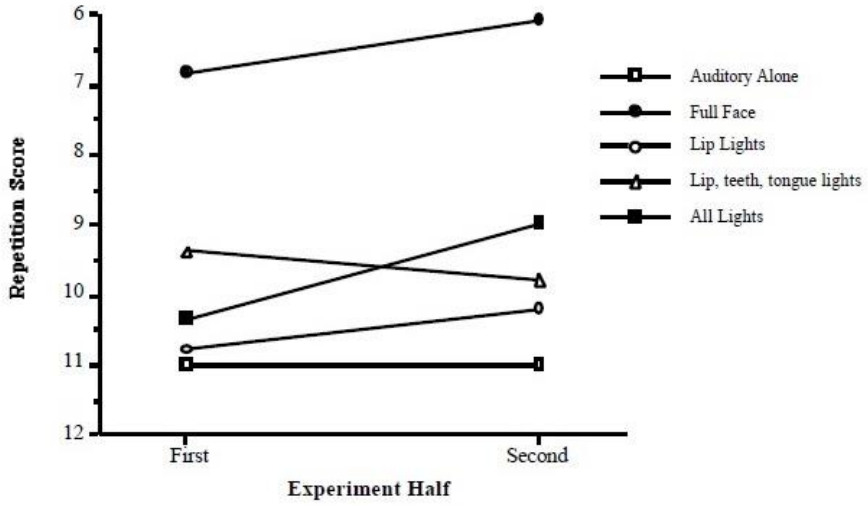

Figure 16. Représentation graphique de certains résultats de Rosenblum et *al*. (1996).

L'image est une représentation graphique du taux de répétitions correctes d'items expérimentaux (*Répétition Score*) en fonction de la phase de l'expérience (*Experiment Half*), première partie (*First*) et deuxième partie (*Second*), et en fonction de la condition de présentation d'items, présentation auditive (*Auditory Alone*), visage (*Full Face*), lèvres et points lumineux (*Lip Lights*), points lumineux sur les lèvres, dents et langue (*Lip, teeth, tongue lights*), points lumineux sur le visage (*All Lights*). On note que la performance a été la meilleure dans la condition « visage », même si toutes les conditions audio-visuelles différaient significativement de la condition de présentation auditive. L'effet simple du facteur « phase de l'expérience » a été également observé ; la performance a été globalement meilleure dans la deuxième moitié de l'expérience. Toutefois, aucun effet d'interaction n'a été trouvé. (L'image de Rosenblum et *al*. (1996).)

3.2.2 Bouche et régions extra-buccales

Les études utilisant la méthode de points lumineux et de peinture ultraviolette, avec laquelle il est possible de réduire considérablement l'information visuelle, semblent montrer le rôle crucial des lèvres (voir aussi Benoît, Guiard-Marigny, Le Goff, & Adjoudani, 1996) et plus largement de la région buccale dans la perception audio-visuelle de la parole. D'autres études ont produit des résultats en accord avec une telle conclusion. Par exemple, Huyse, Berthomier et Leybaert (2013) ont étudié l'effet de la dégradation de l'information labiale sur la perception audio-visuelle chez les enfants porteurs d'implants cochléaires et les enfants sans déficits auditifs. Ayant utilisé le paradigme de la dégradation de l'information auditive et de

l'information visuelle, avec des stimuli audio-visuellement congruents et incongruents, les auteurs ont trouvé que la dégradation de l'information labiale menait à une réduction significative du gain audio-visuel, ainsi qu'à une plus grande influence de l'information auditive sur les réponses des sujets. Le même pattern a été observé aussi bien chez les enfants sans déficits auditifs que chez les enfants porteurs des implants cochléaires. De plus, la recherche a montré que la vision de la bouche de l'orateur est suffisante aussi bien pour la perception de la parole à partir de la modalité visuelle seule que pour induire l'effet McGurk-MacDonald (Haitinen, Manninen, Sams, & Surakka 2000 ; Jordan & Thomas, 2011 ; Rosenblum, Yakel, & Green, 2000). Dans la continuité de ces travaux, de nombreux auteurs se sont intéressés à l'efficacité de la vision de la région buccale seule comparativement à la vision du visage entier dans la perception audio-visuelle (Berger, Garner, & Sudman, 1971 ; Greenberg & Bode, 1968 ; Ijsseldijk, 1992 ; Marassa & Lansing, 1995 ; Stone, 1957). Une telle approche permettrait de déterminer si la région buccale est porteuse de tous les indices visuels qui sont essentiels dans ce domaine. Les résultats de ces études sont relativement consistants. En effet, si Ijsseldijk (1992) a trouvé une efficacité moindre de la région buccale seule, les autres auteurs ont rapporté un effet similaire du visage entier et de la région buccale sur la perception audio-visuelle de la parole.

Les résultats des études ayant comparé deux types de format visuel, l'un comportant le visage entier et l'autre comportant la région buccale seule, suggèrent ainsi que les indices visuels apportés par la région buccale sont suffisants pour une extraction optimale d'informations sur l'aspect acoustique des sons produits. Toutefois, Thomas et Jordan (2004) exposent des points méthodologiques pouvant être problématiques pour une telle interprétation des résultats. Premièrement, dans la plupart des études (Berger et *al*., 1971 ; Greenberg & Bode, 1968 ; Marassa & Lansing, 1995), le format de la région buccale comportait d'autres composantes extra-buccales telles que la mâchoire et le larynx. Deuxièmement, le format comportant la région buccale a été le plus souvent créé en appliquant un masque sur le reste de l'image. Or, le masque, qui induit des contrastes de luminosité assez importants, pourrait affecter l'attention visuelle du sujet et, consécutivement, le traitement de l'information visuelle. Ceci pourrait éventuellement expliquer les résultats d'Ijsselsijk (1992). Or, les détails de la création du format présentant la région buccale seule n'ont pas été donnés par l'auteur. Troisièmement, si l'information visuelle limitée à la région buccale a été présenté en dehors du contexte facial dans l'étude d'Ijsselsijk (1992), ces résultats pourraient également refléter l'interruption dans le traitement holistique. En effet, selon certains éléments empiriques, le

traitement de l'information faciale dans la perception audio-visuelle de la parole pourrait se faire de façon holistique (ce point est développé dans la section 3.2 d ce chapitre).

Pour contourner les problèmes méthodologiques exposés ci-dessus, Thomas et Jordan (2004) ont ainsi créé trois formats, un comportant la région buccale seule sans contrastes visuels, un autre comportant le visage entier de l'orateur dans lequel seule la région buccale était active et un dernier avec le visage entier de l'orateur normalement/entièrement actif (pour les exemples du matériel visuel de l'étude de Thomas et Jordan (2004), voir la Figure 17). Les auteurs n'ont trouvé aucune différence dans l'ampleur de l'effet facilitateur de l'input bimodal de la perception de la parole noyée dans le bruit. De tels résultats semblent ainsi confirmer que la région buccale apporte tous les éléments nécessaires pour l'extraction des informations acoustiques sur les éléments de la parole.

Lors de la production de la parole, les mouvements des régions extra-buccales sont en lien avec ceux de la région buccale. Plus précisément, les transitions dynamiques dans les formes des articulateurs de la région buccale et celles observées dans les régions extra-buccales, telles que les joues et la mâchoire, lors de la production verbale orale présenteraient une corrélation très forte, égale ou supérieure à 0,95 ou plus (Munhall & Vatikiotis-Bateson, 1998 ; Vatikiotis-Bateson, Munhall, Hirayama, Lee, & Terzopoulos, 1996). De ce fait, il est légitime de supposer que les mouvements des régions extra-buccales peuvent également être porteurs d'indices utiles pour la perception de la parole. Les éléments empiriques produits sur ce sujet sont en effet compatibles avec cette supposition. Par exemple, Scheinberg (1980) a établi que la perception des phonèmes ayant un aspect similaire quant aux patterns articulatoires dynamiques au niveau de la région buccale pouvait être facilitée par les indices en lien avec le gonflement des joues. Utilisant la technique de masquage, Preminger, Lin, Payen, et Levitt (1998) ont montré que les sujets étaient capables de tirer un certain bénéfice de l'information faciale lors de la perception audio-visuelle de la parole, même en l'absence d'informations provenant de la région buccale. Thomas et Jordan (2004) ont également rapporté un effet facilitateur de l'input audio-visuel dans la perception de la parole avec des formats visuels dans lesquels la bouche restait statique durant la production orale des mots. Finalement, Davis et Kim (2006) ont trouvé que la vision de la seule partie supérieure du visage de l'orateur facilitait la perception de la parole dégradée par le bruit. Par ailleurs, les sujets adultes semblent également capables de lier une des deux vidéos silencieuses présentant la partie supérieure de la tête de l'orateur à la phrase correspondante. Aussi, il semblerait que, malgré une importance majeure des informations visuelles provenant de la région buccale, les mouvements dans les

régions extra-buccales peuvent également apporter certains indices utiles pour le traitement de la parole.

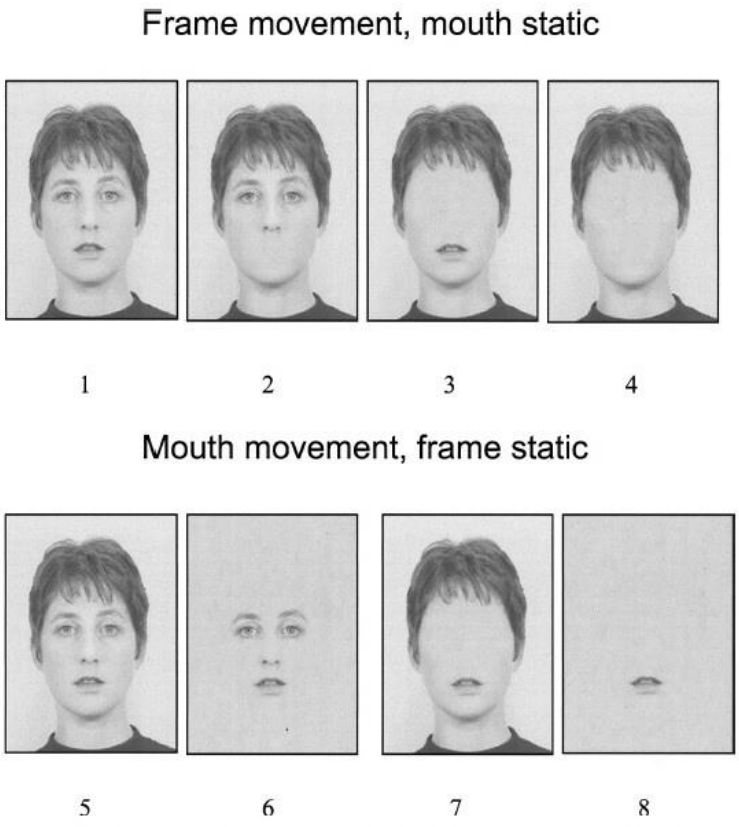

Figure 17. Représentation du matériel visuel relatif aux différentes conditions expérimentales de l'étude de Thomas et Jordan (2004).

L'image représente des exemples du matériel visuel construit par Thomas et Jordan (2004) pour étudier l'apport des mouvements extra-buccaux à la perception de la parole. La partie supérieure de l'image (*Frame movement, mouth static*) comporte les différents types de stimuli utilisés dans la condition où les régions extra-buccales de l'orateur étaient actives, alors que la bouche était maintenue statique. La partie inférieure (*Mouth movement, frame static*) de l'image présente les différents types de matériel visuel dans lequel la bouche de l'orateur était active lors de la production des items expérimentaux, alors que les régions extra-buccales étaient maintenues statiques. Pour chacune des deux conditions, nous voyons les variantes/les différents types du matériel qui diffèrent en fonction de la présence/absence des composantes suivantes : (i) yeux/nez, (ii) régions extra-buccales, (iii) bouche. (L'image de Thomas et Jordan (2004).)

3.2.3 Articulateurs invisibles

Les mouvements articulatoires provenant de la région buccale et ceux des régions extra-buccales semblent ainsi apporter des indices, pouvant être aussi bien temporels qu'acoustiques, qui sont utilisés lors de la perception audio-visuelle de la parole et permettent une meilleure identification des unités de la parole. Badin, Tarabaka, Elisei et Bailly (2010) soulignent que d'importants indices phonémiques, pouvant être particulièrement utiles dans ce domaine, sont portés par la langue. En effet, les mouvements de la langue et sa position par rapport au palais ou encore l'activité laryngienne sont porteurs d'informations pouvant être très utiles dans la désambiguïsation phonémique (Cornett, 1967). Or, les mouvements de la langue sont, en grande partie, cachés, et donc inaccessibles pour la lecture de la parole sur les lèvres. Badin et *al*. (2010) se sont ainsi posé la question sur la capacité humaine à extraire des indices acoustiques à partir des mouvements de la langue quand celle-ci est rendue entièrement visible lors de la production de la parole.

Ayant utilisé le paradigme de la dégradation de l'information auditive par le bruit avec différents SNRs, les auteurs ont exposé les participants à trois conditions de présentation audio-visuelle : (i) coupe sagittale du visage sans langue (AVSL), (ii) coupe sagittale du visage avec langue (AVL), et (iii) visage de l'orateur présenté de profil (AVV)[28] (condition écologique de présentation) (pour des exemples du matériel, voir la Figure 18). Les résultats de l'étude ont montré que la reconnaissance des stimuli expérimentaux, des syllabes, a été facilitée le plus grandement et de façon équivalente par les formats AVV et AVL. Une facilitation moins importante a été observée dans la condition AVSL. Aussi, si la langue semble apporter des indices qui facilitent la perception de la parole davantage qu'un même format sans la langue, la lecture de la parole à partir des indices des mouvements de la langue ainsi que d'autres articulateurs n'est pas plus efficace que la lecture sur les lèvres avec un format écologique (AVV). Badin et *al*. (2010) expliquent la différence entre les conditions AVSL et AVV, qui comportaient les mêmes articulateurs visibles, soit par l'absence de la peau et des déformations à ce niveau survenant lors de la production de la parole dans le format AVSL, soit par le fait qu'un format plus écologique est aussi plus facile à traiter. Une donnée importante quant à la capacité à extraire des indices acoustiques des sons à partir des mouvements de la langue vient des résultats d'une seconde phase de test, proposée pour évaluer un éventuel effet

[28] Notons que les abréviations des conditions expérimentales ont été proposées de notre part pour faciliter la suite de la rédaction. Elles ne correspondent pas aux abréviations présentées dans l'article de Badin et *al*. (2010), écrit en langue anglaise.

d'apprentissage. Badin et *al*. (2010) ont en effet observé que ce dernier était présent avec le format AVL, suggérant ainsi la possibilité de développer les capacités perceptives nous permettant de profiter de la riche information acoustique qui peut être fournie par les mouvements de la langue grâce à l'expérience.

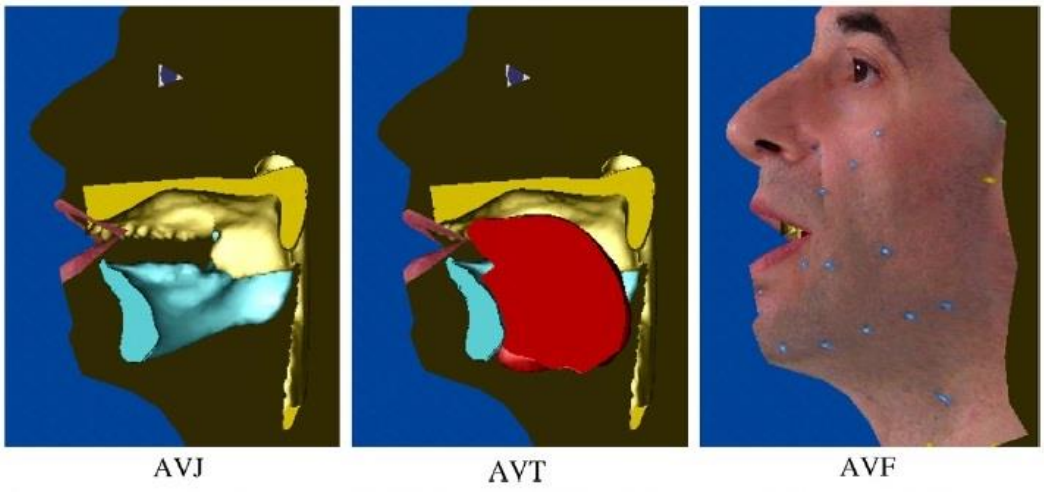

Figure 17. Représentation du matériel visuel utilisé par Badin et *al*. (2010).

L'image représente des exemples des stimuli visuels tels que présentés dans la condition « coupe sagittale du visage sans langue » (audio-video jaw ; AVJ) ; coupe sagittale du visage avec langue (audio-video tangue ; AVT) et visage de l'orateur (audio-video face ; AVF). (L'image de Badin et *al*. (2010).)

3.2.4 Autres aspects de l'information faciale : luminance, acuité, couleur

Outre les dimensions exposées dans les parties précédentes de la section 3.2 du chapitre 3, la recherche s'est également intéressée à l'effet d'autres types de manipulation de l'information visuelle, faciale, sur la perception bimodale de la parole. Ces derniers concernent essentiellement les manipulations de l'acuité de l'information faciale, d'autres types de manipulation sont en lien avec la luminosité et la couleur de l'information visuelle, ou encore avec la taille de l'information faciale et l'angle visuel sous lequel la personne recevant un message verbal oral fixe l'orateur.[29] Finalement, une étude de Jordan et Thomas (2011) a

[29] Notons que l'orientation de l'information faciale et de ses composantes a également été étudiée. Dans la mesure où ce type de manipulation de l'information faciale affecte l'aspect holistique de traitement de l'information faciale, les études ayant exploré cette dimension de la perception audio-visuelle de la parole se sont intéressée à

également exploré l'effet de la réduction de l'information faciale à la moitié du visage sur la perception audio-visuelle de la parole.

Quant aux manipulations dans l'acuité de l'information faciale, Tye-Murray et *al*. (2011) ont réduit le contraste visuel des vidéos présentant un orateur prononçant des syllabes de 98%. Une telle manipulation a donné lieu à une image extrêmement floue, dans laquelle la plupart des détails et la couleur n'étaient plus identifiables. Ayant utilisé le paradigme de la dégradation de l'information auditive par le bruit, les auteurs ont observé une baisse considérable de l'effet facilitateur de la détection des syllabes dans la condition comportant les vidéos de mauvaise qualité comparativement à la condition comportant des vidéos avec une bonne résolution d'image. Toutefois, la performance perceptive des sujets dans la condition audio-visuelle comportant des vidéos à faible acuité était significativement meilleure comparativement à la condition auditive seule. Ceci suggère que l'extraction d'indices temporels des mouvements articulatoires sur l'occurrence des sons produits est possible même dans le cas où l'information faciale est considérablement dégradée. Dans ce domaine, des résultats similaires ont été obtenus par Munhall, Kroos, Jozan et Vatikiotis-Bateson (2004). Ces auteurs ont procédé à une dégradation de l'acuité de l'information faciale en manipulant les fréquences spatiales du matériel visuel (pour des exemples du matériel visuel utilisé dans les différentes conditions de présentation audio-visuelle par Munhall et *al*. (2004), voir la Figure 19). Utilisant également le paradigme de la dégradation de l'information auditive par le bruit Munhall et *al*. (2004) ont observé un effet facilitateur de la perception de la parole induit par un input bimodal pour une gamme assez large des fréquences spatiales qui était le plus élevé pour des fréquences moyennes (stimuli obtenus par application du filtre de 11-cycles/visage pour fréquences centrales ; voir l'exemple D de la Figure 19). Contrairement à Tye-Murray et *al*. (2011) qui se sont intéressés aux seuils de détection des syllabes, Munhall et *al*. (2004) ont pris en compte l'effet facilitateur de l'input bimodal dans l'identification des mots clefs dans des phrases. Dans la mesure où la reconnaissance des mots est bien plus complexe sur le plan acoustique que la seule détection d'un item, les résultats de Munhall et *al*. (2004) pourraient suggérer que l'être humain est capable d'extraire des indices aussi bien temporels qu'acoustiques à partir d'une information faciale considérablement dégradée. Outre cette supposition, les résultats de Munhall et *al*. (2004) indiquent qu'une haute résolution spatiale de

un éventuel lien entre le traitement bimodal de la parole et le traitement des visages. Cette problématique est traitée au niveau de la section 3.3 du chapitre 3.

l'information visuelle n'est pas nécessaire pour un traitement optimal de cette dernière dans le cadre de la perception bimodal de la parole.

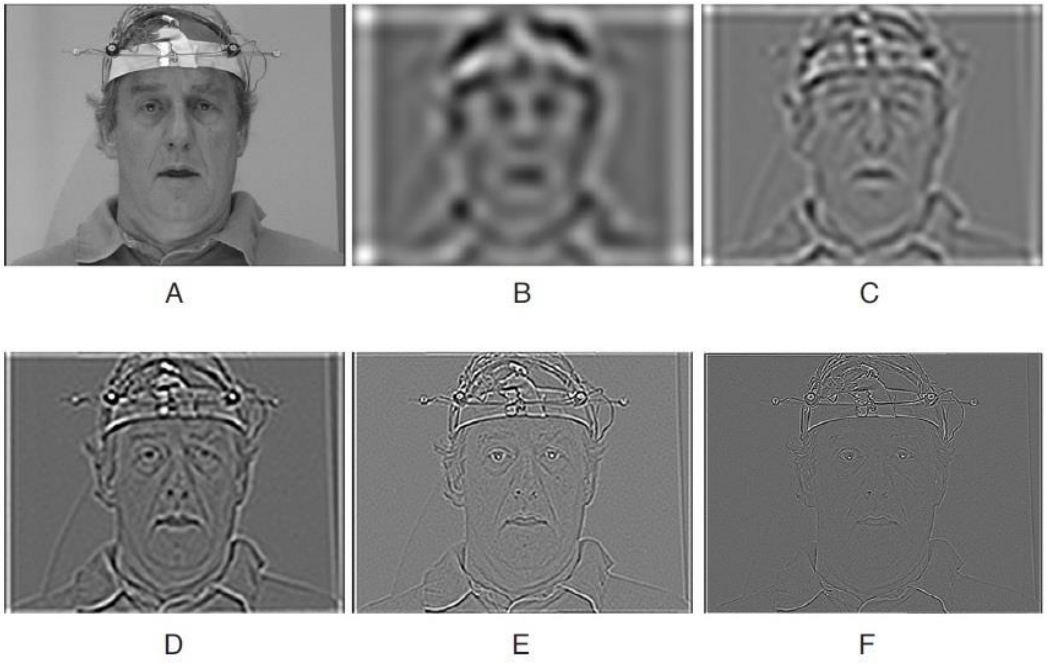

Figure 19. Représentation des exemples des stimuli visuels de Munhall et *al*. (2004).

Les images de l'orateur correspondent aux différentes conditions de présentation audio-visuelle des stimuli. En A, la vidéo non dégradé de l'orateur, les autres exemples sont des images dégradées différant dans la fréquence spatiale. Les stimuli dégradés ont été construits en appliquant des filtres pour différentes fréquences spatiales mesurées en termes du nombre de cycles centrés sur le visage. En B l'image de 2,7 cycles/visage, en C 5,5 cycles/visage, en D 11 cycles/visage, en E 22 cycles/visage et en F 44,1 cycles/visage. Comparativement à la présentation auditive seule, l'identification des stimuli expérimentaux a été significativement meilleure pour l'ensemble des conditions audio-visuelles à l'exception des conditions B et E. (L'image de Munhall et *al*. (2004).)

L'effet de l'information visuelle sur la perception bimodale de la parole semble ainsi assez résistant aux manipulations de la résolution spatiale du matériel visuel. Des observations semblables ont été faites pour d'autres types de manipulations affectant l'information visuelle. En effet, Jordan, McCotter et Thomas (2000) ont établi que l'effet facilitateur de la perception de la parole induit par un input bimodal n'était pas sensible à la présence/absence de la couleur

de l'information faciale. Ayant manipulé aussi bien les contrastes lumineux que la couleur, McCotter et Thomas (2003) ont conclu que ce sont les perturbations dans la luminance de l'information faciale, plutôt que des manipulations de la couleur, qui sont susceptibles d'influencer le traitement de l'information visuelle lors de la perception bimodale de la parole. Une telle conclusion concorde avec les résultats des études précédemment citées. Dans la même lignée de recherche, le traitement audio-visuel de la parole s'est avéré résistant également à des manipulations de la distance entre l'orateur et la personne recevant le message parlé et à des diminutions de la taille de l'information faciale (Jordan & Sergeant, 1998, 2000), ainsi qu'aux variations dans l'angle visuel entre la personne recevant un message verbal et la position de l'information faciale de l'orateur (Jordan, Sergeant, Martin, Thomas, & Thow, 1997 ; Jordan & Thomas, 2001). Finalement, l'étude de Jordan et Thomas (2011) a également trouvé que le traitement bimodal de la parole a été aussi efficace dans les conditions où uniquement la moitié du visage de l'orateur a été visible que dans les conditions avec le visage entier. Aussi, les résultats des études exposées dans cette section montrent globalement une robustesse du traitement audio-visuel de la parole qui résiste bien à une large gamme de manipulation de l'input visuel si ce-dernier garde son aspect écologique.

3.3 Lien entre le traitement de la parole et le traitement des visages

Les visages sont des objets extrêmement importants pour la vie sociale de l'homme. En effet, ils sont porteurs de trois types d'informations au rôle essentiel dans la régulation des relations sociales, de l'identité de l'individu, des expressions faciales et, nous l'avons vu tout le long de ce texte, des informations relatives à la parole (voir Bruce & Young, 2012). Le traitement perceptif des visages a ses propres caractéristiques qui le distinguent de celui des autres objets. Plus précisément, le traitement des visages est davantage marqué par le traitement holistique, c'est-à-dire le traitement du visage dans sa globalité, que par le traitement analytique qui est le traitement des composantes faciales isolées (pour une revue, voir Piepers & Robins, 2012). Traiter un visage de façon holistique signifie que l'information sur ses parties constituantes est intégrée de façon à percevoir le visage en tant qu'un tout, en d'autres termes, en tant qu'une configuration particulière des parties formant un ensemble perceptif unifié et cohérent (e.g., Tanaka & Farah, 1993 ; 2003 ; Taubert, Apthorp, Aagten-Murphy, & Alais, 2011 ; Valentine, 1988). Cette prévalence du traitement holistique est significative pour la perception des visages, alors que la perception des autres catégories d'objets repose essentiellement sur le traitement analytique de leurs composantes (e.g., Biederman, 1987).

Dans la mesure où l'information sur les mouvements articulatoires dans le cadre de la perception bimodale de la parole est portée par le visage de l'orateur, de nombreux auteurs ont étudié l'existence d'une éventuelle similitude dans la nature du traitement entre celui impliqué dans la perception audio-visuelle de la parole et le traitement des visages. Dit autrement, la question qui se pose dans ce cadre est de savoir si le traitement de l'information faciale lors de la perception bimodale de la parole est holistique, dépendant d'un ensemble structuré qu'est le visage, ou analytique, c'est-à-dire centré sur une partie/composante du visage, en l'occurrence sur la partie essentielle pour la perception bimodale de la parole, la bouche.

3.3.1 Marqueurs comportementaux révélateurs du traitement holistique de l'information visuelle dans le cadre de la perception audio-visuelle de la parole

Les études comportementales sur la nature du traitement de l'information visuelle dans le cadre de la perception bimodale de la parole ont été faites pratiquement exclusivement au moyen des marqueurs comportementaux de traitement holistique que la recherche sur la perception des visages a mis en évidence. Ces marqueurs se caractérisent par une baisse de performance perceptive en termes de reconnaissance de l'identité et/ou des expressions faciales dans les conditions où l'information faciale relative à la configuration des composantes du visage, qui est indispensable pour le traitement holistique du visage, est manipulée. Le raisonnement sous-jacent à l'application d'une telle approche méthodologique dans les études portant sur la perception bimodale de la parole est le suivant : Si, lors de la perception bimodale de la parole, l'information visuelle est traitée de façon holistique, alors toute perturbation dans la configuration de l'information faciale induisant une perturbation du traitement holistique de l'information faciale, devrait également induire une baisse de performance dans la perception bimodale (et visuelle seule) de la parole. Autrement dit, les marqueurs du traitement holistique communément observés dans les indices comportementaux relatifs à la perception des visages devraient également apparaître au niveau du comportement reflétant la perception audio-visuelle de la parole. Dans ce contexte, il convient de préciser que la recherche sur la perception audio-visuelle de la parole a utilisé uniquement certains de ces marqueurs, notamment l'effet de l'inversion faciale (*face inversion effect*), l'illusion de Tatcher (*Tatcher illusion*), l'effet des visages à composantes mélangées (*scrambled faces*) et l'effet de la supériorité du tout sur ses parties (*part to whole effect*).

3.3.1.1 Effet de l'inversion faciale

Le marqueur comportemental du traitement holistique des visages le plus utilisé dans la recherche et le mieux connu est l'effet de l'inversion faciale. Il signe le fait que la reconnaissance de l'identité (Goldstein, 1965 ; Rakover & Teucher, 1997 ; Van Belle, De Graef, Verfaillie, Robinson, & Lefèvre, 2010 ; Yin, 1969) ainsi que des expressions faciales (e.g., Prkachin, 2003, Sato, Kochiyama, & Yoshikawa, 2011) est meilleure avec les visages présentés dans une orientation droite, écologique, comparativement à une présentation inversée où les visages ont subi une rotation de 180°. L'effet de l'inversion est bien plus important pour les visages que pour les objets d'autres catégories (e.g., Diamond & Carey, 1986 ; Leder & Carbon, 2006 ; Robbins & McKone, 2007) et est ainsi considéré comme un marqueur comportemental du traitement holistique de l'information faciale.[30]

L'effet de l'inversion faciale est également le marqueur le plus fréquemment utilisé dans l'étude de la nature du traitement de l'information visuelle lors de la perception de la parole. L'approche méthodologique la plus fréquemment appliquée dans ce domaine est celle qui consiste à comparer la performance dans la perception de la parole présentée de façon visuelle ou audio-visuelle entre la condition présentant le visage de l'orateur orienté de façon écologique et la condition où le visage est inversé (Jordan & Bevan, 1997 ; Massaro & Cohen, 1996 ; Thomas & Jordan, 2002). Massaro et Cohen (1996) ont ainsi constaté que l'inversion du visage de l'orateur affectait la perception des syllabes audio-visuellement congruentes. En revanche, Jordan et Bevan (1997) ont rapporté que l'inversion faciale affectait le pattern habituel de la performance perceptive uniquement dans le cas où les stimuli étaient audio-visuellement incongruents. La différence dans les résultats de Massaro et Cohen (1996) d'une part, et ceux de Jordan et Bevan (1997) d'autre part pourrait s'expliquer par les caractéristiques des stimuli. En effet, les consonnes de chaque paire de syllabes utilisées par Massaro et Cohen (1996) avaient le même mode d'articulation. L'information sur le point d'articulation apportée par l'information visuelle avait ainsi un rôle de désambiguïsation pour les qualités acoustiques des consonnes, alors que ce n'était pas le cas pour tous les stimuli utilisés par Jordan et Bevan (1997). Aussi, le poids/l'utilité de l'information visuelle dans la perception des stimuli dans ces études pourrait être explicatif/explicative des différences dans les résultats.

[30] Notons que l'information sur la configuration des composantes faciales, impliquant les relations spatiales entre ces éléments qui permet de les intégrer dans un percept unifié et cohérent, serait perturbée/moins facilement traitée suite à une manipulation des paramètres spatiaux qui caractérisent le visage et permettent son traitement holistique, en l'occurrence suite à son inversion (e.g., Farah, Tanaka, & Drain, 1995 ; Van Belle et *al.*, 2010).

Dans la même lignée de recherche, l'étude de Thomas et Jordan (2002) a cherché à évaluer dans quelle mesure la dégradation de la résolution spatiale de l'information visuelle influence l'effet de l'inversion faciale dans la perception audio-visuelle de la parole. Ces auteurs ont émis l'hypothèse selon laquelle l'effet de l'inversion faciale devrait être plus important dans ce contexte dans les conditions où l'information visuelle est dégradée, car le traitement analytique des composantes faciales, en l'occurrence de la bouche, dépend plus fortement de la résolution de l'image que le traitement holistique de la configuration des composantes faciales. Aussi, en dégradant la résolution spatiale des vidéos, on forcerait le sujet à traiter l'information faciale davantage de façon holistique, ce qui devrait se solder par une baisse de performance dans la perception audio-visuelle de la parole si le traitement de l'information visuelle dans ce contexte implique vraiment une composante holistique. Conformément à cette hypothèse, Thomas et Jordan (2002) ont trouvé l'effet de l'inversion faciale dans la performance perceptive des stimuli audio-visuellement incongruents uniquement dans le cas où l'information visuelle était dégradée. Dans ces conditions, Thomas et Jordan (2002) ont également rapporté l'effet de l'inversion faciale dans la perception visuelle seule des stimuli expérimentaux, des syllabes.

Plus récemment, Vatakis et Spence (2008) ont étudié l'effet de l'inversion faciale sur la performance dans la tâche de jugement temporel pour des stimuli audio-visuels. Cette tâche, qui demande aux participants d'estimer l'ordre temporel d'occurrence de deux stimuli provenant de deux modalités sensorielles distinctes, auditive et visuelle, est fréquemment utilisée pour évaluer les mécanismes de fusion audio-visuelle[31]. Ces auteurs ont constaté que l'inversion faciale affectait le jugement subjectif de la simultanéité des deux inputs et ceci uniquement quand les stimuli utilisés étaient relatifs à la parole produite par un être humain, mais pas dans le cas de la production des stimuli musicaux et de la production des vocalises par un autre primate (pour des exemples des stimuli expérimentaux de Vatakis et Spence (2008), voir la Figure 20). Un tel résultat pourrait indiquer que le traitement de l'information visuelle dans la perception bimodale de la parole est, au moins dans une certaine mesure, holistique. Toutefois, le même effet d'inversion a été trouvé aussi bien pour les stimuli visuels présentant le visage entier de l'orateur que pour ceux qui comportait uniquement la partie inférieure du visage. Or, dans une information faciale aussi partielle, la dimension de la configuration des composantes est très réduite voire absente. Si le traitement sous-jacent au jugement subjectif de la simultanéité des inputs auditif et visuel dépendait de la configuration de l'information faciale,

[31] Toutefois, revoir la section 1.4.1 du chapitre 1 pour les limites de la validité de cette méthode.

on pourrait s'attendre à ce que l'inversion d'une composante faciale relativement isolée du reste du visage ne l'affecte pas. En effet, Rosenblum et *al*. (2000) ont montré que la perception de la fusion illusoire induite par des stimuli de type McGurk-MacDonald n'est pas affectée par l'inversion de la seule composante buccale. Cependant, le format présentant la bouche dans l'étude de Vatakis et Spence (2008) comportait certaines régions extra-buccales (voir la Figure 20) qui auraient pu, peut-être, créer suffisamment de contexte pour déclencher le traitement holistique, à un degré suffisant pour trouver un effet de l'inversion de ce format visuel dans les estimations subjectives de la simultanéité des inputs auditif et visuel dans le cadre de la perception bimodale de la parole.

De manière générale, les résultats des études qui ont utilisé l'effet de l'inversion faciale pour identifier la nature (holistique ou analytique) du traitement de l'information faciale dans la perception bimodale de la parole n'apportent pas de réponse claire à cette question. En effet, l'effet de l'inversion faciale ne semble pas systématique dans la perception bimodale de la parole. Toutefois, la construction des conditions expérimentales dans les études de Massaro et Cohen (1996), Jordan et Bevan (1997) et également Thomas et Jordan (2002) est telle qu'elle crée un contexte dans lequel l'information auditive peut jouer un rôle dominant dans la perception de la parole car clairement audible et non ambiguë. Cet élément aurait pu, éventuellement, réduire ou même annuler l'effet de l'inversion faciale. Toutefois, dans la mesure où l'effet de l'inversion faciale a été observé sous certaines conditions et dans certaines dimensions comportementales dans les études présentées dans cette section, les résultats suggèrent globalement une possible implication du traitement holistique dans le cadre de la perception audio-visuelle de la parole.

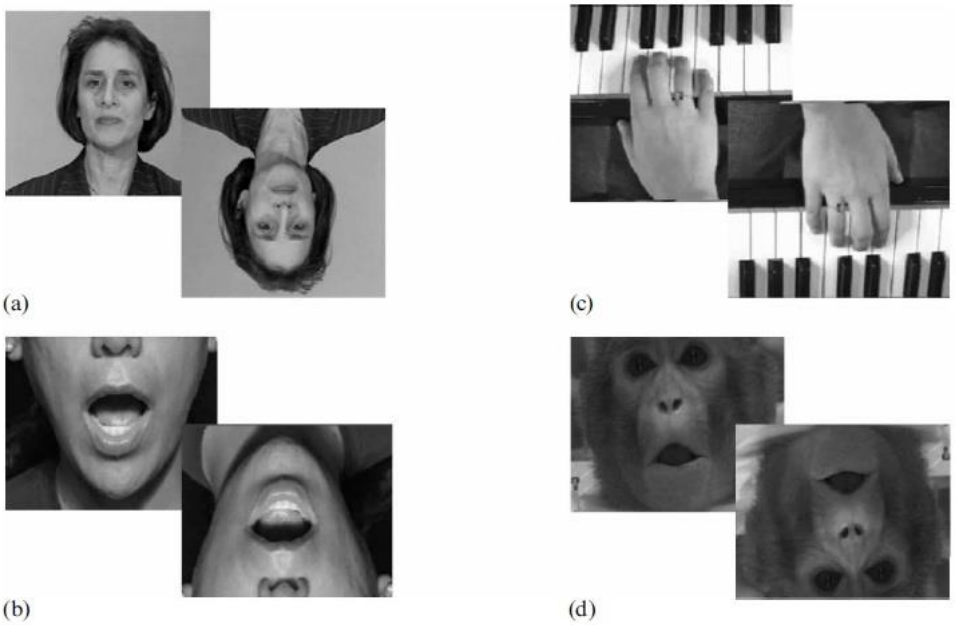

Figure 20. Représentation du matériel visuel utilisé dans les différentes conditions expérimentales par Vatakis et Spence (2008).

L'image représente les quatre conditions expérimentales présentant les différents types de stimuli visuels relatifs à la production de différents types de stimuli auditifs. En (a) et (b) il s'agit de stimuli linguistiques, en (c) de stimuli musicaux et en (d) des cris d'animaux. Chaque vidéo a été présentée de façon droite et de façon inversée. (L'image de Vatakis et Spence (2008).)

3.3.1.2 Illusion de Thatcher

Un autre phénomène démonstrateur de l'importance du traitement holistique des visages est l'illusion de Thatcher qui a été créée en inversant les composantes faciales, notamment la bouche et les yeux, initialement dans une photographie de Margaret Thatcher (Thompson, 1980). Une telle manipulation, qui interrompt les relations spatiales entre les composantes faciales et consécutivement le traitement holistique du visage, donne lieu à un visage grotesque. Toutefois, cet effet est beaucoup moins évident si un visage thatcherisé est inversé. Dans la mesure où l'inversion faciale perturbe également le traitement holistique du visage, l'aspect presque normal d'un visage thatcherisé inversé est considéré comme confirmant la supposition selon laquelle l'information sur la configuration du visage a un moindre poids/est traitée moins efficacement si ce dernier a subi une inversion verticale (e.g., Barlett & Searcy, 1993 ; Bruce & Young, 2012 ; Carbon, Schweinberger, Kaufmann, & Leder, 2005).

L'étude princeps ayant eu recours à l'illusion de Thatcher pour étudier la nature du traitement de l'information faciale dans le cadre de la perception de la parole est celle de Rosenblum et *al*. (2000). Ayant utilisé le paradigme de McGurk-MacDonald, Rosenblum et *al*. (2000) ont montré que la thatcherisation du visage de l'orateur pouvait grandement affecter l'occurrence de l'effet McGurk-MacDonald chez les sujets, mais uniquement dans les conditions où le visage thatcherisé n'a pas subi de rotation. Ces résultats, qui ont été répliqués très récemment par Eskelund, MacDonald et Andersen (2015), suggèrent ainsi que l'information sur la configuration du visage joue un rôle dans le traitement des stimuli visuels de la parole. En conséquence, le traitement de l'information visuelle, quand celle-ci est présentée dans un contexte du visage entier, semble impliquer le traitement holistique. Une telle conclusion est confirmée davantage par le fait que le format présentant la bouche seule de l'orateur, et donc ne comportant aucune information sur la configuration de son visage, n'affectait pas l'occurrence de l'effet McGurk-MacDonald (Rosenblum et *al*., 2000). (Pour des exemples des stimuli utilisés dans l'étude de Rosenblum et *al*. (2000), voir la Figure 21 et la Figure 22). Par ailleurs, le fait que l'inversion faciale simple n'affecte pas la perception des stimuli McGurk-MacDonald, suggère que le rôle du traitement holistique dans le cadre de la perception bimodale de la parole est quelque peu limité, ou en tout cas moindre que dans le cadre de la perception de l'identité ou des expressions faciales. Néanmoins, sur ce point, les résultats de l'étude d'Eskelund et *al*. (2015) ne sont pas concordants avec ceux de Rosenblum et *al*. (2001). En effet, Eskelund et *al*. (2015) ont observé une diminution de l'effet McGurk-MacDonald dans les conditions où le visage d l'orateur était inversé. Aussi, si la perception audio-visuelle de la parole semble impliquer une certaine dimension du traitement holistique en présence de l'information faciale intégrale de l'orateur, le rôle exact de ce type de traitement dans ce contexte précis n'est pas clairement établi par les résultats des études qui ont eu recours au paradigme de l'illusion de Tatcher[32].

[32] Notons que l'implication du traitement holistique dans la perception audio-visuelle de la parole n'est pas clairement établie tout court. Les résultats des études portant sur ce sujet sont relativement divergents et ne peuvent que suggérer une possibilité que l'information faciale soit traitée de façon holistique lorsque notre tâche est d'identifier les caractéristiques acoustiques de la production verbale orale (voir l'ensemble de la section 3.3.1 du présent chapitre).

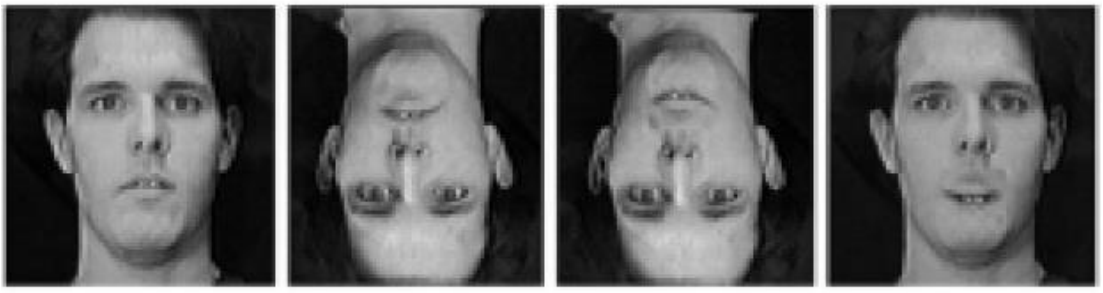

Figure 21. Représentation des stimuli visuels comportant l'information faciale intégrale de l'étude de Rosenblum et *al*. (2000).

De gauche vers la droite, on voit successivement des images statiques des vidéos comportant (i) à l'extrême gauche, le visage normalement orienté de l'orateur, (ii) au centre gauche, le visage inversé de l'orateur, (iii) au centre droit, le visage tatcherisé et inversé de l'orateur, et (iv) à l'extrême droite, le visage tatcherisé de l'orateur orienté normalement. (L'image de Rosenblum et *al*. (2000).)

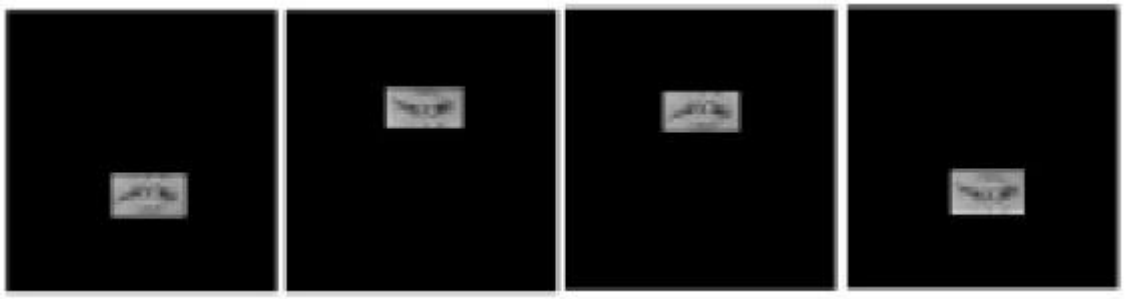

Figure 22. Représentation des stimuli visuels comportant l'information faciale réduite à la bouche seule de l'orateur de l'étude de Rosenblum et *al*. (2000).

De gauche vers la droite, on voit successivement des images statiques des vidéos comportant (i) à l'extrême gauche, la bouche du visage normalement orienté de l'orateur, (ii) au centre gauche, la bouche du visage inversé de l'orateur, (iii) au centre droit, la bouche du visage tatcherisé et inversé de l'orateur, et (iv) à l'extrême droite, la bouche visage tatcherisé de l'orateur orienté normalement. Les images de la bouche correspondent aux stimuli faciaux présentés dans la Figure 21. (L'image de Rosenblum et *al*. (2000).)

3.3.1.3 Effet des visages à composantes mélangées

Une autre manière de manipuler l'aspect holistique d'un visage en affectant la configuration de ses composantes consiste à mélanger les composantes faciales, c'est-à-dire à les organiser spatialement d'une façon qui ne soit pas typique d'un visage humain. Les stimuli ainsi établis portent le nom des visages à composantes mélangées (*scrambled faces*). Conformément aux exemples précédents, la dénaturation des relations spatiales entre les composantes faciales dans les visages à composantes mélangées perturberait le traitement holistique, ce qui, à terme, provoque des difficultés dans la perception de tels visages (e.g., Donnelly, Humphreys, & Sawyer, 1994 ; George, Evans, Fiori, Davidoff, & Renault, 1996 ; Taubert, Agten-Murphy, & Parre, 2012).

Les visages à composantes mélangées représentent ainsi un autre type de stimuli pouvant être utilisés dans la recherche sur la nature du traitement de l'information faciale lors de la perception audio-visuelle de la parole. Aussi, Hietanen, Manninen, Sams et Suraka (2001) ont élaboré 4 formats de visages à composantes mélangées, différant entre eux dans la disposition spatiale des composantes, ainsi que deux formats dans lesquels certaines composantes étaient isolées du contexte facial, mais leur organisation spatiale n'était pas manipulée. Ayant utilisé le paradigme de McGurk-MacDonald, les auteurs ont rapporté que tous les formats visuels avaient produit l'effet McGurk-MacDonald. Toutefois, ce-dernier a été sensiblement moins fréquent dans la condition du format de visage à composantes mélangées avec une organisation non symétriques des composantes, qui était le format présentant la plus forte violation de la configuration faciale (pour des exemples des stimuli utilisés par Hietanen et *al*. (2001), voir la Figure 23).

A la base de tels résultats, Hietanen et *al*. (2001) ont conclu que l'information sur la configuration du visage, et par conséquent le traitement holistique, peuvent être impliqués, dans une certaine mesure, dans l'intégration audio-visuelle lors de la perception de la parole. Toutefois, ce type de traitement n'y semble pas jouer un rôle déterminant. Notons finalement que les résultats de l'étude de Hietanen et *al*. (2001) ont également montré que la fréquence d'occurrence de l'effet McGurk-MacDonald variait en fonction de l'orateur et en fonction des consonnes utilisées dans les combinaisons audio-visuellement incongruentes des stimuli McGurk-MacDonald. Ce dernier point est conforme aux observations de Rosenblum et *al*. (2000). Il suggère qu'outre les indices de la composante buccale seule et de son aspect cinématique, d'autres types d'indices faciaux pourraient avoir un rôle dans l'intégration audio-visuelle de certaines consonnes. Il semble possible que ces indices soient en lien avec la

configuration du visage de l'orateur. Toutefois, il pourrait également s'agir d'indices apportés par les mouvements dans les régions extra-buccales dont la cohérence avec les mouvements de la bouche est inexistante ou fortement perturbé dans les visages à composantes mélangées ou encore les visages tatcherisés.

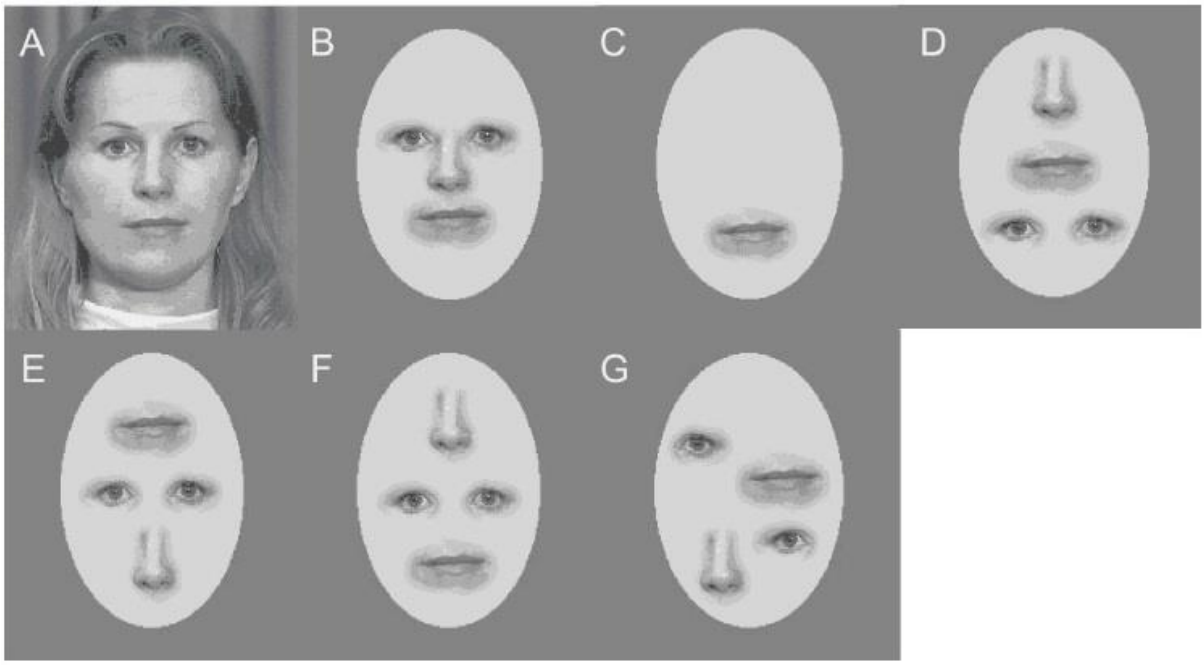

Figure 23. Exemple du matériel visuel utilisé dans les différentes conditions de présentation audio-visuelle par Hietanen et *al*. (2001).

En A, le visage intégral d'un des orateurs de l'étude de Hietanen et *al*. (2001). En B et en C, les formats visuels comportant certaines composantes faciales dépourvues du contexte facial mais non mélangées. En D, E, F et G les formats visuels du visage de l'orateur dans lequel les composantes faciales ont été mélangées. Ils diffèrent entre eux dans l'organisation spatiale des composantes. L'occurrence de l'effet de McGurk-MacDonald était diminuée avec le format G, présentant une forte violation de l'organisation spatiale des composantes faciale car biaisant le principe de symétrie. (L'image de Hietanen et *al*. (2001).)

3.3.1.4 Effet de supériorité du tout (visage) sur une partie (bouche)

Dans le contexte de la perception des visages, l'effet de supériorité du tout (visage) sur une partie désigne le fait qu'une composante faciale est traitée plus efficacement/reconnue plus facilement dans le contexte du visage entier que présentée de façon isolée (Tanaka & Farah,

1993). Dans la mesure où le fait de présenter une composante dans le contexte du visage permet au traitement holistique d'opérer, alors que ceci n'est pas le cas dans la condition où la composante est présentée de manière isolée, l'effet de supériorité du tout (visage) sur une partie est également considéré comme un des marqueurs du traitement holistique des visages.

Dans le cadre de la perception audio-visuelle de la parole, de nombreuses études ont comparé l'efficacité de la vision de la bouche seule à celle du visage entier (Berger et *al*., 1971 ; Greenberg & Bode, 1968 ; Hietanen et *al*., 2001 ; Ijsseldijk, 1992 ; Marassa & Lansing, 1995 ; Stone, 1957 ; Thomas & Jordan, 2004 ; Rosenblum et *al*., 2000) (revoir aussi la section 3.2.2 de ce chapitre). A l'exception de Hietanen et *al*. (2001) et Thomas et Jordan (2004), les auteurs de ces études n'avaient pas pour objectif explicit d'étudier le lien entre le traitement holistique des visages et le traitement de l'information faciale dans la perception audio-visuelle d la parole. Toutefois, dans la mesure où leur démarche méthodologique est conforme au paradigme partie *vs* tout, les résultats de ces études peuvent apporter un éclairage supplémentaire sur la façon dont l'information visuelle est traitée lors de la perception audio-visuelle de la parole. Dans l'absolu, ces études devraient également permettre de répondre à la question concernant la nature de l'information visuelle qui est utilisée le plus efficacement pour la perception de la parole.

La comparaison entre l'efficacité de la bouche seule et celle du visage entier dans le cadre de la perception de la parole a abouti à des résultats divergents. En effet, Ijsseldijk (1992) a trouvé un moindre effet facilitateur de la perception de la parole induit par un input audio-visuel avec la composante visuelle « bouche » qu'avec la composante visuelle « visage entier ». Un tel résultat est ainsi conforme aux observations faites dans le cadre de la majorité des études ayant exploré la présence d'autres marqueurs du traitement holistique de l'information faciale lors de la perception audio-visuelle de la parole. En revanche, Hietanen et *al*. (2001) ont observé le pattern inverse avec l'information réduite à la composante buccale seule qui facilitait l'occurrence de l'effet McGurk-MacDonad davantage que le visage entier de l'orateur. Une tendance, pourtant non significative, vers une supériorité du format « bouche » comparativement au format du visage entier a également été observé par Thomas et Jordan (2004). Dans cette dernière étude qui a utilisé le paradigme de la dégradation de l'information auditive par le bruit, un nombre très réduit d'items aurait pu accentuer l'effet d'apprentissage dans un design expérimental à mesure répétées et noyer ainsi un effet du type de format visuel sur le gain audio-visuel. Par ailleurs, le taux de reconnaissance correcte d'items expérimentaux révélait un pattern très proche de l'effet plafond. Ceci appuie davantage l'hypothèse selon

laquelle la bouche aurait pu s'avérer plus efficace que le visage entier dans le contexte de l'étude de Thomas et Jordan (2004).

Les autres études n'ont pas trouvé des différences entre le format « visage » et le format « bouche » dans la perception audio-visuelle de la parole. Toutefois, comme le soulignent Thomas et Jordan (2004), la région buccale dans ces études était relativement variable, comportant souvent d'autres régions (larynx, mâchoire, joues) dont les mouvements auraient pu également contribuer à la perception de la parole. Par ailleurs, la pertinence de l'utilisation du procédé de masquage pour établir le format visuel « bouche », pourtant très fréquente, est contestée par Thomas et Jordan (2004). D'une part, l'utilisation de régions noires qui masquent le visage de l'orateur introduit d'importants contrastes lumineux dans les stimuli visuels. Selon Thomas et Jordan (2004), ces derniers, pourraient jouer sur l'orientation de l'attention visuelle en déviant le regard de la région exposée vers les régions noires[33]. D'autre part, l'interprétation des stimuli visuels dans lesquels un élément visuel est entouré d'autres régions/éléments est ambiguë (Cavedon, 1980 ; Nelson & Palmer, 2001), ce qui pourrait également impacter la perception audio-visuelle de la parole.

Dans son ensemble, le paradigme de tout *vs* partie pourrait présenter un avantage sur les autres approches qui ont été utilisées pour étudier la nature de la perception audio-visuelle de la parole pour plusieurs raisons. Premièrement, en dissociant la partie (la bouche) du tout (le visage), il n'introduit pas de perturbations au niveau du traitement holistique potentiellement déclenché lors de la perception de la parole, mais supprime la possibilité d'un tel traitement avec un format visuel centré sur la bouche seule. Deuxièmement, il n'affecte ni l'aspect écologique de l'information visuelle ni la cohérence entre les mouvements buccaux et les mouvements des régions extra-buccales pouvant également contribuer à la perception de la parole. Troisièmement, avec les manipulations du mouvement dans les régions extra-buccales et dans la région buccale, il permet d'étudier le rôle respectif du contexte facial et du mouvement des régions extra-buccales dans la perception audio-visuelle de la parole. Néanmoins, à l'exception de l'étude de Thomas et Jordan (2004), les études précédentes n'ont pas exploité tous les potentiels du paradigme tout vs partie dans l'étude audio-visuelle de la parole. Par ailleurs, les différences et limites méthodologiques des études précédentes, telles qu'exposées plus haut, empêchent la mise en place de conclusions valides.

[33] Notons cependant qu'il s'agit ici d'une supposition émise par Marassa et Lansing (1995) qui n'a cependant pas été validée empiriquement.

3.3.2 Corrélats neuronaux

Quelques éléments pertinents quant à la question de la nature du traitement des indices acoustiques visuels portés par le visage de l'orateur lors de la perception audio-visuelle de la parole peuvent être trouvés dans les études ayant exploré l'implication des différentes régions cérébrales dans le cadre en question. Certaines études de ce domaine ont rapporté l'activation du gyrus fusiforme droit lors de la perception bimodale de la parole (e.g., Capek, Bavelier, Corina, Newman, Jezzard, & Neville, 2004 ; Dick et *al*., 2010 ; Kawase, Yamaguchi, Ogawa, Suzuki, Suzuki, Itoh, … Fujii, 2005). Une partie du gyrus fusiforme droit étant la région cérébrale particulièrement spécialisée dans le traitement holistique des visages (Furl, Garrido, Dolan, Driver, & Duchain, 2011 ; Kanwisher, McDermont, & Chun, 1997 ; Kanwisher & Yovel, 2006), cette donnée empirique suggère que la perception audio-visuelle de la parole implique en effet ce type de traitement, au moins dans une certaine mesure. Toutefois, cette supposition est contredite par la plupart des études s'intéressant aux aspects cérébraux fonctionnels lors de la perception bimodale de la parole qui n'ont pas rapporté d'activation du gyrus fusiforme droit dans un contexte pareil[34]. Dans cette même lignée d'éléments empiriques, certains rapports neuropsychologiques des patients prosopagnosiques, étant atteints uniquement dans le traitement des visages, et des patients alexiques, présentant une atteinte limitée au seul traitement de l'information visuelle de la parole produite oralement, indiquent l'existence d'une double dissociation entre le traitement des visages et le traitement de la parole dans sa modalité visuelle (Campbell, Landis, & Regard, 1986). Cette double dissociation signerait une indépendance des deux types de traitements. En revanche, certains autres patients prosopagnosiques étaient décrits comme présentant aussi bien des déficits du traitement des visages que du traitement des indices visuels de la parole portés par les visages (Campbell, 1992 ; de Gelder & Vroomer, 1998). Les symptômes observés chez les patients atteints de lésions cérébrales sont cependant un élément empirique délicat à interpréter. En effet, les lésions peuvent être très variables dans leur étendue corticale et sous-corticale. Néanmoins, le fait qu'une double dissociation puisse être observée indique une forte probabilité que le traitement de l'information visuelle lors de la perception bimodale de la parole n'implique pas nécessairement le traitement holistique du visage de l'orateur.

Hormis l'approche de l'imagerie cérébrale fonctionnelle et des rapports des symptômes neuropsychologiques, une étude a abordé la question de la nature du traitement de l'information

[34] Notons toutefois que les détails concernant le format de l'information visuelle (bouche ou visage) ne sont le plus souvent pas rapportés dans ces études.

faciale lors de la perception audio-visuelle de la parole au moyen de l'approche des potentiels évoqués. En effet, l'étude d'Eskelund et *al*. (2015), déjà mentionnée dans la section 3.3.1.2 du présent chapitre, a également exploré la réponse cérébrale face à des stimuli de type McGurk-MacDonald dont la composante visuelle avait subi une tatcherisation et/ou une inversion faciale. Plus précisément, Eskelund et *al*. (2015) ont pris en compte la négativité de discordance, l'onde qui traduit un changement dans un stimulus ou plus précisément dans la perception d'un stimulus. Aussi, dans le contexte du paradigme de McGurk-MacDonald, la négativité de discordance est utilisée pour détecter le changement dans la perception d'une syllabe, induite par la fusion illusoire entre un stimulus auditif et un stimulus visuel qui lui est incogruent, alors que le stimulus auditif n'a pas changé. Eskelund et *al*. (2015) ont rapporté que la tatcherisation du visage de l'orateur, qui provoquait une forte baisse dans l'effet McGurk-MacDonald (la composante comportementale) était également accompagnée d'une absence de la négativité de discordance. Un tel résultat est conforme à l'hypothèse selon laquelle la perception audio-visuelle de la parole implique, au moins dans une certaine mesure, le traitement holistique de l'information visuelle. En revanche, les résultats pour les conditions d'inversion faciale sont étonnants. En effet, si l'effet McGurk-MacDonald était plus faible (mais toujours présent) avec ce type de stimuli, il n'était pas accompagné de la négativité de discordance. Les auteurs expliquent ce phénomène en suggérant que la négativité de discordance nécessite un certain seuil en termes de grandeur dans la différence perçue dans le cadre du paradigme McGurk-MacDonald. En effet, le stimulus de type McGurk-MacDonald utilisé dans l'expérience d'Eskelund et *al*. (2015) consistait en /ba/ auditif et /va/ visuel, /b/ et /v/ étant acoustiquement assez proches. L'explication d'Eskelund et *al*. (2015) semble ainsi plausible, mais reste à être confirmée empiriquement. Globalement, les résultats de l'expérience d'Eskelund et *al*. (2015) apportent quelques éléments à l'appui de l'hypothèse selon laquelle le traitement de l'information visuelle dans la perception bimodale de la parole comporte une dimension holistique.

Finalement, les résultats des études s'intéressant à l'asymétrie fonctionnelle inter-hémisphérique dans la perception visuelle de la parole peuvent également apporter un éclairage quant à la problématique concernant la nature du traitement de l'information visuelle dans la perception audio-visuelle de la parole. En effet, Jordan, Sheen, Abedipour et Paterson (2014) ainsi que Jordan et Thomas (2007) ont procédé à une présentation dichotomique[35] du visage de

[35] La présentation dichotomique des items visuels consiste à les projeter à plus de 2° à gauche ou à droite par rapport au point de fixation du sujet (procédure de Jordan & Thomas, 2007). Aussi, les items sont présentés dans

l'orateur prononçant des syllabes respectivement dans l'hémichamp visuel gauche et droit des participants. Les stimuli étaient présentés dans leur modalité visuelle seule. Les résultats des deux études sont concordants et montrent un avantage de l'hémichamp visuel droit, et donc de l'hemisphère cérébral gauche, pour la reconnaissance des stimuli expérimentaux. L'hémisphère cérébral gauche étant spécialisé dans le traitement analytique, de tels résultats suggèrent que l'approche optimale à la perception des unités phonologiques de la parole à partir de la modalité visuelle de la parole n'implique pas le traitement holistique de l'information visuelle.

Pris dans leur ensemble, les éléments empiriques en lien avec l'organisation cérébrale structurelle et fonctionnelle du traitement de l'information visuelle dans le cadre de la perception de la parole sont relativement peu nombreux et, à l'image des résultats des études comportementales, assez divergents. Aussi, même si certains éléments suggèrent une possible implication du traitement holistique dans ce cadre précis, l'état actuel de recherche sur ce sujet ne permet pas d'aboutir à des conclusions valides. Par ailleurs, même si le traitement holistique de l'information visuelle intervient dans la perception audio-visuelle de la parole, il n'est pas certain que ce soit la stratégie optimale pour traiter la parole dans son seul aspect phonologique.

3.4 Comportement oculaire dans la perception audio-visuelle de la parole

Certaines études dans le domaine de la perception audio-visuelle de la parole ont exploré la façon dont l'information visuelle, faciale, est traitée dans ce cadre en prenant en compte le comportement oculaire de la personne percevant le message. L'aspect du comportement oculaire le plus souvent analysé dans ces études sont les fixations oculaires dans les différentes régions faciales. Dans la mesure où les fixations oculaires dans une certaine région du stimulus visuel sont censées généralement indiquer les moments durant lesquels s'opère l'extraction de l'information visuelle de cette région (pour plus de détails, voir Liversedge, Gilchrist, & Everling, 2013), ce type d'études nous permet d'apprécier la façon de collecter l'information visuelle ainsi que les régions faciales les plus fortement impliquées dans la perception audio-visuelle de la parole.

---

un seul hémichamp visuel, gauche ou droit. Conformément à l'organisation structurelle du nerf optique, la présentation dichotomique permet que l'input sensoriel de la rétine soit projeté dans un seul hémisphère cérébral, gauche pour les projections dans l'hémichamp visuel droit et droit pour les projections dans l'hémichamp visuel gauche.

Les résultats des études ayant exploré le comportement oculaire durant la perception audio-visuelle de la parole présentent un degré de concordance relativement élevé. En effet, toutes les études de ce domaine montrent que la perception audio-visuelle de la parole des stimuli linguistiques normalement audibles s'accompagne des fixations oculaires situées principalement dans la région buccale et également dans la région des yeux (Buchan, Paré, & Munhall, 2008 ; Everdell, Marsh, Yurick, Munhall, & Paré, 2007 ; Lansing & McConkie, 2003 ; Paré, Richler, Ten Hove, & Munhall, 2003 ; Vatikiotis-Bateson, Eigsti, Yano, & Munhall, 1998). Ce pattern des fixations oculaires semble varier en fonction de l'acuité de l'input auditif. En effet, dans les conditions où l'input auditif est dégradé, Buchan et *al*. (2008) ont trouvé que la durée des fixations de la région des yeux baisse, alors la durée des fixations du nez augmente (pour les exemples des stimuli visuels et des régions d'intérêt (*areas of interest* (*AOI*)) de l'étude de Buchan et *al*. (2008), voir la Figure 24). Étonnamment, dans l'étude de Buchan et *al*. (2008), la durée des fixations de la région buccale a été également inférieure dans le contexte de dégradation de l'input auditif comparativement au contexte où l'input auditif était normalement audible. Ce résultat diffère de celui trouvé par Vatikotis-Bateson et *al*. (1998). Ces auteurs ont notamment rapporté que l'ajout du bruit au signal acoustique s'accompagnait d'une augmentation de la durée des fixations dans la région buccale de l'orateur. Dans la même veine de problématique, Lansing et McConkie (2003) ont étudié les patterns des fixations oculaires durant la perception visuelle et audio-visuelle des phrases. Les résultats de cette étude ont montré que le regard de la personne percevant les phrases était orienté essentiellement vers les yeux de l'orateur durant les pauses vocales, et vers la bouche de l'orateur durant la production vocale. Ce changement dynamique dans le pattern des fixations oculaires dans la région des yeux et la région buccale apparaît avant le début du voisement (mouvement du regard des yeux vers la bouche) et après la fin du voisement (mouvement du regard de la bouche vers les yeux de l'orateur). Conformément aux observations de Vatikiotis-Bateson et *al*. (1998), Lansing et McConkie (2003) ont trouvé que le degré auquel le regard était attiré par la bouche de l'orateur était plus élevé dans les conditions de haute incertitude quant à l'input auditif.

Un autre facteur affectant le pattern des fixations oculaires dans le visage de l'orateur lors de la perception audio-visuelle de la parole est la stabilité de l'identité de l'orateur. Buchan et *al*. (2008) ont effectivement rapporté que si le pattern de collection d'informations à partir du visage de l'orateur était relativement robuste, le fait de varier l'identité de l'orateur augmentait faiblement mais significativement la durée des fixations oculaires dans la région

buccale. Everdell et *al*. (2007) ont mis en évidence une autre caractéristique du comportement oculaire lors de la perception audio-visuelle de la parole normalement audible. Il s'agit de la nature asymétrique du pattern des fixations oculaires dans ce cadre, avec une légère tendance à fixer le côté droit du visage de l'orateur, essentiellement au niveau de la région des yeux. Une tendance similaire, même si non significative, a été observée également par Buchan et *al*. (2008). Cette caractéristique a été observée uniquement avec les visages dynamiques, articulant des items linguistiques, mais pas avec les visages statiques et les objets (Everdell et *al*., 2007). Finalement, Paré et *al*. (2003) ont exploré l'effet de la manipulation de l'orientation du regard de la personne percevant un message verbal oral sur l'occurrence de l'effet McGurk-MacDonald. Les résultats ont montré que la déviation dans les fixations oculaires par rapport au visage de l'orateur devait être supérieure à 10° à 20° pour affecter l'effet en question de façon significative. Par ailleurs, Paré et *al*. (2003) ont trouvé que l'effet de McGurk-MacDonald était supprimé dans les conditions où les fixations oculaires étaient déviées de plus de 60° du visage de l'orateur.

Globalement, les résultats des études ayant exploré les stratégies visuelles du traitement de l'information faciale lors de la perception audio-visuelle de la parole montrent que ce type de perception engage essentiellement et presque exclusivement la région buccale, la région des yeux et la région du nez. Les patterns des fixations oculaires dans ce cadre apparaissent ainsi comme assez robustes. Finalement, les résultats de l'étude de Paré et *al*. (2003) mettent également en évidence le fait que la perception audio-visuelle de la parole est également opérationnelle dans les cas où l'attention visuelle est légèrement déviée du visage de l'orateur. Un tel résultat prouve que la parole visuelle peut être encodée par vision périphérique ainsi que le caractère hautement performant des mécanismes de la perception audio-visuelle de la parole chez l'humain.

Figure 24. Exemple des stimuli visuels avec les régions d'intérêt de l'étude de Buchan et *al*. (2008).

L'image représente des stimuli visuels comportant des orateurs différents. Les régions d'intérêt consistaient en la région buccale de l'orateur, et de son œil gauche et droit. (L'image de Buchan et *al*. (2008).)

Résumé du chapitre

La recherche centrée autour des différentes caractéristiques de l'input visuel et leur impact sur la perception audio-visuelle de la parole a abouti à des conclusions qui, au vu des résultats des études, présentent des degrés de validité variables. En effet, d'une part un accord relativement élevé dans les résultats des études portant sur certaines problématiques de ce domaine suggère un degré relativement élevé de validité des conclusions. Ainsi, la perception audio-visuelle de la parole paraît être un phénomène robuste, résistant bien à des manipulations de l'input visuel dans les dimensions telles que la taille, la couleur, l'acuité, la luminance, l'angle de vue et l'identité de l'orateur. L'humain semble être très performant dans le traitement des indices visuels relatifs aux mouvements des articulateurs lors de la perception de la parole, pouvant profiter des indices minimes, réduits aux seuls mouvements des lèvres. La région buccale apportant les indices relatifs aux mouvements, la forme et la position des lèvres, de la langue et des dents a été identifiée comme l'élément visuel crucial dans la perception de la parole. Cette région attire le plus fortement le regard de la personne percevant la parole, même

si certaines autres régions faciales sont également fixées/traitées lors dans ce cadre précis. Il s'agit notamment des yeux et du nez de l'orateur. Le comportement oculaire lors de la perception bimodale de la parole semble varier aussi bien en fonction de l'acuité de l'input auditif qu'en fonction des moments de voisement et les moments de pauses dans une suite continue de la parole. Plus précisément, une baisse dans l'acuité de l'input visuel s'accompagne d'une baisse de la durée des fixations oculaires au niveau des yeux et une hausse de la durée des fixations oculaires au niveau du nez. Quant aux transitions entre les périodes de voisement et celles de pauses, le regard de la personne percevant la parole semble se diriger vers la bouche peu avant le début de voisement et quitter cette région pour s'orienter vers les yeux de l'orateur peu après la fin de voisement. Finalement, la perception bimodale de la parole pourrait également profiter des indices visuels provenant des articulateurs invisibles, notamment de la langue. Toutefois, un certain degré d'expérience perceptive/d'entraînement au traitement de ce type d'indices semble nécessaire pour tirer un bénéfice dépassant potentiellement celui obtenu par la vision des articulateurs visibles.

Contrairement aux études qui ont produit des résultats d'un degré de concordance relativement élevé dont les conclusions sont présentées ci-dessus, la recherche sur un éventuel apport supplémentaire de la vision du visage entier comparativement à la bouche seule et sur une éventuelle implication du traitement holistique de l'input visuel dans la perception bimodale de la parole a produit des résultats relativement divergents. De tels résultats ne peuvent pas aboutir à un consensus clair quant à la conclusion concernant ces deux problématiques. Aussi, la plupart des études comparant l'efficacité de la vision de la bouche seule à celle de la vision du visage entier de l'orateur dans la perception bimodale de la parole n'a trouvé aucune différence entre les deux formats visuels. Toutefois, la mise en place de l'input « bouche » étant problématique dans ces études (non restriction du format à la seule région buccale et utilisation de masques visuels pouvant affecter le processus perceptif ; revoir les sections 3.2.2 et 3.3.1.4 pour plus de détails), la validité de ces résultats semble problématique. Par ailleurs, l'étude d'Ijsseldijk (1992) a montré une supériorité du visage entier comparativement à la composante faciale buccale seule, sans pour autant donner de détails sur les formats visuels utilisés dans l'expérience. D'autre part, deux études ayant réduit le format « bouche » strictement à la composante buccale sans y introduire de régions opaques masquant le visage de l'orateur montrent une tendance, pas toujours significative, d'une supériorité de la composante buccale seule sur le format comportant le visage entier (Hietanen et *al.*, 2001 ; Thomas & Jordan, 2004). Quant aux études portant sur une éventuelle implication du traitement

holistique du visage de l'orateur dans la perception audio-visuelle de la parole, les résultats des études comportementales suggèrent que ce type de perception implique en effet un certain degré de traitement holistique de l'information faciale. En revanche, la quasi-totalité de ces études ne répond pas à la question de la façon optimale de traitement de l'information visuelle pour la perception de la parole dans ses seuls aspects phonologiques. En effet, le traitement holistique du visage semble être hautement automatisé, se déclenchant au stade de la détection d'un visage (Taubert et *al*., 2011). De ce fait, les marqueurs du traitement holistique des visages qui ont été observés dans le cadre de la perception bimodale de la parole pourraient s'expliquer par le fait que, dans ce cadre précis, l'information faciale déclenche deux types de traitement qui sont menés de façon parallèle. Plus précisément, il pourrait s'agir de lancer un traitement holistique, ayant pour objectif d'identifier l'identité de l'individu et ses expressions faciales, ainsi qu'un traitement analytique visant à identifier les unités phonologiques produites. Une telle approche serait certainement avantageuse dans un cadre de conversation quotidienne où notre objectif premier est de comprendre le sens du message parlé qui dépend fortement des facteurs contextuels (dimension pragmatique) que des indices supra-segmentaux (dimension prosodique). En revanche, si la tâche de l'individu est de traiter la parole dans son aspect purement phonologique, l'homme pourrait bénéficier davantage d'un traitement analytique, centré sur les indices visuels de la région buccale seule, comme semblent suggérer les résultats de Hietanen et *al*. (2001) et de Thomas et Jordan (2004). Certains éléments empiriques sur les corrélats neuronaux du traitement de l'input visuel lors de la perception de la parole vont dans le sens de cette hypothèse. Il s'agit, par exemple, de l'absence du recrutement systématique du gyrus fusiforme droit dans la perception bimodale de la parole ainsi que d'une possible supériorité de l'hémisphère cérébral gauche, spécialisé dans le traitement analytique, dans le traitement de la parole dans son aspect phonologique pur (revoir la section 3.3.2). En somme, d'autres études sont nécessaires pour élucider la question du format visuel permettant une extraction optimale d'indices acoustiques et celle de l'implication du traitement holistique de l'input visuel dans la perception bimodale de la parole.

4 Questions de recherche et hypothèses

4.1 Contribution de l'information visuelle relative à la région buccale, aux régions extra-buccales et au contexte facial à la perception de la parole

L'objectif du présent travail de thèse est de contribuer à la recherche sur les deux problématiques exposées ci-dessus, à la fin du chapitre précédent. Plus précisément, il s'agit d'exploiter au mieux le paradigme du tout (visage) *vs* partie (bouche), à l'image de ce qui a été fait précédemment par Thomas et Jordan (2004) (revoir le dernier paragraphe de la section 3.2.1.4), pour répondre aux questions suivantes : (i) Quelle est la contribution respective de la région buccale et du visage entier de l'orateur à la perception audio-visuelle de la parole ? (ii) Quelle est la contribution des régions extra-buccales à la perception audio-visuelle de la parole ? (iii) Quelle est la contribution du contexte facial à la perception audio-visuelle de la parole.

Ces questions d'étude s'accompagnent des hypothèses suivantes : (i) Si la perception audio-visuelle de la parole implique un traitement holistique du visage de l'orateur qui, à terme, profite à l'extraction d'indices acoustiques à partir de l'input visuel, alors (i) la contribution du visage entier de l'orateur devrait être plus importante que celle de la bouche seule. Par ailleurs, (ii) l'ajout d'un contexte facial, même statique, à la région buccale de l'orateur devrait permettre une meilleure performance dans le traitement audio-visuel de la parole que la région buccale seule. En revanche, si la supériorité du visage entier par rapport à la région buccale seule a été constatée en raison d'un supplément d'indices apportés par la vision des mouvements dans les régions extra-buccales dans le format « visage » (Ijsseldijk, 1992), alors (i) on devrait pouvoir observer une contribution significative de la vision des seuls mouvements dans les régions extra-buccales à la perception de la parole, et (ii) l'apport d'un format comportant un visage entièrement actif à la perception de la parole devrait être plus important que celui d'un format comportant le visage entier de l'orateur dans lequel les mouvements dans les régions extra-buccales ont été supprimés (le contexte facial a été rendu statique). Finalement, en lien avec les résultats de Hietanen et *al*. (2001) ainsi que ceux de Jordan et *al*. (2014), il semble également possible de supposer que le traitement de l'aspect acoustique de la parole se fait davantage de façon analytique et que la présentation du visage entier de l'orateur pourrait être désavantageuse pour ce type de traitement. Ainsi, notre troisième variante des hypothèses en lien avec la problématique de la nature du traitement de l'information visuelle lors de la perception bimodale de la parole est la suivante : (i) La vision de la bouche seule devrait être plus efficace pour la perception audio-visuelle de la parole que la vision du visage entier. (ii) La simple

présence d'un contexte facial, même statique, devrait s'avérer désavantageuse pour ce type de perception.

4.2 Caractéristiques du format « bouche »

Thomas et Jordan (2004) ayant exposé le possible problème de l'utilisation de régions obscures dans l'établissement du format visuel centré sur la région buccale de l'orateur, notre quatrième question d'étude est de savoir si leur supposition peut être considérée comme valide. En d'autres termes, la question (iv) de notre étude est de savoir si un format visuel « buche », comportant des régions obscures qui masquent le visage de l'orateur, est moins efficace pour la perception de la parole qu'un format « bouche » ne comportant pas ces régions. Au sujet de cette question, nous émettons l'hypothèse que le format visuel « bouche » comportant des régions obscures donnera lieu à de moins bonnes performances dans la perception audio-visuelle de la parole que le format sans régions obscures. Une des raisons, avancées par Thomas et Jordan (2004), pour lesquelles le format visuel « bouche » comportant des régions obscures ne serait pas aussi efficace que le format « bouche » sans ces régions est que les régions opaques attireraient le regard de la personne percevant la parole.

Thomas et Jordan (2004) n'ayant pas testé leur hypothèse, notre cinquième et sixième questions de recherche sont de savoir comment les caractéristiques de bas niveau des stimuli visuels réduisant l'information faciale à la seule région buccale affectent (v) le traitement visuel de l'information faciale et (vi) la perception audio-visuelle de la parole. Dans la lignée des suppositions avancées par Thomas et Jordan (2004), notre hypothèse (v) est ainsi que l'application d'un masque opaque, couvrant le visage de l'orateur tout en exposant la bouche, attirera le regard de la personne percevant la parole de façon bimodale. Plus précisément, nous pensons que les fixations oculaires les participants auront une moindre tendance à fixer la région buccale avec un format « bouche » comportant les régions opaques qu'avec un format « bouche » sans régions opaques. Notre hypothèse (vi) est que la perception audio-visuelle de la parole sera meilleure avec un format « bouche » ne comportant pas de régions obscure qu'avec un format « bouche » comportant ce type de régions.

4.3 Dimension développementale dans la perception audio-visuelle de la parole et les patterns du comportement oculaire

Une autre dimension des questions traitées dans le cadre du présent travail de thèse concerne l'aspect développemental du traitement des visages. En effet, de nombreuses études ont établi que le développement du le traitement analytique (le traitement des composantes faciales) est plus rapide que celui du traitement holistique (traitement du visage en tant qu'un tout) qui atteint sa maturité en adolescence (e.g., Bruce, Campbell, Doherty-Sneddon, Import, Langton, McAuley, & Wright, 2000 ; Carey, Diamond, & Woods, 1980 ; Mondloch, Le Grand, & Mauer, 2002). Aussi, si le traitement de l'information faciale lors de la perception de la parole implique également le traitement holistique du visage de l'orateur, les enfants chez qui le traitement holistique n'est pas encore optimal devraient tirer un moindre bénéfice de l'information faciale dans ce type de perception que les adultes.

Notre septième question de recherche est ainsi de savoir si l'impact du format de l'information visuelle sur la perception audio-visuelle de la parole varie en fonction de l'âge. Dans la lignée des résultats de la majorité des études développementales, notre hypothèse (vii) est que, de façon globale, l'apport de l'information visuelle à la perception bimodale de la parole sera moindre chez les enfants que chez les adultes. Par ailleurs, nous sommes plus particulièrement intéressés de savoir si (viii) le traitement audio-visuel de la parole se faisant à l'aide de l'information faciale intégrale évolue durant la période allant de l'enfance à l'adolescence. Conformément à l'hypothèse selon laquelle l'information faciale est traitée de façon holistique dans la perception bimodale de la parole et aux résultats des études qui ont mis en évidence que le traitement holistique des visages connaît une période de maturation relativement longue, allant possiblement jusqu'à l'adolescence, nous formulons l'hypothèse (viii) de la façon suivante : La différence dans l'apport de l'information visuelle à la perception bimodale de la parole entre les enfants et les adultes sera particulièrement importante avec les formats visuels comportant le visage entier de l'orateur.

Finalement, les études qui se sont intéressées au développement du comportement oculaire lors de la perception des visages ont montré que les enfants présentent des fixations oculaires bien moins concentrées au niveau des yeux, du nez et de la bouche que les adultes, les fixations oculaires des enfants couvrant davantage l'ensemble du visage, y compris le contour facial (Bronson, 1994 ; Kato & Konishi, 2013). Des différences entre les adultes et les enfants pourraient ainsi exister également dans le traitement visuo-attentionnel de l'information faciale lors de la perception bimodale de la parole. Notre neuvième (ix) question de recherche

concerne ainsi le développement du comportement oculaire de l'enfance tardive à l'âge adulte. Conformément aux résultats des études sur le développement du comportement oculomoteur dans la perception des visages, notre hypothèse (ix) est la suivante : Le traitement visuel de l'information faciale lors de la perception audio-visuelle de la parole sera moins centré sur les yeux, le nez et la bouche de l'orateur chez l'enfant que chez les adultes. Pour les problématiques de (vi) à (viii), notre objectif est également d'identifier l'âge auquel le traitement concerné par la problématique atteint le niveau adulte. Dans le cadre de notre étude, nous prenons ainsi en compte quatre populations, les enfants, les pré-adolescents, les adolescents et les adultes. A la fin, pour ce qui est des différences dans les deux formats « bouche », elles concernent les caractéristiques de bas niveau de l'input visuel. Nous pensons ainsi (ix) qu'elles affecteront le traitement visuel et consécutivement la perception audio-visuelle de la parole de la même façon chez les enfants que chez les adultes.

4.4 Degré de dégradation de l'input auditif

Pour finir, nous prenons en compte un autre facteur affectant la perception bimodale de la parole, le degré de dégradation de l'information auditive. Les variations dans cette dimension pourraient notamment affecter le traitement visuo-atentionnel des stimuli visuels et possiblement influencer leur impact sur ce type de perception. Nos questions de recherche (x) et (xi) sont ainsi : (x) Comment le degré de dégradation d l'information auditive affecte le comportement oculaire des personnes percevant la parole de façon bimodale ? (xi) Quel est l'effet des variations dans l'intelligibilité du signal auditif sur l'apport de différents types de stimuli visuels, décrits précédemment, à la perception de la parole ?

Conformément aux résultats de Buchan et *al*. (2008), notre hypothèse (x) est que l'augmentation dans le degré de dégradation de l'information auditive s'accompagnera d'une plus forte concentration des fixations oculaires dans la région de la bouche et du nez. Quant à la question (xi) notre hypothèse est conforme aux résultats des études précédentes. Plus précisément, notre hypothèse (xi) est que l'apport des différents formats des stimuli visuels à la perception de la parole sera plus important dans la condition de dégradation forte du l'information auditive comparativement à la condition de dégradation faible. Si l'ajout du bruit affecte le comportement oculaire de la même façon pour les stimuli de type « visage » que ceux de type « bouche » et si un attrait plus important de regard vers la région buccale facilite la perception de la parole, alors il est également possible (xii) que l'apport des stimuli du format

« bouche » soit augmenté davantage que celui des autres types de stimuli visuels dans la condition de haute dégradation de l'information auditive. Les différences du degré de dégradation de l'information auditive pourraient ainsi expliquer certaines différences des résultats des études ayant comparé l'efficacité du format visuel « bouche » au format « visage » dans la perception de la parole.

Sur le plan développemental, il est possible que les variations du degré de la dégradation de l'information auditive affectent le comportement oculaire des enfants différemment que celui des adultes. Plus précisément, il est possible (xiii) que l'attrait du regard vers le nez et la bouche de l'orateur dans la condition de haute dégradation de l'information auditive soit plus marqué chez les adultes que chez les enfants. Ceci pourrait possiblement expliquer les résultats de Ross et *al*. (2011) qui ont mis en évidence que le pic du gain audio-visuel apparaissant à un SNR plus bas chez les adultes que chez les enfants. Notre hypothèse (xiii) est ainsi que le regard des adultes sera attiré davantage par le nez et la bouche dans la condition de dégradation forte de l'information auditive que ce ne sera le cas chez les enfants.

5 Méthode

Pour répondre aux questions d'études précédemment exposées, deux expériences ont été menées avec des participants de 4 groupes d'âge, adultes (de 18 à 35 ans), adolescents (de 13 à 15 ans), pré-adolescents (de 10 à 12 ans) et enfants (de 7 à 9 ans).

5.1 Expérience 1

5.1.1 Participants

Seize sujets dans chaque groupe d'âge ont participé à l'expérience 1 pour un total de 64 participants. Les participants n'ont rapporté aucun trouble psychiatrique et/ou neurologique. Leur vue était normale ou corrigée. Ils ont rapporté une absence de problèmes auditifs et de problèmes relatifs à la fatigue oculaire. Pour tous les participants, le français était la langue maternelle. Deux participants du groupe « enfants » (*M*=8 ans et 2 mois, *SD*=0,335), trois participants du groupe « adolescents » (*M*=13 ans et 9 mois, *SD*=0,301) et deux participants du groupe « adultes » (*M*=23,6 ans, *SD*=2,301) étaient bilingues. Les participants du groupe « pré-adolescents » (*M*=11 ans et 1 mois, *SD*=0,294) étaient tous monolingues. Tous les participants majeurs ont signé un consentement éclairé (pour le formulaire de consentement éclairé pour les

participants majeurs, voir l'Annexe 1). Pour chaque participant mineur, un consentement éclairé d'un des parents a été obtenu (pour le formulaire de consentement éclairé pour les parents des participants mineurs, voir l'Annexe 2). Les formulaires de consentement ont été établis conformément aux normes du code de déontologie des psychologues. Les données sur l'âge exact des participants ont été recueillies et codées dans un fichier spécifiquement dédié à cette fin.

5.1.2 Matériel

Les essais critiques de l'expérience 1 consistaient en 16 syllabes de type consonne-voyelle. Elles étaient constituées à partir des 16 consonnes simples de la langue française (/ba/, /da/, /fa/, /ga/, /ka/, /la/, /ma/, /na/, /pa/, /ʁa/, /sa/, /ʃa/, /ta/, /va/, /za/, /ʒa/) et de la voyelle /a/. L'expérience 1 comportait 4 conditions de présentation des stimuli expérimentaux, une auditive (condition AU) et trois audio-visuelles. Les conditions audio-visuelles différaient entre elles dans le format de présentation de l'information visuelle. Aussi, dans une condition, le format visuel comportait le visage entier de l'orateur (condition audio-vidéo « visage » ou AVV), dans les deux autres conditions audio-visuelles, le format de présentation de l'information faciale comportait uniquement la bouche de l'orateur. Dans la condition « audio-vidéo bouche-masquage » (AVB-M) le format visuel a été établi par l'application d'un masque noir sur l'ensemble du visage de l'oratrice de façon à ce qu'uniquement la bouche soit visible. Dans la condition « audio-vidéo bouche-extraction » (AVB-E), le format visuel a été établi par extraction de la bouche du visage de l'oratrice. Pour éviter l'introduction de contrastes lumineux dans ce format, la bouche était présentée sur un fond de la couleur de la peau de l'oratrice. (Pour les exemples du matériel visuel employé dans chaque condition expérimentale de l'expérience 1, voir la Figure 25.) Outre les stimuli des essais critiques, l'expérience 1 comportait deux stimuli dédiés à la démonstration du matériel expérimental. Il s'agissait des syllabes /oua/ et /ɲa/.

Pour rendre les stimuli visuels des conditions de présentation audio-visuelles les plus neutres possible, l'oratrice a été instruite à prononcer les syllabes de façon monotone, évitant toute accentuation marquée. Par ailleurs, les mouvements faciaux lors de l'articulation des syllabes étaient limités aux seuls mouvements d'articulateurs (aucune vidéo ne comportait des mouvements de sourcils ou de clignements d'yeux). Dans les vidéos, l'oratrice apparaissait sur un fond clair, son visage n'était pas maquillé. Toutes les manipulations graphiques des vidéos ont été réalisées avec l'Adobe Premiere. La longueur des vidéos était de 2s, les syllabes étant

centrées de façon à ce que le début de voisement se situe vers 0,5s, la fin de voisement se situe vers 1,3s, la fermeture totale de la bouche de l'oratrice se situe vers 1,5s. Précisons encore que l'oratrice était de langue maternelle française. Elle a signé le formulaire d'autorisation de l'utilisation des images filmées à des fins scientifiques (pour le contenu du formulaire, voir l'Annexe 3).

Sur le plan acoustique, les stimuli expérimentaux ont été dégradés par le bruit rose[36] à deux degrés de dégradation différents, SNR-6 (où le bruit rose était de 6dB plus fort que le signal acoustique des syllabes) et SNR-12 (où le bruit rose était de 12dB plus fort que le signal acoustique des syllabes). La présentation du bruit correspondait aux vidéos des syllabes (elle commençait avec le début de chaque vidéo et se terminait avec la fin de chaque vidéo). Les manipulations acoustiques des stimuli ont été réalisées au moyen du logiciel Adobe Audition.

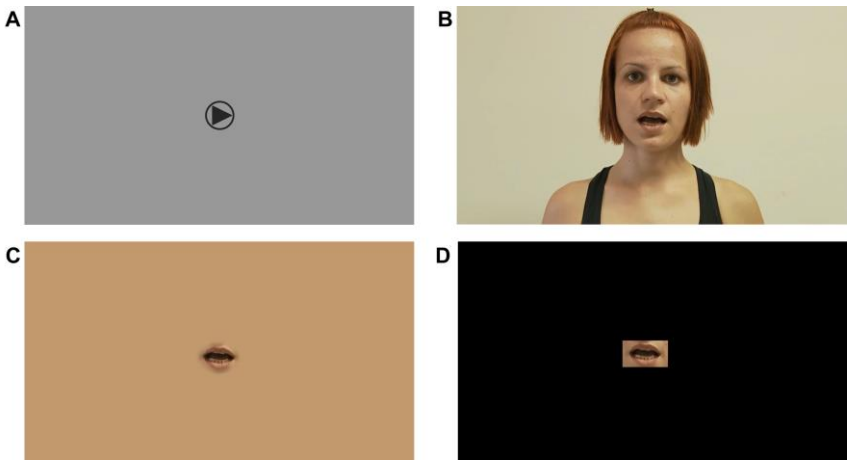

Figure 25. Les exemples du type de matériel visuel utilisé dans les différentes conditions de présentation audio-visuelle de l'expérience 1.

En A, la condition AU, en B a condition AVV, en C la condition AVB-E et en D la condition AVB-M.

[36] Le choix du bruit rose a été fait après le prétest du matériel lors duquel le bruit blanc avait été utilisé pour dégrader les syllabes. Or, ce type de bruit a souvent été jugé comme trop intrusif, essentiellement par les participants du groupe « enfants ».

5.1.3 Procédure

La procédure expérimentale de l'expérience 1 consistait en deux étapes. La première étape était l'étape de démonstration du matériel expérimental. Ainsi, les syllabes /oua/ et /ɲa/ ont été présentées pour les quatre formats de présentation audio(-visuelle), AO, AVV, AVB-M et AVB-E. Chaque format pour chaque syllabe a également été présenté dans les deux conditions de dégradation. Les participants ont été instruits à répéter à haute voix la syllabe qu'ils ont perçue. Il leur a été demandé de donner une seule réponse de manière spontanée. Par ailleurs, il leur a été expliqué qu'ils disposaient de 2s pour donner la réponse. Une démonstration de l'intervalle de 2s séparant deux stimuli consécutifs a été faite.

La deuxième étape de la procédure expérimentale était l'étape de test qui comportait les essais critiques. Lors de cette étape, les stimuli étaient présentés par bloc. Chaque bloc correspondait à un format de présentation audio(-visuelle) d'items. L'ordre de la présentation des blocs dans les protocoles individuels était établi par la procédure de contre-balancement partiel. Aussi, quatre combinaisons de présentation ont été mises en place. Dans l'ensemble des combinaisons, chaque condition est apparue une fois à la 1ère, à la 2ème, à la 3ème et à la 4ème place. Chaque condition a été également précédée et suivie au moins une fois par chaque autre condition. Par ailleurs, chaque bloc était divisé en deux sous-blocs qui correspondaient, chacun, à un degré de dégradation de l'information auditive. L'ordre des sous-blocs a été également contrebalancé de façon à ce que chaque ordre de présentation de blocs commence par le sous-bloc correspondant au SNR-6 pour la moitié des participants de chaque groupe, et par le sous-bloc correspondant au SNR-12 pour l'autre moitié des participants. Ainsi, huit patterns de présentation des blocs et des sous-blocs ont été établis au total (voir l'Annexe 4). Pour chaque groupe d'âge, 2 participants ont été exposés à chacun des huit patterns de présentation des stimuli expérimentaux pour un total de 16 participants par groupe.

A l'intérieur de chaque sous-bloc, les stimuli des essais critiques étaient présentés de façon aléatoire. L'intervalle inter-stimuli était de 2s. Finalement, le premier stimulus de chaque bloc consistait en une syllabe de l'étape de démonstration (/oua/ ou /ɲa/ ; présentée de façon alternée sur l'ensemble des conditions d'un protocole et contrebalancée pour l'ensemble des protocoles individuels). Il n'était pas dégradé acoustiquement et avait pour rôle d'illustrer le type de format de présentation audio(-visuelle) qui allait suivre. Les participants avaient reçu l'instruction de répéter également ce premier stimulus.

Sur le plan visuel, les stimuli expérimentaux ont été présentés à l'aide du logiciel Tobii Studio 1.3 sur l'écran de l'eye-tracker Tobii 1750. Sur le plan acoustique, les stimuli ont été

administrés à l'aide d'un casque audio. L'intensité du signal au niveau des oreilles des participants correspondait au niveau conventionnel de 65dB. Les participants étaient assis à une distance d'environ 60cm de l'écran. Une calibration à 5 points a été effectuée pour chaque participant au début de l'étape de test. Durant cette étape, l'enregistrement du comportement oculaire a été fait par l'eye-tracker Tobii 1750. Dans le cadre du présent travail de thèse, uniquement les protocoles individuels comportant les enregistrements valides pour au moins 75% de l'ensemble des stimuli du protocole ont été maintenus : (i) groupe « enfants » (*M*=85,630 ; *SD*=2,479) ; (ii) groupe « pré-adolescents » (*M*=83,862 ; *SD*=3,081) ; (iii) groupe « adolescents » (*M*=91,219 ; *SD*=2,053) ; (iv) groupe « adultes » (*M*=89,846 ; *SD*=2,932). Le critère du pourcentage minimum des données valides dans un enregistrement du comportement oculaire fixé dans le présent travail de thèse n'a pas été atteint pour 9 protocoles du groupe « enfants », 6 protocoles du groupe « pré-adolescents », 4 protocoles du groupe « adolescents » et 5 protocoles du groupe « adultes ». De plus, lors de chaque passation, les réponses verbales du participant ont été enregistrées par un microphone placé à une distance de 50 à 60cm au-dessus de la tête du participant[37]. A l'issue de l'étape de test, les participants ont été informés sur les objectifs exacts de l'expérience.

Pour finir, les réponses verbales données par les participants ont été codées par deux juges de langue maternelle française. Les juges ne connaissaient ni les objectifs de l'expérience ni les stimuli expérimentaux. Pour chaque item expérimental, quatre catégories de réponse ont été constituées comportant la syllabe correspondant (i) à l'item en question, (ii) à une syllabe acoustiquement proche de l'item en question, (iii) à une syllabe acoustiquement éloignée de l'item en question, (iv) à la catégorie « Autre réponse ». Par exemple, pour l'item expérimental /ba/, les items de réponses correspondant aux quatre catégories en question étaient les suivants (i) /ba/, (ii) /pa/, (iii) /ga/, (iv) « Autre réponse ». L'ensemble des items expérimentaux avec les quatre catégories de réponse pour chaque item des essais critiques a été présenté sur une feuille de codage de format A4. (Pour un exemple des stimuli d'un protocole individuel, voir l'Annexe 5. Pour la feuille de codage des réponses correspondant à l'exemple présenté en Annexe 5, voir l'Annexe 6. Pour le tableau présentant le la syllabe acoustiquement proche de chaque syllabe cible et les syllabes acoustiquement éloignées de la syllabe cible, qui a servi à la mise en place des feuilles de codage des réponses des participants, voir l'Annexe 7.). Pour chaque item des essais critiques, l'ordre de présentation des trois premières catégories de réponse était aléatoire.

[37] Précisons que les modalités du placement du microphone ont été établies suite au prétest lors duquel plusieurs emplacements du microphone ont été évalués.

La quatrième catégorie occupait systématiquement la dernière place. Par ailleurs, les essais non critiques (/oua/ et /ɲa/) étaient également présentés sur la feuille de codage. Ils étaient signalés en rouge. Les juges étaient instruits que ces items correspondaient aux essais non critiques de l'expérience. Les juges ont écouté chaque enregistrement audio des réponses verbales des participants une fois de manière non interrompue. Le codage des réponses données par les participants se faisait de façon simultanée. Le dispositif utilisé par les juges lors du codage des réponses des participants consistait en un casque audio (l'intensité du son était réglée à 65dB au niveau des oreilles), la feuille de codage portant le code indiquant la piste audio correspondante, le ficher de l'ensemble des pistes audio comportant, chacune, toutes les réponses d'un participant.

Les réponses des juges ont été analysées au moyen du test de kappa de Choen. L'analyse a été effectuée aussi bien pour chaque protocole individuel que pour l'ensemble des protocoles individuels dans chaque groupe. Dans le cadre du présent travail de thèse, uniquement les protocoles présentant un degré d'accord inter-juges excellent (κ>0,900) ont été retenus : (i) groupe « enfants » (0,921<κ<1,000 ; κ(groupe)=0,933) ; (ii) groupe « pré-adolescents » (0,930<κ<1,000 ; κ(groupe)=0,951) ; (iii) groupe « adolescents » (0,918<κ<1,000 ; κ(groupe)=0,935) ; (iv) groupe « adultes » (0,937<κ<1,000 ; κ(groupe)=0,960). Le critère du degré d'accord inter-juges minimum fixé dans le présent travail de thèse n'a pas été atteint pour 6 protocoles du groupe « enfants », 3 protocoles du groupe « pré-adolescents », 1 protocole du groupe « adolescents » et 3 protocoles du groupe « adultes ».

A la base du critère en rapport avec la qualité des enregistrements du comportement oculaire des participants et celui en rapport avec l'accord inter-juges, 33 protocoles ont été éliminés au total dans l'expérience 1, 15 dans le groupe « enfants », 7 protocoles dans le groupe « pré-adolescents », 5 protocoles dans le groupe « adolescents » et 6 protocoles dans le groupe « adultes ».

5.2 Expérience 2

5.2.1 Participants

Comme dans l'expérience 1, l'expérience 2 a été réalisée avec des participants de 4 groupes d'âge : (i) « enfants » (*M*=7 ans et 11 mois ; *SD*=0,286) ; (ii) « pré-adolescents » (*M*=11 ans et 3 mois ; *SD*=0,205) ; (iii) « adolescents » (*M*=13 ans et 10 mois ; *SD*=0,251) ; (iv) « adultes » (*M*=24,2 ans ; *SD*=2,718). Seize sujets par groupe ont participé à l'expérience 2.

Aucun des sujets de l'expérience 2 n'avait participé à l'expérience 1. Certains participants étaient bilingues : 1 participant dans le groupe « pré-adolescents », 3 participants dans le groupe « adolescents » et 2 participants dans le groupe « adultes ». Le groupe « enfants » ne comportait pas de participants bilingues. Toutes les modalités de recrutement des participants dans l'expérience 2 étaient identiques à celles précédemment exposées pour l'expérience 1.

5.2.2 Matériel

Les stimuli expérimentaux de l'expérience 2 ont été établis de la même manière que ceux de l'expérience 1. Ils différaient néanmoins des stimuli de l'expérience 1 dans le format de l'information visuelle pour les conditions de présentations audio-visuelle. En effet, l'expérience 2 comportait également 3 modes de présentation audio-visuelle avec les formats visuels suivants : AVV, audio-vidéo « visage bouche active » (format comportant le visage de l'oratrice dans lequel seule la bouche bougeait lors de l'articulation des items, le contexte facial étant statique) (AVV-BA), audio-vidéo « visage régions extra-buccales actives » (format dans lequel la bouche de l'oratrice a été effacée ; on ne voyait que les mouvements dans les régions extra-buccales lors de l'articulation des items) (AVV-EBA)[38]. (Pour les exemples de matériel visuel employé dans chaque condition expérimentale de l'expérience 2, voir la Figure 26.)

[38] Pour le format AVV-EBA, où uniquement les mouvements dans les régions extra-buccales ont été présentés, une variante de format visuel a été pré-testée. Elle comportait la bouche de l'oratrice qui restait statique durant la prononciation des syllabes. Toutefois, les participants les plus jeunes présentaient des difficultés à comprendre pourquoi la bouche ne s'ouvrait pas. La variante du format AVV-EBA, retenue dans notre expérience, leur a été expliquée comme ayant été établie au moyen d'une « gomme magique » qui nous a permis d'effacer la bouche de l'oratrice pour les « embêter ».

Figure 26. Les exemples du type de matériel visuel utilisé dans les différentes conditions de présentation audio-visuelle de l'expérience 2.

En A, la condition AU, en B la condition AVV, en C la condition AVV-BA et en D la condition AVV-EBA.

5.2.3 Procédure

Les modalités de la procédure de l'expérience 2 étaient identiques à celles de l'expérience 1. Quant au pourcentage des données recueillies, les enregistrements du comportement oculaire des participants dans l'expérience 2 présentaient les caractéristiques suivantes : (i) groupe « enfants » (M=86,153 ; SD=2,030) ; (ii) groupe « pré-adolescents » (M=83,792 ; SD=2,859) ; (iii) groupe « adolescents » (M=89,304 ; SD=2,185) ; (iv) groupe « adultes » (M=92,048 ; SD=1,977).). Le critère du pourcentage minimum des données valides dans un enregistrement du comportement oculaire fixé dans le présent travail de thèse n'a pas été atteint pour 10 protocoles du groupe « enfants », 4 protocoles du groupe « pré-adolescents », 5 protocoles du groupe « adolescents » et 3 protocoles du groupe « adultes ». Quant à l'accord inter-juges pour le codage des réponses des participants de l'expérience 2, il présentait les caractéristiques suivantes : (i) groupe « enfants » ($0,929<\kappa<1,000$; κ(groupe)=0,936) ; (ii) groupe « pré-adolescents » ($0,931<\kappa<1,000$; κ(groupe)=0,942) ; (iii) groupe « adolescents » ($0,940<\kappa<1,000$; κ(groupe)=0,945) ; (iv) groupe « adultes » ($0,938< \kappa<1,000$; κ(groupe)=0,943). Le critère du degré d'accord inter-juges minimum fixé dans le présent travail de thèse n'a pas été atteint pour 6 protocoles du groupe « enfants », 3 protocoles du groupe « pré-adolescents », 2 protocoles du groupe « adolescents » et 1 protocole du groupe « adultes ». Au total, 31 protocoles individuels ont été éliminés dans l'expérience 2, 13

protocoles dans le groupe « enfants », 8 protocoles dans le groupe « pré-adolescents », 5 protocoles dans le groupe « adolescents » et 5 protocoles dans le groupe « adultes ».

6 Résultats

Dans le design expérimental classique utilisé dans les expériences 1 et 2, trois types de variables dépendantes ont été retenus : (i) le nombre total des répétitions correctes d'items, appelé, dans la suite du document, la performance totale ; (ii) le gain audio-visuel (gain AV) en termes de la différence dans le nombre total des répétitions correctes entre chaque condition AV et la condition AU[39] ; (iii) la durée des fixations oculaires dans les différentes régions d'intérêt (AOI).

Le nombre des répétitions correctes pour chaque participant a été établi à partir du codage de ses réponses par les juges. Trois cas de figure ont été retenus : (i) l'item de la réponse a été unanimement codé comme correspondant à l'item de l'essai critique en question par les 2 juges (1 point a été attribué dans ce cas) ; (ii) l'item de la réponse a été codé comme correspondant à l'item de l'essai critique en question par 1 des 2 juges (0,5 point a été attribué dans ce cas) ; (iii) l'item de la réponse a été codé comme ne correspondant pas à l'item de l'essai critique en question par les 2 juges (0 point a été attribué dans ce cas). Quant aux données relatives au comportement oculaire des participants, uniquement les données enregistrées dans la fenêtre temporelle de 1,5s ont été retenues pour analyse. Cette fenêtre temporelle débutait avec le point de départ de chaque vidéo et se terminait au moment de la fermeture de la bouche de l'oratrice, ce moment ayant été estimé de façon moyenne pour l'ensemble des vidéos.

6.1 Expérience 1

6.1.1 Performance totale

Les données de la performance totale des participants de l'expérience 1 ont été analysées au moyen de l'ANOVA à design mixte comportant un facteur inter-sujets, le facteur Groupe d'âge, et 2 facteurs de type intra-sujets, les facteurs Format de présentation de l'information audio-visuelle et le SNR, correspondant aux 2 degrés de dégradation de l'information auditive.

[39] Cette méthode de l'évaluation du gain AV a été choisie car elle semble la moins sujette aux exagérations du gain AV aux SNR les plus bas, à l'effet du plafond et à une surcompensation de cet effet pour les SNR les plus élevés qui caractérisent les autres méthodes de l'évaluation du gain AV (pour plus de détails, voir Ross et *al*., 2007).

(Voir le Tableau 1 pour les moyennes et les écart-types pour la performance totale de l'expérience 1.)

Tableau 1. *Moyennes (M) et écarts-types (SD) pour la performance totale des participants des 4 groupes d'âge dans les différentes conditions expérimentales de l'expérience 1.*

	Groupe d'âge	*M*	*M [%]*	*SD*	*SD [%]*	*N*
	Adultes	4,938	30,859	1,879	11,741	16
	Adolescents	3,625	22,656	1,928	12,049	16
AU_SNR-12	Préadolescents	3,188	19,922	1,905	11,906	16
	Enfants	2,563	16,016	1,590	9,940	16
	Total	**3,578**	**22,363**	**1,990**	**12,441**	**64**
	Adultes	8,938	55,859	1,692	10,574	16
	Adolescents	7,063	44,141	1,482	9,261	16
AU_SNR-6	Préadolescents	7,938	49,609	1,982	12,389	16
	Enfants	7,688	48,047	1,138	7,115	16
	Total	**7,906**	**49,414**	**1,706**	**10,665**	**64**
	Adultes	8,250	51,563	1,571	9,816	16
	Adolescents	7,688	48,047	1,493	9,331	16
AVV_SNR-12	Préadolescents	7,250	45,313	2,352	14,702	16
	Enfants	6,938	43,359	1,436	8,976	16
	Total	**7,531**	**47,070**	**1,781**	**11,134**	**64**
	Adultes	10,625	66,406	1,544	9,649	16
	Adolescents	9,000	56,250	1,414	8,839	16
AVV_SNR-6	Préadolescents	9,250	57,813	1,571	9,816	16
	Enfants	9,313	58,203	1,778	11,115	16
	Total	**9,547**	**59,668**	**1,671**	**10,442**	**64**
	Adultes	9,063	56,641	1,692	10,574	16
AVB-E_SNR-12	Adolescents	6,313	39,453	1,250	7,813	16

	Préadolescents	6,813	42,578	2,344	14,652	16
	Enfants	6,188	38,672	2,287	14,292	16
	Total	**7,094**	**44,336**	**2,231**	**13,941**	**64**
	Adultes	11,750	73,438	1,880	11,748	16
	Adolescents	9,688	60,547	1,493	9,331	16
AVB-E_SNR-6	Préadolescents	9,563	59,766	2,279	14,246	16
	Enfants	9,188	57,422	1,721	10,757	16
	Total	**10,047**	**62,793**	**2,081**	**13,007**	**64**
	Adultes	9,625	60,156	1,500	9,375	16
	Adolescents	7,438	46,484	1,315	8,219	16
AVB-M_SNR-12	Préadolescents	7,125	44,531	2,446	15,288	16
	Enfants	7,000	43,750	1,826	11,411	16
	Total	**7,797**	**48,731**	**2,079**	**12,995**	**64**
	Adultes	10,750	67,188	1,238	7,739	16
	Adolescents	8,813	55,078	1,601	10,005	16
AVB-M_SNR-6	Préadolescents	9,125	57,031	1,204	7,526	16
	Enfants	9,750	60,938	1,732	10,825	16
	Total	**9,609**	**60,059**	**1,610**	**10,060**	**64**

6.1.1.1 Effet significatif du facteur Groupe

L'effet du facteur Groupe s'est révélé significatif ($F(3,60)=9,700$; $p<0,001$; $\eta p^2=0,980$). Les comparaisons post-hoc ont été effectuées au moyen du *t*-test pour échantillons indépendants. La méthode de Tukey HSD a été appliquée pour la correction de l'intervalle de confiance pour les comparaisons multiples intergroupes. Les comparaisons post-hoc ont ainsi mis en évidence que la différence intergroupe a été significative dans les cas suivants : (i) entre les groupes « adultes » et « adolescents » ($t(15)=4,355$; $p<0,001$) ; (ii) entre les groupes « adultes » et « pré-adolescents » ($t(15)=4,161$; $p<0,002$) ; (iii) entre les groupes « adultes » et « enfants » ($t(15)=4,647$; $p<0,001$). Pour chaque comparaison, la performance totale des

adultes était supérieure à celle obtenue par les participants dans chaque autre groupe d'âge. (Voir la Figure 27 pour une représentation graphique des résultats.)

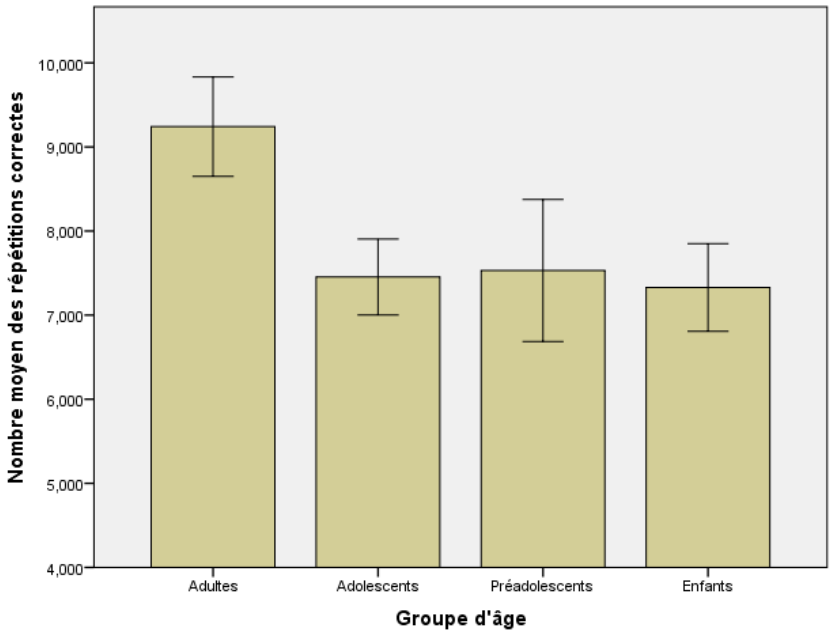

Figure 27. Variations du nombre moyen des répétitions correctes (performance totale) en fonction du groupe d'âge.[40]

6.1.1.2 Effet significatif du facteur Format

L'effet du facteur Format s'est révélé significatif ($F(3,180)=135,432$; $p<0,001$; $\eta p^2=0,693$). Les comparaisons post-hoc ont été effectuées au moyen du *t*-test pour échantillons appariés. Elles ont montré que les différences entre le format AU et chaque autre format AV ont été significatives : (i) AU *vs* AVV ($t(63)=-16,649$; $p<0,001$) ; (ii) AU *vs* AVB-E ($t(63)=-14,070$; $p<0,001$) ; (iii) AU *vs* AVB-M ($t(63)=-17,418$; $p<0,001$). Dans les trois comparaisons significatives, la performance totale obtenue dans chacune des conditions AV a été supérieure à celle obtenue dans la condition AU. (Voir la Figure 28 pour une représentation graphique des résultats.)

[40] Notons que, pour l'ensemble des graphiques, les barres représentent les intervalles de confiance.

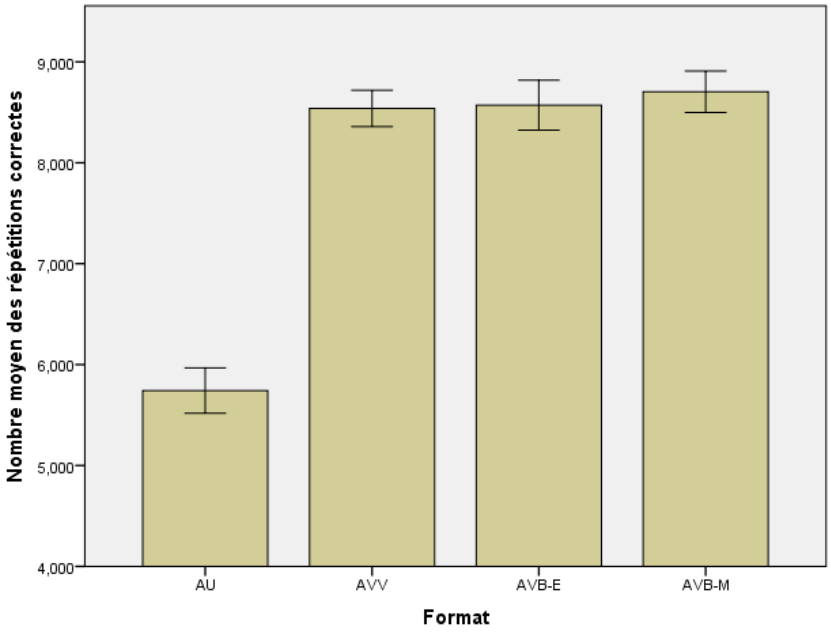

Figure 28. Variations du nombre moyen des répétitions correctes (performance totale) en fonction du format de présentation de l'information audio-visuelle.

6.1.1.3 Effet significatif du facteur SNR

L'effet du facteur SNR s'est révélé significatif ($F(1,60)=365,278$; $p<0,001$; $\eta p^2=0,859$). On constate que la performance totale des participants a été meilleure dans la condition SNR-6 que dans la condition SNR-12. (Voir la Figure 29 pour une représentation graphique des résultats.)

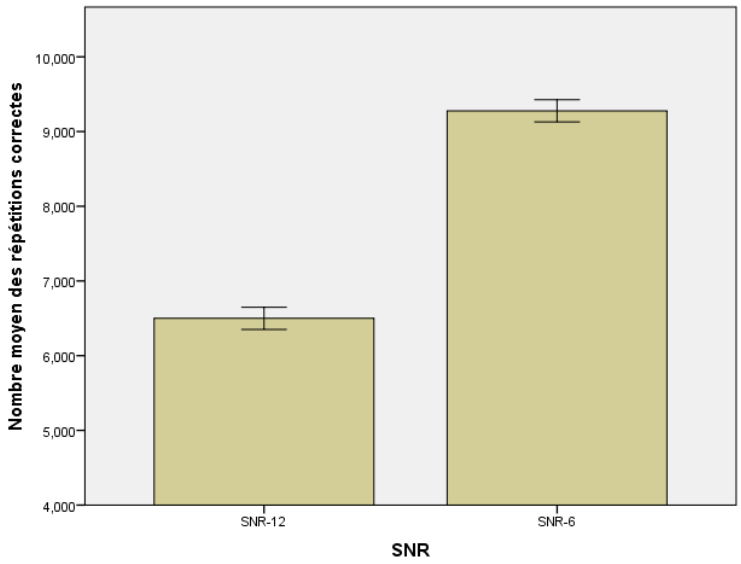

Figure 29. Variations du nombre moyen des répétitions correctes (performance totale) en fonction du degré de dégradation de l'information auditive (SNR).

6.1.1.4 Effet significatif de l'interaction Format x SNR

L'effet de l'interaction Format x SNR s'est révélé significatif ($F(3,180)=25,955$; $p<0,001$; $\eta p^2=0,302$). Les comparaisons post-hoc ont été effectuées au moyen du *t*-test pour échantillons appariés. Elles ont mis en évidence que les différences entre le format AU et chaque autre format AV ont été significatives aussi bien pour le SNR-6 ((i) AU *vs* AVV ($t(63)=-8,485$; $p<0,001$) ; (ii) AU *vs* AVB-E ($t(63)=-10,254$; $p<0,001$) ; (iii) AU *vs* AVB-M ($t(63)=-7,659$; $p<0,001$)) que pour le SNR-12 ((i) AU *vs* AVV ($t(63)=-14,618$; $p<0,001$) ; (ii) AU *vs* AVB-E ($t(63)=-11,482$; $p<0,001$) ; (iii) AU *vs* AVB-M ($t(63)=-17,552$; $p<0,001$)). Pour les deux niveaux de dégradation de l'information auditive, la performance totale des participants dans chaque condition AV a été supérieure à celle obtenue dans la condition AU. En revanche, la performance totale pour les formats AVV, AVB-E et AVB-M variait en fonction du SNR. En effet, pour le SNR-12, la performance totale pour le format AVB-M a été supérieure à celle obtenue pour le format AVB-E ($t(63)=-2,514$; $p<0,014$). Ce pattern a été différent pour le SNR-6, où la performance totale pour le format AVB-E a été supérieure à celle obtenue pour le format AVV ($t(63)=-2,435$; $p<0,018$). Par ailleurs, pour ce niveau de dégradation, la différence entre la performance totale pour le format AVB-E et le format AVB-M a été proche du seuil de signification ($t(63)=1,891$; $p<0,063$). (Voir la Figure 30 pour une représentation graphique des résultats.)

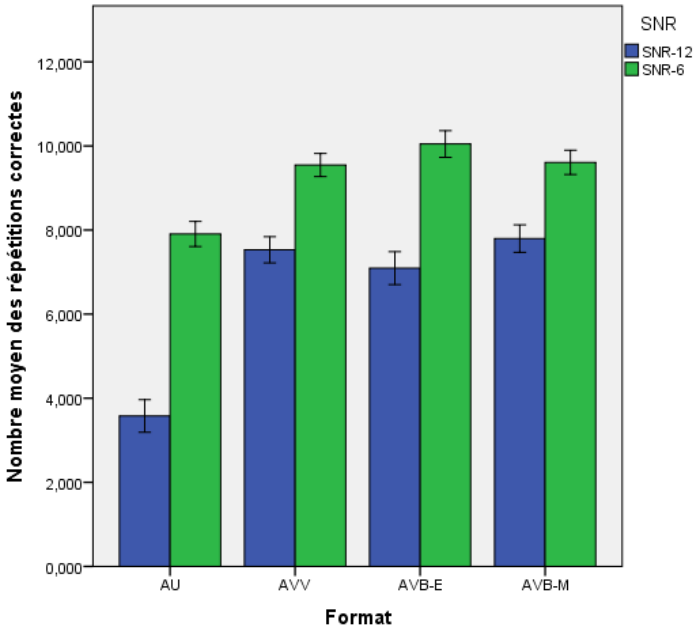

Figure 30. Variations du nombre moyen des répétitions correctes (performance totale) en fonction du format de présentation de l'information audio-visuelle et en fonction du degré de dégradation de l'information auditive (SNR).

6.1.2 Gain AV

Les données relatives au gain AV de l'expérience 1 ont été analysées au moyen de l'ANOVA à design mixte comportant 1 facteur de type inter-sujets, le Groupe d'âge, et 2 facteurs de type intra-sujets, le Format de présentation de l'information visuelle et le SNR. (Voir le Tableau 2 pour les moyennes et les écart-types pour le gain AV de l'expérience 1.)

Tableau 2. *Moyennes (M) et écarts-types (SD) pour le gain AV des participants des 4 groupes d'âge dans les différentes conditions expérimentales de l'expérience 1.*

	Groupe d'âge	*M*	*M [%]*	*SD*	*SD [%]*	*N*
	Adultes	3,313	20,703	2,330	14,563	16
	Adolescents	4,063	25,391	2,620	16,373	16
AVV_SNR-12	Pré-adolescents	4,063	25,391	1,879	11,741	16
	Enfants	4,375	27,344	1,784	11,151	16
	Total	**3,953**	**24,707**	**2,163**	**13,521**	**64**
	Adultes	1,688	10,547	1,580	9,874	16
	Adolescents	1,938	12,109	1,914	11,961	16
AVV_SNR-6	Pré-adolescents	1,313	8,203	1,493	9,332	16
	Enfants	1,625	10,156	1,204	7,526	16
	Total	**1,641**	**10,254**	**1,547**	**9,668**	**64**
	Adultes	4,125	25,781	2,391	14,943	16
	Adolescents	2,688	16,797	2,089	13,054	16
AVB-E_SNR-12	Pré-adolescents	3,625	22,656	1,784	11,151	16
	Enfants	3,625	22,656	3,284	20,524	16
	Total	**3,516**	**21,973**	**2,449**	**15,309**	**64**
	Adultes	2,813	17,578	1,601	10,005	16
	Adolescents	2,625	16,406	1,668	10,427	16
AVB-E_SNR-6	Pré-adolescents	1,625	10,156	1,668	10,427	16
	Enfants	1,500	9,375	1,461	9,129	16
	Total	**2,141**	**13,379**	**1,670**	**10,439**	**64**
	Adultes	4,688	29,297	2,024	12,649	16
	Adolescents	3,813	23,828	1,682	10,513	16
AVB-M_SNR-12	Pré-adolescents	3,938	24,609	2,081	13,004	16
	Enfants	4,438	27,734	1,931	12,069	16
	Total	**4,219**	**26,367**	**1,923**	**12,018**	**64**

	Adultes	1,813	11,328	1,167	7,295	16
	Adolescents	1,750	10,938	2,206	13,788	16
AVB-M_SNR-6	Pré-adolescents	1,188	7,422	2,287	14,292	16
	Enfants	2,063	12,891	1,181	7,384	16
	Total	**1,703**	**10,645**	**1,779**	**11,118**	**64**

6.1.2.1 Effet significatif du facteur SNR

L'effet du facteur SNR s'est révélé significatif ($F(1,60)=56,908$; $p<0,001$; $\eta p^2=0,487$). On constate que le gain AV de la condition SNR-12 a été supérieur à celui de la condition SNR-6. (Voir la Figure 31 pour une représentation graphique des résultats.)

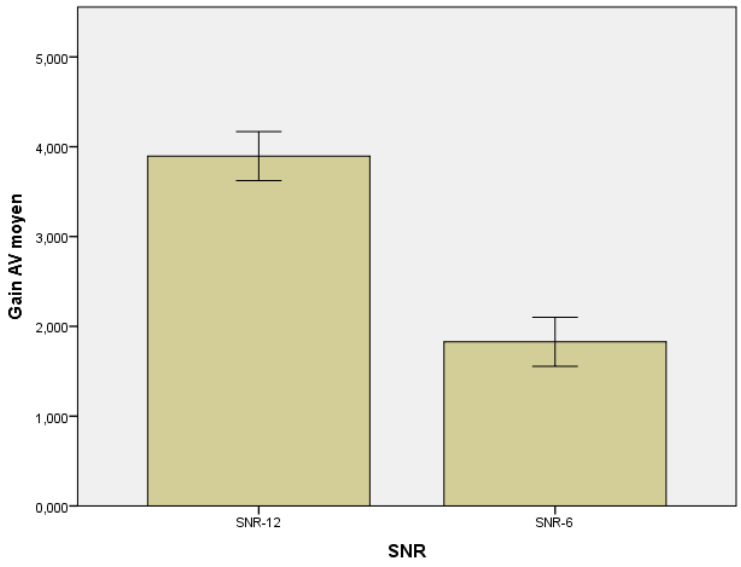

Figure 31. Variations du gain AV moyen en fonction du degré de dégradation de l'information auditive (SNR).

6.1.2.2 Effet significatif de l'interaction Groupe x Format

L'effet de l'interaction Groupe x Format s'est révélé significatif ($F(6,120)=2,268$; $p<0,045$; $\eta p^2=0,102$). Les comparaisons post-hoc intra-sujets, effectuées au moyen du *t*-test pour échantillons appariés, ont mis en évidence que les seules différences de ce type ont apparue

dans le groupe « adultes ». En effet, dans ce groupe, le gain AV obtenu avec le format AVB-E a été supérieur à celui obtenu avec le format AVV (t(15)=-2,930 ; p<0,011). De la même façon, le gain AV obtenu avec le format AVB-M a été supérieur à celui obtenu avec le format AVV (t(15)=-2,449 ; p<0,028). Finalement, les comparaisons post-hoc inter-sujets, réalisées pour toutes les combinaisons des groupes d'âge pour chaque format AV, n'ont mis en évidence aucune différence significative[41]. (Voir la Figure 32 pour une représentation graphique des résultats.)

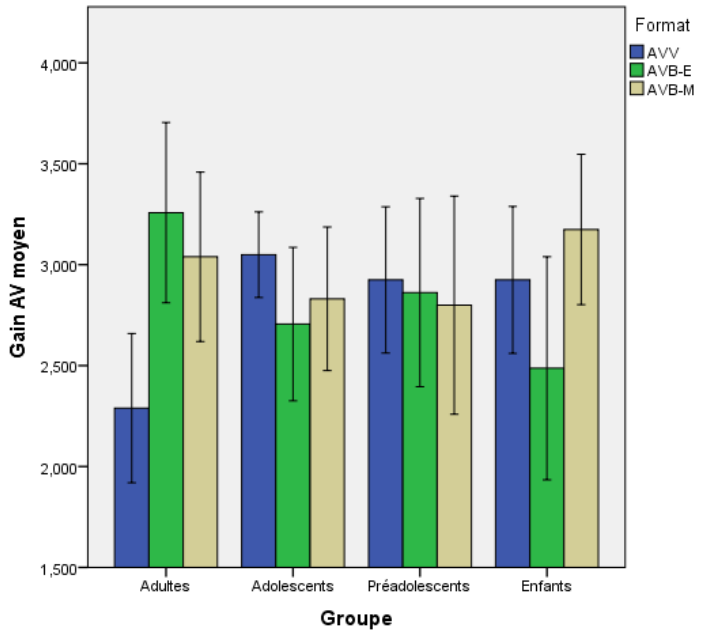

Figure 32. Variations du gain AV moyen en fonction du groupe d'âge et en fonction du format de présentation de l'information visuelle.

6.1.2.3 Effet significatif de l'interaction Format x SNR

L'effet de l'interaction Format x SNR s'est révélé significatif (F(2,120)=7,735; p<0,002; ηp^2=0,114). Les comparaisons post-hoc intra-sujets, effectuées au moyen du t-test pour échantillons appariés, ont mis en évidence qu'au niveau du SNR-12, le gain AV obtenu pour le format AVB-E a été inférieur à celui obtenu pour le format AVB-M (t(63)=-2,514 ;

[41] Notons que la correction de Tukey HSD a été appliquée pour l'ensemble des comparaisons post-hoc intergroupes réalisées sur les données concernant les réponses verbales des participants (performance totale et gain AV) dans les expériences 1 et 2.

p<0,015). En revanche, au niveau du SNR-6, le gain AV le plus élevé a été obtenu pour le format AVB-E. Le gain AV de cette condition a été supérieur à celui obtenu pour le format AVV. Par ailleurs, au niveau du SNR-12, la différence entre le gain AV obtenu pour le format AVB-E et celui obtenu pour le format AVB-M a été proche du seuil de signification (*t*(63)=1,891 ; *p*<0,063) et indiquait une supériorité du format AVB-E comparativement au format AVB-M à ce niveau. (Voir la Figure 33 pour une représentation graphique des résultats.)

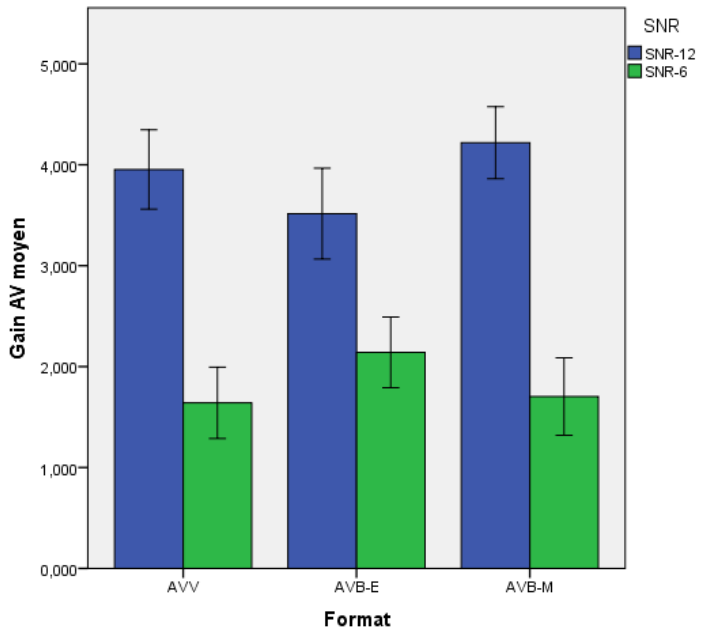

Figure 33. Variations du gain AV moyen en fonction du format de présentation de l'information visuelle et en fonction du degré de dégradation de l'information auditive (SNR).

6.1.3 Comportement oculaire (durée des fixations oculaires)

L'indice du comportement oculaire pris en compte dans les expériences 1 et 2 du présent travail de thèse était la durée des fixations oculaires dans les différentes régions d'intérêt qui correspondaient aux différentes régions faciales, notamment la région buccale, le nez, les yeux, le contour faciale et les régions extra-buccales ou les régions à l'extérieur de la bouche, du nez et des yeux et à l'intérieur du contour du visage (que nous appellerons *intra-contour* dans la suite du document). (Pour une représentation graphique des régions d'intérêt telles qu'elles ont été définies pour l'ensemble des formats AV des expériences 1 et 2, voir la Figure 34.) L'expérience 1 comportant deux formats visuels réduits à la seule région buccale de l'oratrice,

nous avons conduit une ANOVA à design mixte avec un facteur de type inter-sujets, Groupe d'âge, et deux facteurs de type intra-sujets, Format et SNR, sur la durée des fixations oculaires dans la région buccale de l'oratrice pour les trois formats visuels. (Voir le Tableau 3 pour les moyennes et les écarts-types dans les différentes conditions expérimentales correspondant à cette analyse.) Par ailleurs, une ANOVA à design mixte avec un facteur inter-sujet, Groupe d'âge, et deux facteurs intra-sujets, SNR et Régions d'intérêt (*areas of interest* (*AOI*)) a été réalisée sur la durée des fixations oculaires dans les différentes régions faciales pour le format AVV seul[42]. (Voir le Tableau 4 pour les moyennes et les écarts-types dans les différentes conditions expérimentales correspondant à cette analyse.)

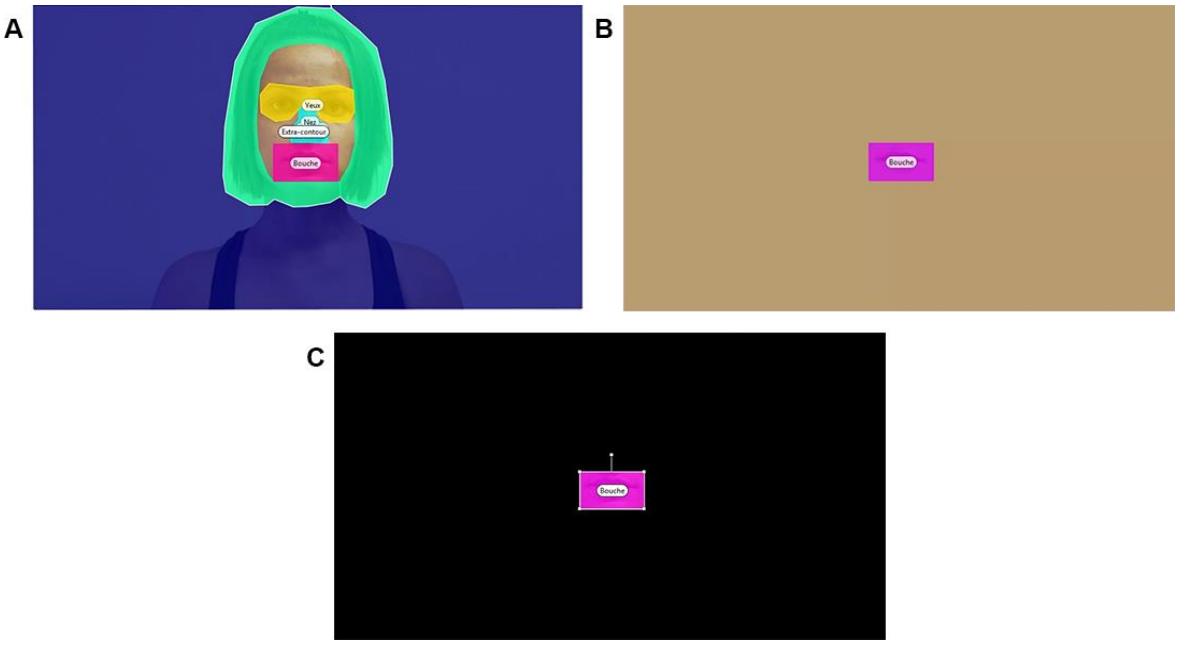

Figure 34. Les régions d'intérêt (AOI) pour les différents types de formats AV des expériences 1 et 2.

En A, les AOIs pour les formats AVV des expériences 1 et 2. En violet, la région buccale, en bleu clair, le nez, en jaune, les yeux, en vert le contour et en couleur peau l'intra-contour. (Pour des raisons de facilitation de l'extraction des données, la zone en dehors du visage de l'oratrice

[42] Précisons que, pour cette analyse, la supposition de la distribution normale des données et celle de l'homogénéité des écarts-types dans les différentes conditions expérimentales ont été violées. Toutefois, l'ANOVA résiste à ce type de violations : (i) dans le cas d'échantillons supérieurs à 10 participants pour le premier type de violation (e.g., Schmider, Ziegler, Danay, Meyer, & Bühner, 2010) ; (ii) pour les échantillons comportant 15 participants ou plus pour le deuxième type de violation (e.g., Ramsey, 1980).

a également été définie comme une AOI (en beau foncé), or, les données de cette AOI n'ont pas été analysées.) En B et en C, la région buccale de l'oratrice (marquée en violet) pour les formats AVB-E et AVB-M de l'expérience 1 respectivement.

Tableau 3. *Moyennes (M) et écarts-types (SD) pour la durée des fixations oculaires dans la région buccale de l'oratrice des participants des 4 groupes d'âge dans les différentes conditions expérimentales de l'expérience 1.*

	Groupe d'âge	*M*	*M [%]*	*SD*	*SD [%]*	*N*
	Adultes	1,217	81,154	,354	23,623	16
	Adolescents	1,184	78,914	,251	16,706	16
AVV_SNR-12	Pré-adolescents	1,080	72,005	,327	21,783	16
	Enfants	1,070	71,347	,261	17,399	16
	Total	**1,138**	**75,855**	**,301**	**20,070**	**64**
	Adultes	1,165	77,635	,344	22,934	16
	Adolescents	1,119	74,572	,244	16,277	16
AVV_SNR-6	Pré-adolescents	1,004	66,941	,261	17,430	16
	Enfants	,972	64,767	,285	19,007	16
	Total	**1,065**	**70,979**	**,291**	**19,369**	**64**
	Adultes	1,359	90,630	,079	5,269	16
	Adolescents	1,286	85,720	,288	19,189	16
AVB-E_SNR-12	Pré-adolescents	1,218	81,174	,240	15,974	16
	Enfants	1,228	81,845	,278	18,541	16
	Total	**1,273**	**84,842**	**,238**	**15,852**	**64**
	Adultes	1,373	91,560	,049	3,280	16
	Adolescents	1,287	85,808	,322	21,466	16
AVB-E_SNR-6	Pré-adolescents	1,193	79,542	,233	15,522	16
	Enfants	1,211	80,760	,376	25,094	16
	Total	**1,266**	**84,418**	**,278**	**18,505**	**64**

	Adultes	,919	61,276	,521	34,755	16
	Adolescents	1,228	81,853	,197	13,137	16
AVB-M_SNR-12	Pré-adolescents	1,073	71,524	,454	30,248	16
	Enfants	1,154	76,956	,370	24,658	16
	Total	**1,094**	**72,902**	**,411**	**27,398**	**64**
	Adultes	,909	60,577	,565	37,688	16
	Adolescents	1,221	81,433	,345	23,023	16
AVB-M_SNR-6	Pré-adolescents	,971	64,755	,515	34,334	16
	Enfants	1,129	75,239	,372	24,814	16
	Total	**1,058**	**70,501**	**,465**	**31,002**	**64**

Tableau 4. *Moyennes (M) et écarts-types (SD) pour la durée des fixations oculaires des participants des 4 groupes d'âge pour les différentes régions faciales de l'oratrice dans l'ensemble conditions expérimentales de l'expérience 1.*

	Groupe d'âge	*M*	*M [%]*	*SD*	*SD [%]*	*N*
	Adultes	1,217	81,154	,354	23,623	16
	Adolescents	1,184	78,914	,251	16,706	16
SNR-12_Bouche	Pré-adolescents	1,080	72,005	,327	21,783	16
	Enfants	1,070	71,347	,261	17,399	16
	Total	**1,138**	**75,855**	**,301**	**20,070**	**64**
	Adultes	,020	1,339	,035	2,342	16
	Adolescents	,009	,575	,011	,704	16
SNR-12_Contour	Pré-adolescents	,036	2,414	,058	3,885	16
	Enfants	,050	3,321	,099	6,605	16
	Total	**,029**	**1,912**	**,061**	**4,063**	**64**
SNR-12_Nez	Adultes	,090	5,969	,201	13,419	16

	Adolescents	,125	8,340	,147	9,810	16
	Pré-adolescents	,104	6,930	,199	13,233	16
	Enfants	,120	8,003	,123	8,210	16
	Total	**,110**	**7,310**	**,167**	**11,154**	**64**
	Adultes	,046	3,091	,136	9,071	16
	Adolescents	,032	2,104	,041	2,738	16
SNR-12_Yeux	Pré-adolescents	,028	1,888	,045	3,009	16
	Enfants	,044	2,926	,075	4,974	16
	Total	**,038**	**2,502**	**,082**	**5,449**	**64**
	Adultes	,033	2,224	,051	3,389	16
	Adolescents	,048	3,196	,058	3,897	16
SNR-12_Intra-contour	Pré-adolescents	,076	5,091	,116	7,710	16
	Enfants	,086	5,739	,092	6,121	16
	Total	**,061**	**4,063**	**,084**	**5,609**	**64**
	Adultes	1,165	77,635	,344	22,934	16
	Adolescents	1,119	74,572	,244	16,277	16
SNR-6_Bouche	Pré-adolescents	1,004	66,941	,261	17,430	16
	Enfants	,972	64,767	,285	19,007	16
	Total	**1,065**	**70,979**	**,291**	**19,369**	**64**
	Adultes	,026	1,703	,047	3,132	16
	Adolescents	,036	2,367	,070	4,643	16
SNR-6_Contour	Pré-adolescents	,055	3,648	,091	6,075	16
	Enfants	,079	5,236	,133	8,893	16
	Total	**,049**	**3,239**	**,091**	**6,077**	**64**
	Adultes	,095	6,346	,211	14,069	16
	Adolescents	,138	9,186	,146	9,764	16
SNR-6_Nez	Pré-adolescents	,143	9,510	,141	9,399	16
	Enfants	,153	10,183	,161	10,701	16
	Total	**,132**	**8,806**	**,165**	**10,968**	**64**

	Adultes	,056	3,740	,105	6,994	16
	Adolescents	,050	3,308	,082	5,461	16
SNR-6_Yeux	Pré-adolescents	,044	2,918	,038	2,521	16
	Enfants	,065	4,358	,098	6,510	16
	Total	**,054**	**3,581**	**,083**	**5,536**	**64**
	Adultes	,060	3,995	,084	5,596	16
	Adolescents	,053	3,506	,059	3,921	16
SNR-6_Intra-contour	Pré-adolescents	,071	4,734	,067	4,482	16
	Enfants	,078	5,225	,076	5,036	16
	Total	**,065**	**4,365**	**,071**	**4,731**	**64**

6.1.3.1 Durée des fixations oculaires dans la région buccale (tous les formats AV)

6.1.3.1.1 Effet significatif du facteur Format

L'effet du facteur Format s'est révélé significatif ($F(2,120)=10,141$; $p<0,001$; $\eta p^2=0,145$). Les comparaisons post-hoc ont été réalisées au moyen du *t*-test pour échantillons appariés. Elles ont mis en évidence que la région buccale a été fixée plus longuement dans le format AVB-E que dans les deux autres formats, AVV ($t(63)=4,541$; $p<0,001$) et AVB-M ($t(63)=4,409$; $p<0,001$). (Voir la Figure 35 pour une représentation graphique des résultats.)

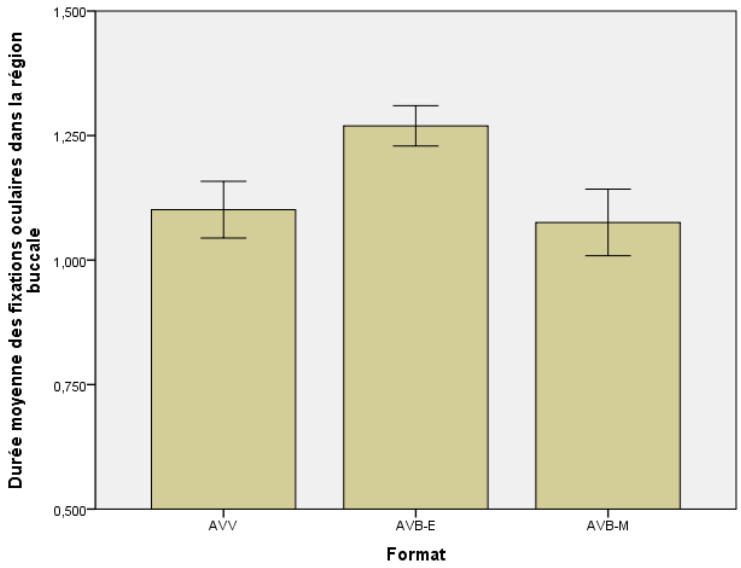

Figure 35. Variations de la durée moyenne (en secondes) des fixations oculaires dans la région buccale de l'oratrice en fonction du format de présentation de l'information visuelle.

6.1.3.1.2 Effet significatif du facteur SNR

L'effet du format SNR s'est révélé significatif ($F(1,60)=10,590$; $p<0,002$; $\eta p^2=0,150$). Sur l'ensemble des formats, la région buccale a été fixée plus longtemps dans la condition de dégradation moyenne/forte de l'information auditive (SNR-12) que dans la condition de dégradation faible (SNR-6). (Voir la Figure 36 pour une représentation graphique des résultats.)

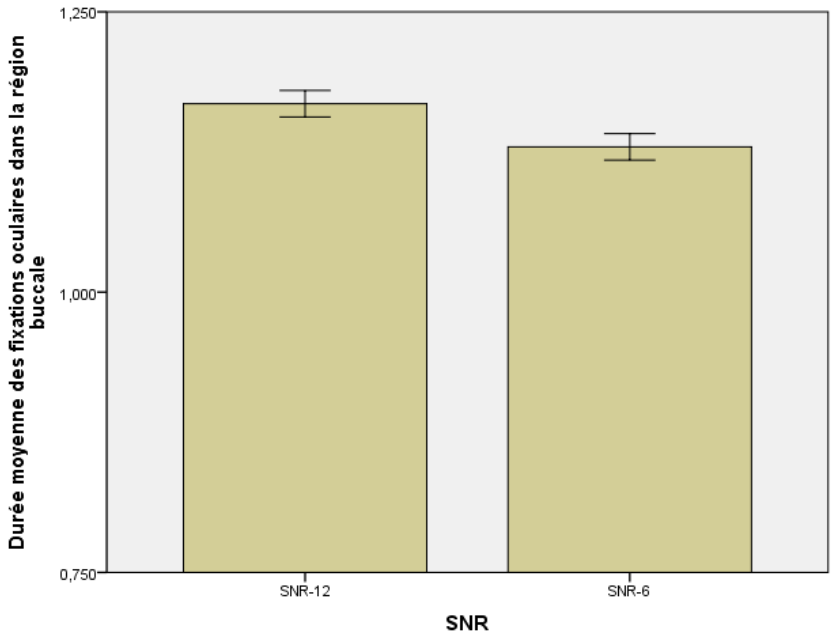

Figure 36. Variations de la durée moyenne (en secondes) des fixations oculaires dans la région buccale de l'oratrice en fonction du degré de dégradation de l'information auditive (SNR).

6.1.3.1.3 Effet significatif de l'interaction Groupe x Format

L'effet de l'interaction Groupe x Format s'est révélé significatif ($F(6,120)=2,468$; $p<0,040$; $\eta p^2=0,110$). Les comparaisons post-hoc de type intra-sujet ont été réalisées au moyen du *t*-test pour échantillons appariés. Elles ont mis en évidence que la région buccale a été fixée plus longtemps dans le format AVB-E que dans le format AVV dans les groupes « enfants » ($t(15)=-2,782$; $p<0,015$) et « pré-adolescents » ($t(15)=-3,504$; $p<0,004$). Dans le groupe « adultes », la région buccale a été fixée plus longtemps dans le format AVB-E que dans le format AVB-M ($t(15)=3,592$; $p<0,004$). Dans ce groupe, la différence a été proche du seuil de signification également pour la comparaison AVB-E *vs* AVV ($t(15)=-2,039$; $p<0,060$). Les comparaisons post-hoc de type inter-sujets (avec correction de Tukey HSD), réalisées pour l'ensemble des combinaisons des groupes d'âge pour chaque format AV, n'ont pas révélé de différences significatives. (Voir la Figure 37 pour une représentation graphique des résultats.)

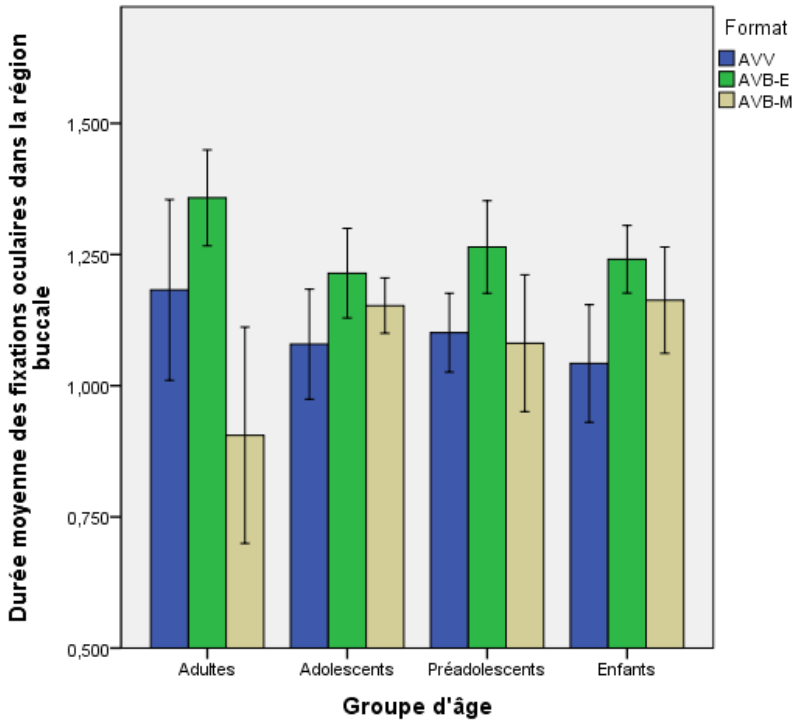

Figure 37. Variations de la durée moyenne (en secondes) des fixations oculaires dans la région buccale de l'oratrice en fonction du groupe d'âge et en fonction du format de présentation de l'information visuelle.

6.1.3.2 Durée des fixations oculaires dans les différentes régions faciales (format AVV)

6.1.3.2.1 Effet significatif du facteur AOI

L'effet du facteur AOI s'est révélé significatif ($F(4,240)=459,855$; $p<0,001$; $\eta p^2=0,885$). Les comparaisons post-hoc ont été réalisées au moyen du *t*-test pour échantillons appariés. Elles ont mis en évidence que la région buccale a été fixée plus longtemps que le nez ($t(63)=18,491$; $p<0,001$), les yeux ($t(63)=25,143$; $p<0,001$), l'intra-contour ($t(63)=24,714$; $p<0,001$) et le contour ($t(63)=27,256$; $p<0,001$). Par ailleurs, le nez a été fixé plus longtemps que les yeux ($t(63)=4,412$; $p<0,001$), l'intra-contour ($t(63)=3,053$; $p<0,040$) et le contour ($t(63)=-3,565$; $p<0,007$). (Voir la Figure 38 pour une représentation graphique des résultats.)

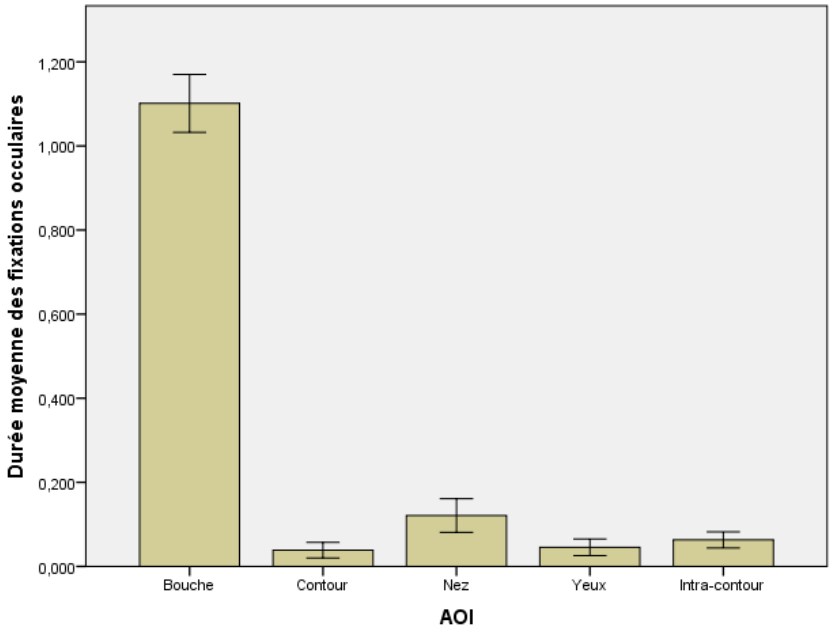

Figure 38. Variations de la durée moyenne (en secondes) des fixations oculaires en fonction de l'AOI (en fonction de la région faciale de l'oratrice).

6.1.3.2.2 Effet significatif de l'interaction AOI x SNR

L'effet de l'interaction AOI x SNR s'est révélé significatif ($F(4,240)=8,285$; $p<0,001$; $\eta p^2=0,121$). Les comparaisons post-hoc ont été réalisées au moyen du *t*-test pour échantillons appariés. Elles ont mis en évidence que le pattern de la durée des fixations oculaires dans les différentes régions faciales (décrit ci-dessus) variait quelque peu en fonction du bruit. En effet, conformément au pattern décrit pour l'effet du facteur AOI, les différences dans la durée des fixations oculaires entre la bouche et les autres régions faciales ont été toutes significatives, pour les deux niveaux de dégradation de l'information auditive. (Pour SNR-6 : (i) Bouche *vs* Nez ($t(63)=17,333$; $p<0,001$) ; (ii) Bouche *vs* Yeux ($t(63)=23,287$; $p<0,001$) ; (iii) Bouche *vs* Intra-Contour ($t(63)=23,442$; $p<0,001$) ; (iv) Bouche *vs* Contour ($t(63)=24,961$; $p<0,001$). Pour SNR-12 : (i) Bouche *vs* Nez ($t(63)=18,496$; $p<0,001$) ; (ii) Bouche *vs* Yeux ($t(63)=24,805$; $p<0,001$) ; (iii) Bouche *vs* Intra-Contour ($t(63)=23,763$; $p<0,001$) ; (iv) Bouche *vs* Contour ($t(63)=27,143$; $p<0,001$).) Il en était de même pour les comparaisons entre le nez et les autres régions faciales. (Pour SNR-12 : (i) Nez *vs* Yeux ($t(63)=4,019$; $p<0,001$) ; (ii) Nez *vs* Intra-Contour ($t(63)=2,341$; $p<0,023$) ; (iii) Nez *vs* Contour ($t(63)=3,644$; $p<0,002$). Pour SNR-6 : (i) Nez *vs* Yeux ($t(63)=4,404$; $p<0,001$) ; (ii) Nez *vs* Intra-Contour ($t(63)=3,349$; $p<0,002$) ; (iii) Nez *vs* Contour ($t(63)=3,346$; $p<0,002$).) La seule différence par rapport au

pattern des résultats décrit précédemment (section 6.1.3.2.1 du présent chapitre) est apparue au niveau du SNR-12 où la durée des fixations oculaires dans la région intra-contour a été supérieure à celle dans la région contour ($t(63)=3,292$; $p<0,003$). (Voir la Figure 39 pour une représentation graphique des résultats.)

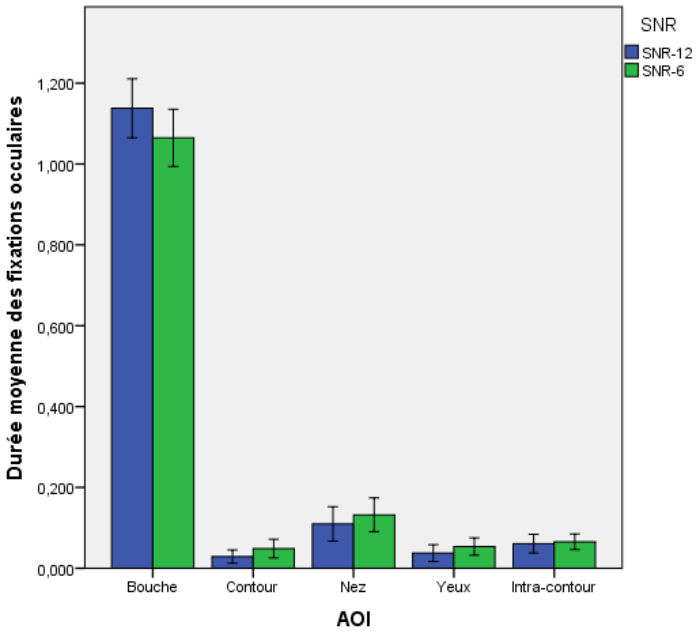

Figure 39. Variations de la durée moyenne (en secondes) des fixations oculaires en fonction de l'AOI (en fonction de la région faciale de l'oratrice) et en fonction du degré de dégradation de l'information auditive (SNR).

6.2 Expérience 2

6.2.1 Performance totale

Les données de la performance totale des participants de l'expérience 1 ont été analysées au moyen de l'ANOVA à design mixte comportant un facteur inter-sujets, le facteur Groupe d'âge, et 2 facteurs de type intra-sujets, les facteurs Format de présentation de l'information audio-visuelle et le SNR, correspondant aux 2 degrés de dégradation de l'information auditive. (Voir le Tableau 5 pour les moyennes et les écart-types pour la performance totale de l'expérience 2.)

Tableau 5. *Moyennes (M) et écarts-types (SD) pour la performance totale des participants des 4 groupes d'âge dans les différentes conditions expérimentales de l'expérience 2.*

	Groupe d'âge	*M*	*M [%]*	*SD*	*SD [%]*	*N*
	1	5,000	31,250	2,191	13,693	16
	2	3,813	23,828	2,105	13,153	16
AU_SNR-12	3	3,438	21,484	1,590	9,940	16
	4	3,625	22,656	1,628	10,174	16
	Total	**3,969**	**24,805**	**1,952**	**12,197**	**64**
	1	8,750	54,688	1,438	8,985	16
	2	7,875	49,219	1,544	9,649	16
AU_SNR-6	3	6,250	39,063	2,145	13,405	16
	4	7,938	49,609	1,611	10,070	16
	Total	**7,703**	**48,145**	**1,900**	**11,873**	**64**
	1	9,438	58,984	1,896	11,852	16
	2	7,813	48,828	1,559	9,741	16
AVV_SNR-12	3	6,875	42,969	2,187	13,669	16
	4	7,438	46,484	1,788	11,173	16
	Total	**7,891**	**49,316**	**2,063**	**12,896**	**64**
	1	11,563	72,266	1,861	11,630	16
	2	9,813	61,328	1,515	9,470	16
AVV_SNR-6	3	8,875	55,469	2,630	16,437	16
	4	9,375	58,594	1,088	6,799	16
	Total	**9,906**	**61,914**	**2,083**	**13,021**	**64**
	1	8,750	54,688	2,049	12,809	16
	2	7,688	48,047	1,401	8,756	16
AVV-BA_SNR-12	3	6,875	42,969	1,668	10,427	16
	4	7,125	44,531	2,062	12,885	16

	Total	**7,609**	**47,559**	**1,916**	**11,974**	**64**
	1	11,375	71,094	1,962	12,263	16
	2	9,938	62,109	1,692	10,574	16
AVV-BA_SNR-6	3	8,750	54,688	2,696	16,848	16
	4	10,563	66,016	1,209	7,558	16
	Total	**10,156**	**63,477**	**2,147**	**13,419**	**64**
	1	6,813	42,578	2,198	13,735	16
	2	5,313	33,203	1,852	11,574	16
AVV-EBA_SNR-12	3	5,063	31,641	2,351	14,696	16
	4	4,563	28,516	2,190	13,687	16
	Total	**5,438**	**33,984**	**2,267**	**14,168**	**64**
	1	9,500	59,375	1,862	11,637	16
	2	7,875	49,219	1,784	11,151	16
AVV-EBA_SNR-6	3	6,688	41,797	1,250	7,813	16
	4	6,938	43,359	1,652	10,325	16
	Total	**7,750**	**48,438**	**1,960**	**12,249**	**64**

6.2.1.1 Effet significatif du facteur Groupe

L'effet du facteur Groupe s'est révélé significatif ($F(3,60)=2098,488$; $p<0,001$; $\eta p^2=0,972$). Les comparaisons post-hoc ont été réalisées au moyen du *t*-test pour échantillons indépendants. La correction de Tukey HSD a été appliquée pour les comparaisons multiples de type inter-sujets. Les comparaisons post-hoc ont mis en évidence que la performance totale des adultes a été supérieure à celle des enfants ($t(15)=3,655$; $p<0,004$), des pré-adolescents ($t(15)=4,929$; $p<0,001$) et des adolescents ($t(15)=2,968$; $p<0,027$). (Voir la Figure 40 pour une représentation graphique des résultats.)

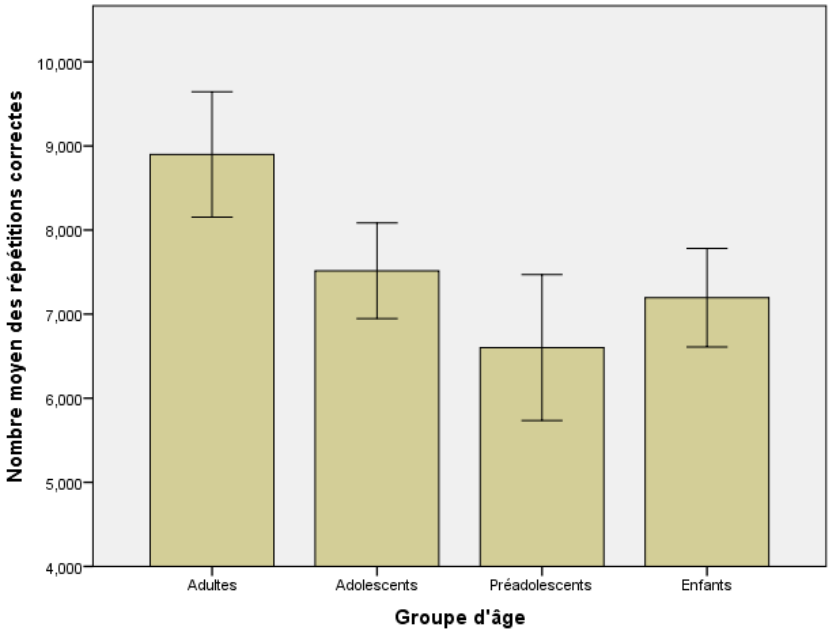

Figure 40. Variations du nombre moyen des répétitions correctes (performance totale moyenne) en fonction du groupe d'âge.

6.2.1.2 Effet significatif du facteur Format

L'effet du facteur Format s'est révélé significatif ($F(3,180)=136,431$; $p<0,001$; $\eta p^2=0,695$). Les comparaisons post-hoc ont été effectuées au moyen du *t*-test pour échantillons appariés. Elles ont mis en évidence que la performance totale pour les trois formats de présentation de l'information audio-visuelle a été supérieure à celle obtenue pour le format AU ((i) AU *vs* AVV ($t(63)=-16,037$; $p<0,001$) ; (ii) AU *vs* AVV-BA ($t(63)=-15,159$; $p<0,001$) ; (iii) AU *vs* AVV-EBA ($t(63)=-3,461$; $p<0,007$)). Par ailleurs, la performance totale pour les formats AVV et AVV-BA a été supérieure à celle obtenue pour le format AVV-EBA ((i) AVV *vs* AVV-EBA ($t(63)=12,596$; $p<0,001$) ; (ii) AVV-BA *vs* AVV-EBA ($t(63)=12,932$; $p<0,001$)). (Voir la Figure 41 pour une représentation graphique des résultats.)

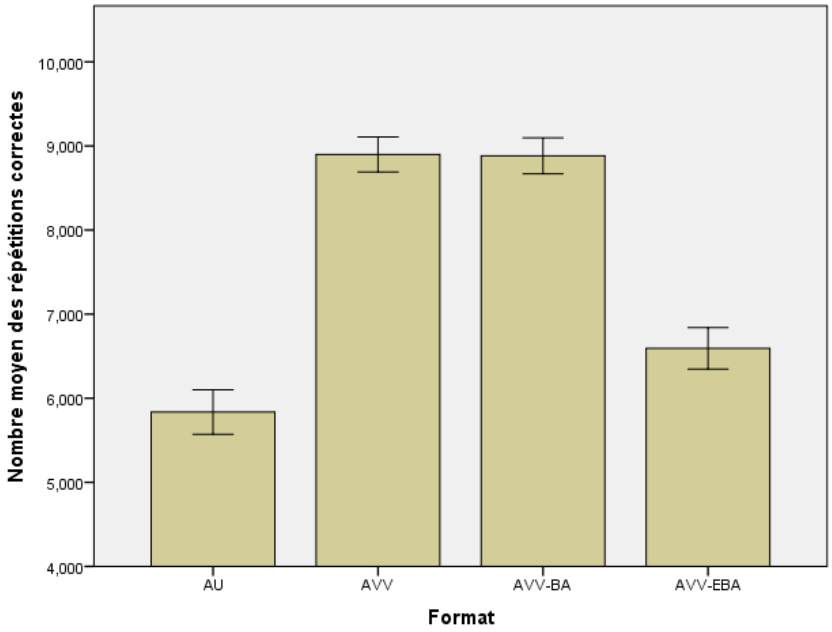

Figure 41. Variations du nombre moyen des répétitions correctes (performance totale moyenne) en fonction du format de présentation de l'information audio-visuelle.

6.2.1.3 Effet significatif du facteur SNR

L'effet du facteur SNR s'est révélé significatif ($F(1,60)=356,457$; $p<0,001$; $\eta p^2=0,856$). Globalement, la performance totale a été supérieure dans la condition du SNR-6 comparativement à celle du SNR-12. (Voir la Figure 42 pour une représentation graphique des résultats.)

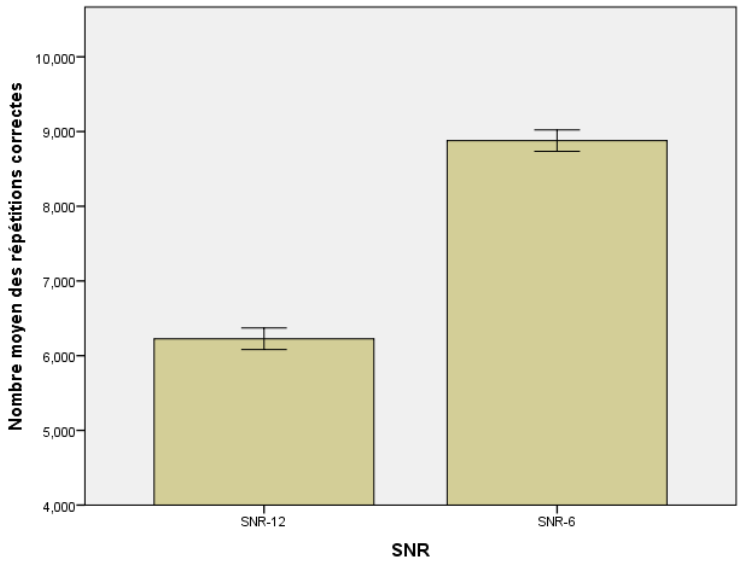

Figure 42. Variations du nombre moyen des répétitions correctes (performance totale moyenne) en fonction du degré de dégradation de l'information auditive (SNR).

6.2.1.4 Effet significatif de l'interaction Format x SNR

L'effet de l'interaction Format x SNR s'est révélé significatif ($F(3,180)=11,874$; $p<0,001$; $\eta p^2=0,165$). Les comparaisons post-hoc, réalisées au moyen du *t*-test pour échantillons appariés, ont mis en évidence que l'effet du facteur Format variait en fonction du facteur SNR de façon suivante : (i) Pour le SNR-12, le pattern des différences inter-formats était identique à celui propre à l'effet du facteur Format ((i) AU *vs* AVV ($t(63)=-14,973$; $p<0,001$) ; (ii) AU *vs* AVV-BA ($t(63)=-13,748$; $p<0,001$) ; (iii) AU *vs* AVV-EBA ($t(63)=-4,875$; $p<0,001$) ; (iv) AVV *vs* AVV-EBA ($t(63)=10,553$; $p<0,001$) ; (v) AVV-BA *vs* AVV-EBA ($t(63)=9,447$; $p<0,001$)). (ii) Pour le SNR-6, uniquement la différence dans la performance totale entre les formats AU et AVV-EBA n'a pas été significative. Les autres différences inter-formats correspondaient au pattern propre à l'effet du facteur Format ((i) AU *vs* AVV ($t(63)=-9,555$; $p<0,001$) ; (ii) AU *vs* AVV-BA ($t(63)=-11,615$; $p<0,001$) ; (iii) AVV *vs* AVV-EBA ($t(63)=9,394$; $p<0,001$) ; (iv) AVV-BA *vs* AVV-EBA ($t(63)=9,308$; $p<0,001$)). (Voir la Figure 43 pour une représentation graphique des résultats.)

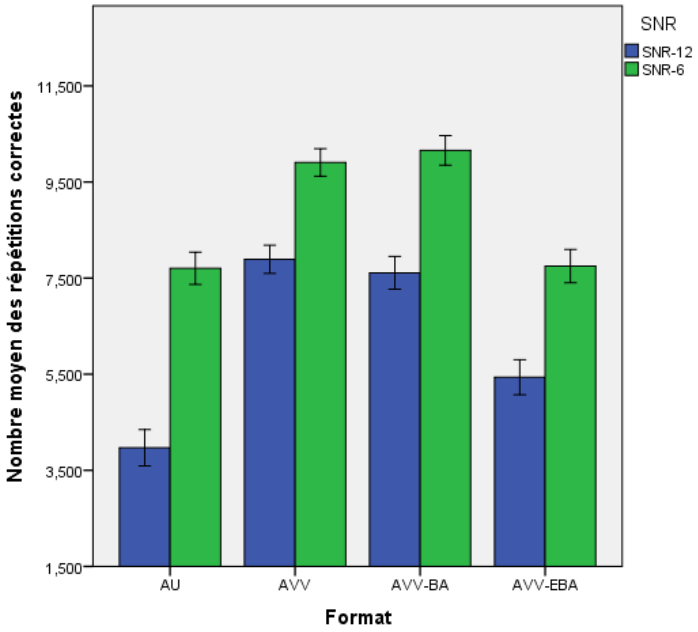

Figure 43. Variations du nombre moyen des répétitions correctes (performance totale moyenne) en fonction du format de présentation de l'information audio-visuelle et en fonction du degré de dégradation de l'information auditive (SNR).

6.2.2 Gain AV

Les données relatives au gain AV de l'expérience 1 ont été analysées au moyen de l'ANOVA à design mixte comportant 1 facteur de type inter-sujets, le Groupe d'âge, et 2 facteurs de type intra-sujets, le Format de présentation de l'information visuelle et le SNR. (Voir le Tableau 6 pour les moyennes et les écart-types pour le gain AV de l'expérience 2.)

Tableau 6. *Moyennes (M) et écarts-types (SD) pour le gain AV des participants des 4 groupes d'âge dans les différentes conditions expérimentales de l'expérience 2.*

	Groupe d'âge	*M*	*M [%]*	*SD*	*SD [%]*	*N*
	Adultes	4,438	27,734	2,250	14,063	16
	Adolescents	4,000	25,000	2,338	14,613	16
AVV_SNR-12	Pré-adolescents	3,438	21,484	2,065	12,904	16
	Enfants	3,813	23,828	1,759	10,997	16
	Total	**3,922**	**24,512**	**2,095**	**13,096**	**64**
	Adultes	2,813	17,578	1,759	10,997	16
	Adolescents	1,938	12,109	2,144	13,399	16
AVV_SNR-6	Pré-adolescents	2,625	16,406	1,746	10,915	16
	Enfants	1,438	8,984	1,504	9,401	16
	Total	**2,203**	**13,770**	**1,845**	**11,529**	**64**
	Adultes	3,750	23,438	2,720	17,002	16
	Adolescents	3,875	24,219	2,217	13,858	16
AVV-BA_SNR-12	Pré-adolescents	3,438	21,484	1,711	10,697	16
	Enfants	3,500	21,875	1,862	11,637	16
	Total	**3,641**	**22,754**	**2,118**	**13,240**	**64**
	Adultes	2,625	16,406	1,708	10,674	16
	Adolescents	2,063	12,891	1,806	11,289	16
AVV-BA_SNR-6	Pré-adolescents	2,500	15,625	1,673	10,458	16
	Enfants	2,625	16,406	1,668	10,427	16
	Total	**2,453**	**15,332**	**1,690**	**10,560**	**64**
	Adultes	1,813	11,328	3,124	19,528	16
	Adolescents	1,500	9,375	2,309	14,434	16
AVV-EBA_SNR-12	Pré-adolescents	1,625	10,156	2,156	13,477	16
	Enfants	,938	5,859	2,048	12,802	16

	Total	**1,469**	**9,180**	**2,410**	**15,063**	**64**
	Adultes	,750	4,688	2,082	13,010	16
	Adolescents	,000	,000	1,826	11,411	16
AVV-EBA_SNR-6	Pré-adolescents	1,063	6,641	1,482	9,261	16
	Enfants	-1,000	-6,250	2,556	15,975	16
	Total	**,203**	**1,270**	**2,132**	**13,325**	**64**

6.2.2.1 Effet significatif du facteur Format

L'effet du facteur Format s'est révélé significatif ($F(2,120)=113,402$; $p<0,001$; $\eta p^2=0,654$). Les comparaisons post-hoc ont été réalisées au moyen du *t*-test pour échantillon appariés. Elles ont mis en évidence que le gain AV a été supérieur pour le format AVV ($t(63)=14,083$; $p<0,001$) et pour le format AVV-BA ($t(63)=13,349$; $p<0,001$) comparativement au format AVV-EBA. (Voir la Figure 44 pour une représentation graphique des résultats.)

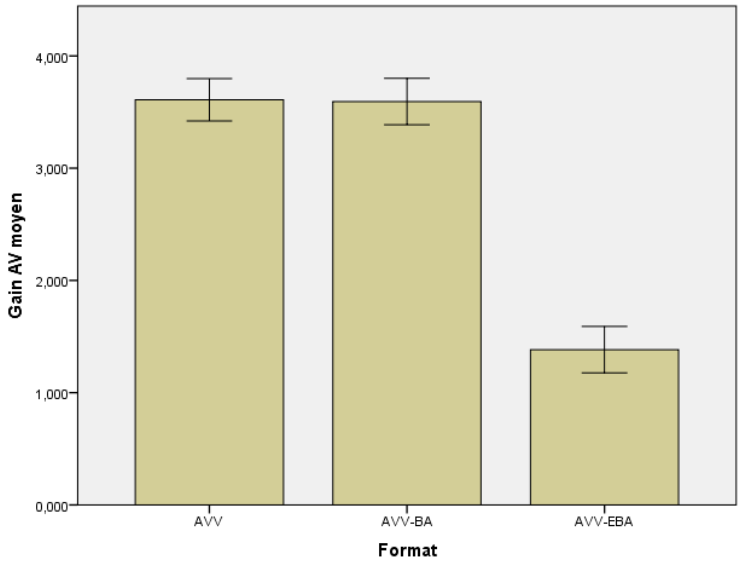

Figure 44. Variations du gain AV moyen en fonction du format de présentation de l'information visuelle.

6.2.2.2 Effet significatif du facteur SNR

L'effet du facteur SNR s'est révélé significatif ($F(1,60)=25,904$; $p<0,001$; $\eta p^2=0,302$). Le gain AV a été globalement supérieur dans la condition du SNR-12 comparativement à la condition du SNR-6. (Voir la Figure 45 pour une représentation graphique des résultats.)

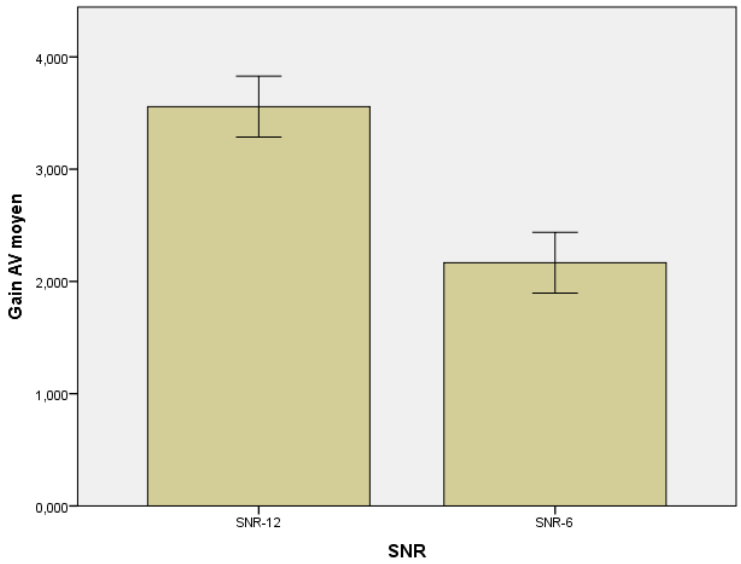

Figure 45. Variations du gain AV moyen en fonction du degré de dégradation de l'information auditive (SNR).

6.2.3 Comportement oculaire (durée des fixations oculaires)

La durée des fixations oculaires situées au niveau du visage de l'oratrice a été analysée au moyen de l'ANOVA à design mixte, comportant un facteur de type inter-sujets, le facteur Groupe d'âge, et trois facteurs de type intra-sujets, le facteur Format de présentation de l'information visuelle, le facteur SNR et le facteur AOI[43]. (Voir le Tableau 7 pour les moyennes et les écart-types pour les fixations oculaires situées au niveau du visage de l'oratrice de l'expérience 2.)

[43] Les données relatives au comportement oculaire (durée des fixations oculaires dans les différentes régions faciales) de l'expérience 2 présentaient la violation des mêmes suppositions de l'ANOVA que celles de l'expérience 1. (Voir la section 6.1.3 du présent chapitre pour la justification de l'utilisation de l'ANOVA dans notre cas de figure.)

Tableau 7. *Moyennes (M) et écarts-types (SD) pour la durée des fixations oculaires des participants des 4 groupes d'âge pour les différentes régions faciales de l'oratrice dans l'ensemble conditions expérimentales de l'expérience 2.*

	Groupe d'âge	*M*	*M [%]*	*SD*	*SD [%]*	*N*
	Adultes	1,273	84,889	,238	15,896	16
	Adolescents	1,158	77,183	,305	20,366	16
AVV_SNR-12_Bouche	Pré-adolescents	1,128	75,228	,294	19,582	16
	Enfants	,957	63,789	,439	29,270	16
	Total	**1,129**	**75,272**	**,339**	**22,630**	**64**
	Adultes	,026	1,733	,054	3,587	16
	Adolescents	,013	,862	,017	1,102	16
AVV_SNR-12_Contour	Pré-adolescents	,034	2,286	,068	4,545	16
	Enfants	,063	4,192	,136	9,083	16
	Total	**,034**	**2,268**	**,081**	**5,425**	**64**
	Adultes	,029	1,951	,051	3,423	16
	Adolescents	,051	3,401	,064	4,287	16
AVV_SNR-12_Nez	Pré-adolescents	,050	3,326	,056	3,711	16
	Enfants	,072	4,791	,071	4,758	16
	Total	**,051**	**3,367**	**,062**	**4,106**	**64**
	Adultes	,049	3,251	,113	7,555	16
	Adolescents	,014	,934	,026	1,759	16
AVV_SNR-12_Yeux	Pré-adolescents	,031	2,038	,046	3,086	16
	Enfants	,093	6,219	,257	17,125	16
	Total	**,047**	**3,110**	**,143**	**9,507**	**64**
	Adultes	,030	2,004	,039	2,613	16
	Adolescents	,070	4,651	,148	9,887	16
AVV_SNR-12_Intra-contour	Pré-adolescents	,048	3,200	,046	3,045	16
	Enfants	,140	9,325	,227	15,126	16

	Total	**,072**	**4,795**	**,142**	**9,456**	**64**
	Adultes	1,260	83,997	,168	11,210	16
	Adolescents	1,057	70,438	,339	22,626	16
AVV_SNR-6_Bouche	Pré-adolescents	,990	65,972	,408	27,225	16
	Enfants	1,023	68,171	,354	23,599	16
	Total	**1,082**	**72,144**	**,339**	**22,605**	**64**
	Adultes	,017	1,148	,036	2,376	16
	Adolescents	,019	1,260	,029	1,952	16
AVV_SNR-6_Contour	Pré-adolescents	,032	2,116	,057	3,781	16
	Enfants	,055	3,656	,105	7,002	16
	Total	**,031**	**2,045**	**,064**	**4,284**	**64**
	Adultes	,030	2,014	,035	2,332	16
	Adolescents	,090	5,990	,111	7,425	16
AVV_SNR-6_Nez	Pré-adolescents	,120	7,971	,133	8,857	16
	Enfants	,090	5,990	,096	6,380	16
	Total	**,082**	**5,491**	**,103**	**6,896**	**64**
	Adultes	,030	2,000	,046	3,048	16
	Adolescents	,043	2,893	,065	4,362	16
AVV_SNR-6_Yeux	Pré-adolescents	,034	2,259	,048	3,208	16
	Enfants	,130	8,672	,270	17,988	16
	Total	**,059**	**3,956**	**,145**	**9,688**	**64**
	Adultes	,041	2,706	,052	3,437	16
	Adolescents	,081	5,396	,149	9,930	16
AVV_SNR-6_Intra-contour	Pré-adolescents	,061	4,087	,059	3,925	16
	Enfants	,079	5,278	,061	4,081	16
	Total	**,066**	**4,367**	**,089**	**5,927**	**64**
	Adultes	1,286	85,733	,172	11,466	16
AVV-BA_SNR-12_Bouche	Adolescents	1,066	71,060	,347	23,146	16
	Pré-adolescents	1,065	71,014	,340	22,669	16

	Enfants	1,029	68,597	,248	16,539	16
	Total	**1,112**	**74,101**	**,297**	**19,828**	**64**
	Adultes	,079	5,254	,159	10,621	16
	Adolescents	,036	2,377	,096	6,433	16
AVV-BA_SNR-12_Contour	Pré-adolescents	,072	4,817	,136	9,063	16
	Enfants	,079	5,242	,132	8,785	16
	Total	**,066**	**4,423**	**,131**	**8,723**	**64**
	Adultes	,024	1,570	,039	2,618	16
	Adolescents	,067	4,444	,080	5,324	16
AVV-BA_SNR-12_Nez	Pré-adolescents	,069	4,589	,064	4,271	16
	Enfants	,064	4,250	,057	3,783	16
	Total	**,056**	**3,713**	**,063**	**4,207**	**64**
	Adultes	,003	,174	,005	,320	16
	Adolescents	,031	2,048	,041	2,721	16
AVV-BA_SNR-12_Yeux	Pré-adolescents	,021	1,386	,027	1,818	16
	Enfants	,032	2,125	,042	2,791	16
	Total	**,022**	**1,433**	**,034**	**2,247**	**64**
	Adultes	,017	1,148	,024	1,585	16
	Adolescents	,053	3,540	,061	4,080	16
AVV-BA_SNR-12_Intra-contour	Pré-adolescents	,051	3,429	,050	3,304	16
	Enfants	,075	5,008	,099	6,609	16
	Total	**,049**	**3,281**	**,066**	**4,415**	**64**
	Adultes	1,310	87,326	,112	7,447	16
	Adolescents	1,016	67,718	,360	23,972	16
AVV-BA_SNR-6_Bouche	Pré-adolescents	1,098	73,194	,357	23,805	16
	Enfants	1,106	73,736	,226	15,047	16
	Total	**1,132**	**75,494**	**,297**	**19,796**	**64**
	Adultes	,028	1,884	,046	3,069	16
AVV-BA_SNR-6_Contour	Adolescents	,020	1,350	,036	2,429	16

	Pré-adolescents	,041	2,703	,065	4,335	16
	Enfants	,024	1,586	,024	1,602	16
	Total	**,028**	**1,881**	**,045**	**3,000**	**64**
	Adultes	,038	2,537	,045	2,999	16
	Adolescents	,075	4,997	,097	6,474	16
AVV-BA_SNR-6_Nez	Pré-adolescents	,064	4,242	,068	4,547	16
	Enfants	,098	6,550	,100	6,677	16
	Total	**,069**	**4,581**	**,082**	**5,457**	**64**
	Adultes	,009	,582	,014	,962	16
	Adolescents	,046	3,078	,061	4,057	16
AVV-BA_SNR-6_Yeux	Pré-adolescents	,031	2,086	,041	2,733	16
	Enfants	,041	2,744	,094	6,288	16
	Total	**,032**	**2,123**	**,060**	**4,033**	**64**
	Adultes	,014	,923	,020	1,350	16
	Adolescents	,080	5,342	,085	5,638	16
AVV-BA_SNR-6_Intra-contour	Pré-adolescents	,045	3,028	,041	2,729	16
	Enfants	,082	5,467	,077	5,166	16
	Total	**,055**	**3,690**	**,067**	**4,436**	**64**
	Adultes	1,115	74,307	,306	20,401	16
	Adolescents	1,048	69,864	,312	20,807	16
AVV-EBA_SNR-12_Bouche	Pré-adolescents	,977	65,151	,372	24,795	16
	Enfants	,811	54,054	,323	21,564	16
	Total	**,988**	**65,844**	**,341**	**22,738**	**64**
	Adultes	,098	6,506	,153	10,202	16
	Adolescents	,031	2,047	,046	3,073	16
AVV-EBA_SNR-12_Contour	Pré-adolescents	,134	8,917	,191	12,759	16
	Enfants	,099	6,598	,116	7,732	16
	Total	**,090**	**6,017**	**,139**	**9,290**	**64**
AVV-EBA_SNR-12_Nez	Adultes	,042	2,786	,090	6,032	16

	Adolescents	,096	6,369	,085	5,653	16
	Pré-adolescents	,081	5,367	,118	7,871	16
	Enfants	,164	10,958	,141	9,378	16
	Total	**,096**	**6,370**	**,117**	**7,799**	**64**
	Adultes	,010	,667	,018	1,184	16
	Adolescents	,051	3,425	,068	4,552	16
AVV-EBA_SNR-12_Yeux	Pré-adolescents	,029	1,951	,042	2,820	16
	Enfants	,091	6,054	,212	14,137	16
	Total	**,045**	**3,024**	**,115**	**7,669**	**64**
	Adultes	,074	4,924	,089	5,903	16
	Adolescents	,063	4,214	,056	3,732	16
AVV-EBA_SNR-12_Intra-contour	Pré-adolescents	,080	5,331	,061	4,088	16
	Enfants	,117	7,807	,079	5,295	16
	Total	**,084**	**5,569**	**,074**	**4,912**	**64**
	Adultes	1,112	74,124	,259	17,281	16
	Adolescents	,999	66,588	,232	15,498	16
AVV-EBA_SNR-6_Bouche	Pré-adolescents	,879	58,625	,408	27,203	16
	Enfants	,760	50,691	,316	21,076	16
	Total	**,938**	**62,507**	**,331**	**22,091**	**64**
	Adultes	,059	3,947	,119	7,902	16
	Adolescents	,056	3,764	,087	5,770	16
AVV-EBA_SNR-6_Contour	Pré-adolescents	,111	7,422	,133	8,842	16
	Enfants	,156	10,367	,125	8,317	16
	Total	**,096**	**6,375**	**,121**	**8,089**	**64**
	Adultes	,089	5,959	,108	7,199	16
	Adolescents	,104	6,964	,095	6,359	16
AVV-EBA_SNR-6_Nez	Pré-adolescents	,112	7,441	,109	7,257	16
	Enfants	,190	12,700	,199	13,275	16
	Total	**,124**	**8,266**	**,137**	**9,133**	**64**

	Adultes	,020	1,315	,024	1,573	16
	Adolescents	,047	3,122	,063	4,189	16
AVV-EBA_SNR-6_Yeux	Pré-adolescents	,041	2,740	,049	3,243	16
	Enfants	,048	3,228	,121	8,097	16
	Total	**,039**	**2,601**	**,073**	**4,845**	**64**
	Adultes	,038	2,525	,047	3,152	16
	Adolescents	,074	4,917	,055	3,648	16
AVV-EBA_SNR-6_Intra-contour	Pré-adolescents	,067	4,461	,077	5,157	16
	Enfants	,133	8,889	,104	6,903	16
	Total	**,078**	**5,198**	**,080**	**5,352**	**64**

6.2.3.1 Effet significatif du facteur Format

L'effet du facteur Format s'est révélé significatif ($F(2,120)=3,945$; $p<0,023$; $\eta p^2=0,062$). Les comparaisons post-hoc ont été réalisées par le *t*-test pour échantillons appariés. Elles ont mis en évidence que la durée des fixations dans les différentes régions faciales a été plus longue pour le format AVV que pour le format AVV-EBA ($t(63)=2,667$; $p<0,018$). (Ceci implique que les participants ont fixé des régions en dehors du visage de l'oratrice plus longuement dans la condition AVV-EBA que dans la condition AVV.) (Voir la Figure 46 pour une représentation graphique des résultats.)

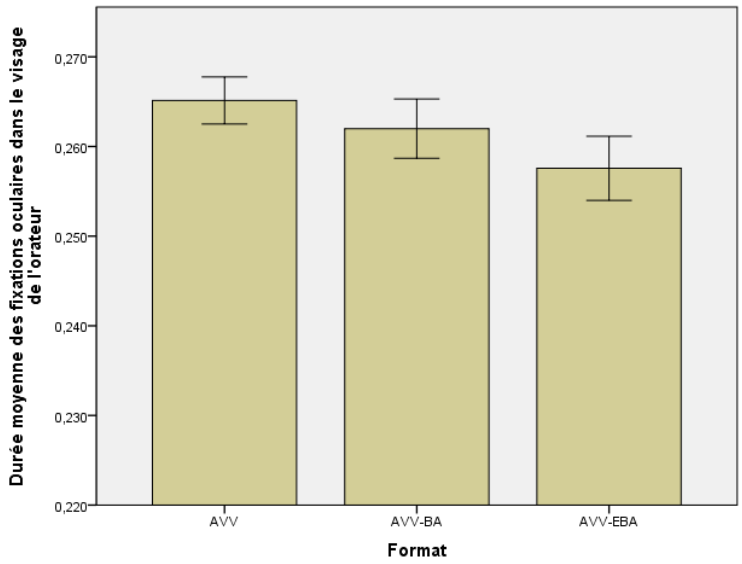

Figure 46. Variations de la durée moyenne (en secondes) des fixations oculaires au niveau du visage de l'oratrice obtenue en fonction du format de présentation de l'information visuelle.

6.2.3.2 Effet significatif du facteur AOI

L'effet du facteur AOI s'est révélé significatif ($F(4,240)=698,401$; $p<0,001$; $\eta p^2=0,921$). Les comparaisons post-hoc ont été effectuées au moyen du *t*-test pour échantillons appariés. Elles ont mis en évidence que la région buccale de l'oratrice a été fixées plus longtemps que son nez ($t(63)=27,333$; $p<0,001$), ses yeux ($t(63)=27,689$; $p<0,001$), l'intra-contour ($t(63)=27,667$; $p<0,001$) et le contour ($t(63)=28,743$; $p<0,001$) de son visage. Par ailleurs, le nez a été fixé plus longtemps que les yeux ($t(63)=4,875$; $p<0,001$) et les yeux moins longtemps que l'intra-contour ($t(63)=-3,000$; $p<0,052$). Cette dernière différence dépassait cependant légèrement le seuil critique de signification. (Voir la Figure 47 pour une représentation graphique des résultats.)

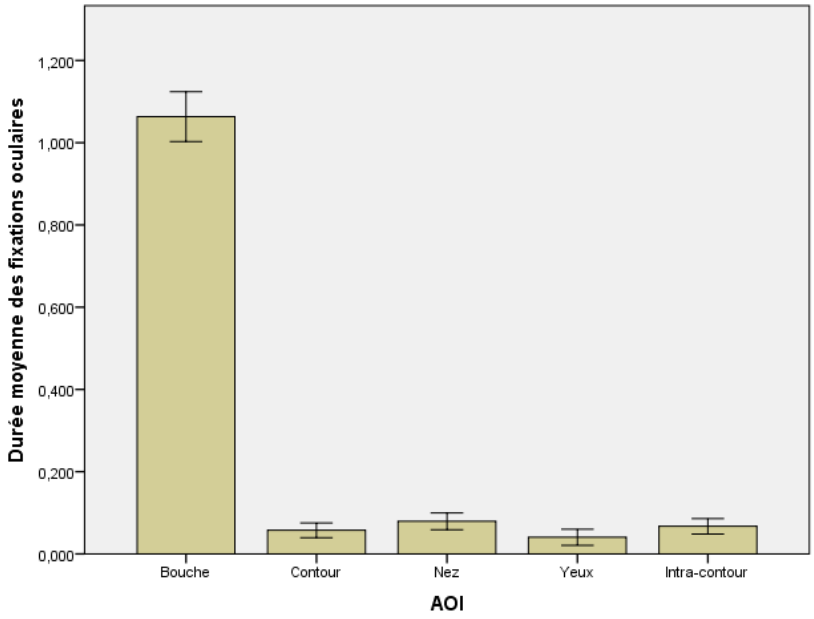

Figure 47. Variations de la durée moyenne (en secondes) des fixations oculaires en fonction des différentes régions faciales de l'oratrice (AOI).

6.2.3.3 Effet significatif de l'interaction Groupe x AOI

L'effet de l'interaction Groupe x AOI s'est avéré significatif ($F(12,177)=3,478$; $p<0,014$; $\eta p^2=0,148$). Les comparaisons post-hoc intra-groupe ont été réalisées au moyen du *t*-test pour échantillons appariés. Elles ont mis en évidence les différences significatives dans la durée des fixations oculaires suivantes : (i) Dans le groupe « adultes », la bouche a été fixée plus longtemps que le nez ($t(15)=27,086$; $p<0,001$), les yeux ($t(15)=28,442$; $p<0,001$), l'intra-contour ($t(15)=26,187$; $p<0,001$) et le contour ($t(15)=23,132$; $p<0,001$). Par ailleurs, l'intra-contour a été fixé plus longtemps que les yeux ($t(15)=-2,155$; $p<0,049$). (ii) Dans le groupe « adolescents », la bouche a été fixée plus longtemps que le nez ($t(15)=13,799$; $p<0,001$), les yeux ($t(15)=14,443$; $p<0,001$), l'intra-contour ($t(15)=12,881$; $p<0,001$) et le contour ($t(15)=14,421$; $p<0,001$). Le nez a été fixé plus longtemps que le contour ($t(15)=2,573$; $p<0,022$) et les yeux ($t(15)=2,631$; $p<0,020$). Finalement, l'intra-contour a été fixé plus longtemps que les yeux ($t(15)=-2,158$; $p<0,049$) et que le contour ($t(15)=-2,916$; $p<0,012$). (iii) Dans le groupe « pré-adolescents », la bouche a été fixée plus longtemps que le nez ($t(15)=10,240$; $p<0,001$), les yeux ($t(15)=11,526$; $p<0,001$), l'intra-contour ($t(15)=10,908$; $p<0,001$) et le contour ($t(15)=10,882$; $p<0,001$). Par ailleurs, les yeux ont été fixés moins longtemps que le nez ($t(15)=-4,019$; $p<0,001$) et l'intra-contour ($t(15)=-4,129$; $p<0,002$). (iv)

Dans le groupe « enfants », la bouche a été fixée plus longtemps que le nez ($t(15)=10,850$; $p<0,001$), les yeux ($t(15)=10,177$; $p<0,001$), l'intra-contour ($t(15)=11,573$; $p<0,001$) et le contour ($t(15)=14,012$; $p<0,001$). Aucune autre différence inter-AOIs n'a été significative dans ce groupe. Pour finir, les comparaisons post-hoc intergroupes pour chaque AOI ont été menées. Elles ont été réalisées au moyen de *t*-test pour échantillons indépendants avec l'application de la correction de Tukey HSD pour comparaisons multiples. Ces comparaisons ont mis en évidence que la bouche a été fixée plus longtemps dans le groupe « adultes » que dans le groupe « enfants » ($t(15)=3,033$; $p<0,022$). En revanche, l'intra-contour a été globalement fixé plus longtemps dans le groupe « enfants » que dans le groupe « adultes » ($t(15)=-3,496$; $p<0,005$). (Voir la Figure 48 pour une représentation graphique des résultats.)

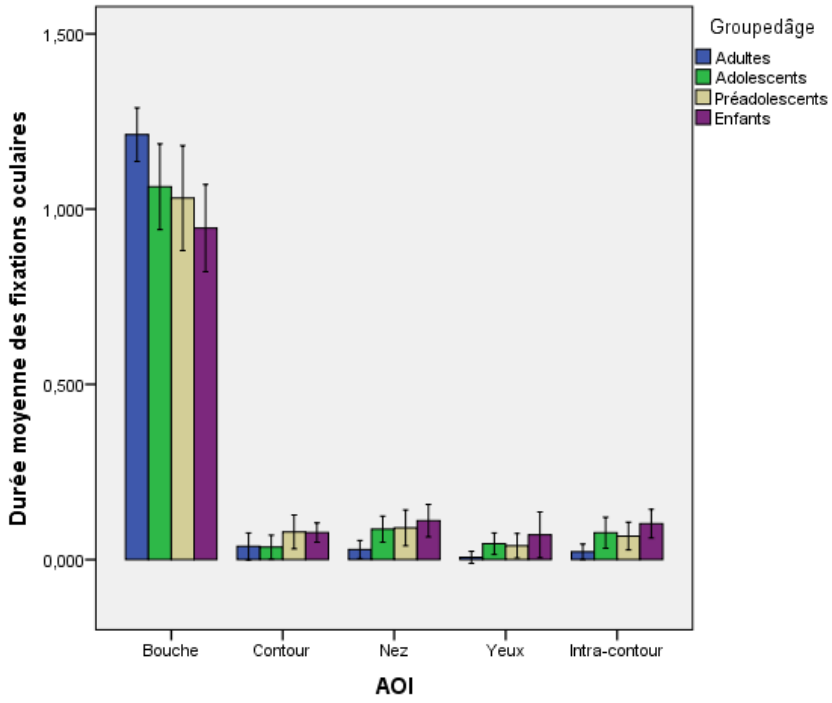

Figure 48. Variations de la durée moyenne (en secondes) des fixations oculaires dans les différentes régions faciales de l'oratrice (AOI) en fonction du groupe d'âge.

6.2.3.4 Effet significatif de l'interaction Format x AOI

L'effet de l'interaction Format x AOI s'est révélé significatif ($F(8,480)=16,187$; $p<0,001$; $\eta p^2=0,212$). Les comparaisons post-hoc ont été réalisées au moyen du *t*-test pour échantillons appariés. Quant aux comparaisons inter-AOIs pour l'ensemble des combinaisons

de deux formats de présentation de l'information visuelle, les patterns suivants ont été identifiés : (i) Dans la condition AVV, la bouche a été fixée plus longtemps que le nez (t(63)=25,486 ; p<0,001), les yeux (t(63)=24,030 ; p<0,001), l'intra-contour (t(63)=23,348 ; p<0,001) et le contour (t(63)=26,464 ; p<0,001). Par ailleurs, le contour a été fixé moins longtemps que le nez (t(63)=-3,218 ; p<0,003) et que l'intra-contour (t(63)=-2,993 ; p<0,005). (ii) Dans la condition AVV-BA, la bouche a été fixée plus longtemps que le nez (t(63)=26,855 ; p<0,001), les yeux (t(63)=28,498 ; p<0,001), l'intra-contour (t(63)=26,729 ; p<0,001) et le contour (t(63)=27,801 ; p<0,001). En outre, la région des yeux s'est révélée être fixée moins longtemps que le contour (t(63)=-2,032 ; p<0,046), le nez (t(63)=-5,320 ; p<0,001) et l'intra-contour (t(63)=-3,383 ; p<0,002). (iii) Dans la condition AVV-EBA, la bouche a été de nouveau fixée plus longtemps que les autres régions faciales, le nez (t(63)=17,262 ; p<0,001), les yeux (t(63)=19,814 ; p<0,001), l'intra-contour (t(63)=19,256 ; p<0,001) et le contour (t(63)=18,389 ; p<0,001). Le nez a été fixé plus longtemps que les yeux (t(63)=5,275 ; p<0,001) et l'intra-contour (t(63)=2,325 ; p<0,024). Finalement, les yeux ont été fixés moins longtemps que l'intra-contour (t(63)=-3,222 ; p<0,003) et le contour (t(63)=-2,558 ; p<0,013).

Les comparaisons intra-formats pour l'ensemble des combinaisons de deux AOIs ont également été effectuées. Les résultats ont révélé que (i) la bouche a été fixée plus longtemps avec les formats AVV (t(63)=5,064 ; p<0,001) et AVV-BA (t(63)=5,125 ; p<0,001) qu'avec le format AVV-EBA ; (ii) les yeux ont été fixés plus longtemps avec le format AVV qu'avec le format AVV-EBA (t(63)=2,424 ; p<0,019) ; (iii) l'intra-contour a été fixé plus longtemps avec le format AVV-BA qu'avec le format AVV-EBA (t(63)=-3,635 ; p<0,002) ; (iv) le nez a été fixé plus longtemps avec le format AVV-EBA qu'avec les deux autres formats, AVV (t(63)=3,211 ; p<0,003) et AVV-BA (t(63)=4,407 ; p<0,001) ; (v) le contour a été fixé plus longtemps avec le format AVV-EBA qu'avec les deux autres formats, AVV (t(63)=4,123 ; p<0,001) et AVV-BA (t(63)=2,840 ; p<0,007). (Voir la Figure 49 pour une représentation graphique des résultats.)

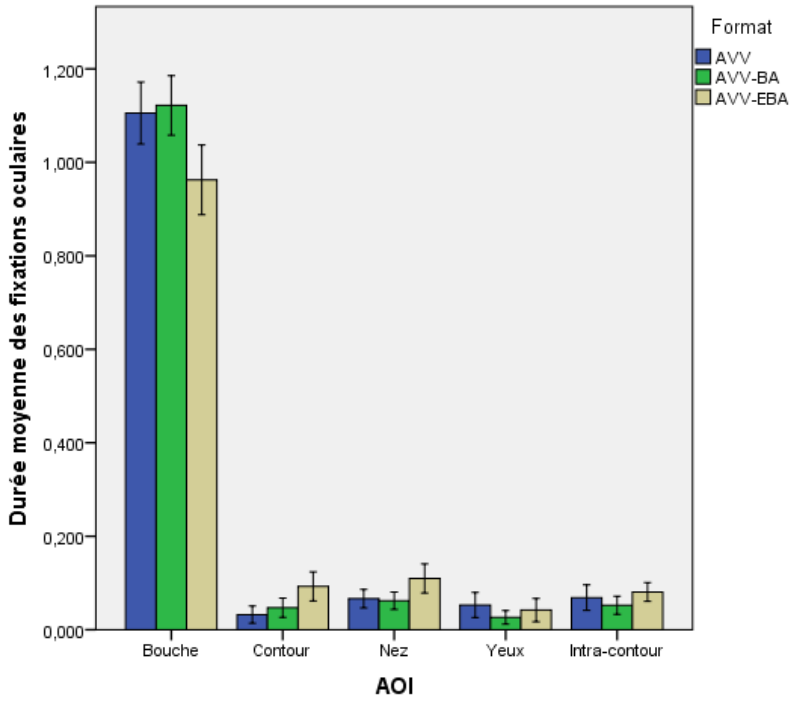

Figure 49. Variations de la durée moyenne (en secondes) des fixations oculaires dans les différentes régions faciales de l'oratrice (AOI) en fonction du format de présentation de l'information visuelle.

6.2.3.5 Effet significatif de l'interaction SNR x AOI

L'effet de l'interaction SNR x AOI s'est révélé significatif ($F(4,240)=2,505$; $p<0,044$; $\eta p^2=0,040$). Les comparaisons post-hoc ont été réalisées au moyen du *t*-test pour échantillons appariés. Les comparaisons intra-SNR pour l'ensemble des combinaisons de deux AOIs ont été effectuées en premier. Les résultats montrent les patterns suivants : (i) Pour le SNR-12, le pattern classique a été trouvé avec la région buccale qui a été fixée plus longtemps que le nez ($t(63)=26,241$; $p<0,001$), les yeux ($t(63)=26,648$; $p<0,001$), l'intra-contour ($t(63)=25,103$; $p<0,001$) et le contour ($t(63)=27,028$; $p<0,001$). Pour ce degré de dégradation de l'information auditive, les yeux ont été fixés moins longtemps que le nez ($t(63)=-3,563$; $p<0,002$), l'intra-contour ($t(63)=-3,374$; $p<0,002$) et le contour ($t(63)=-2,155$; $p<0,036$). (ii) Pour le SNR-6, la bouche s'est également révélée être fixée plus longtemps que les autres régions faciales, le nez ($t(63)=22,421$; $p<0,001$), les yeux ($t(63)=23,764$; $p<0,001$), l'intra-contour ($t(63)=24,194$; $p<0,001$) et le contour ($t(63)=25,601$; $p<0,001$). En outre, le nez a été fixé plus longtemps que les yeux ($t(63)=5,602$; $p<0,001$), l'intra-contour ($t(63)=2,414$; $p<0,020$) et le contour

(t(63)=2,774 ; p<0,008). Finalement, les yeux ont été fixés moins longtemps que l'intra-contour (t(63)=-3,222 ; p<0,003).

Dans un deuxième temps, les comparaisons intra-AOIs pour les deux SNRs ont été effectuées. Les résultats ont révélé que le nez a été globalement fixé moins longtemps dans la condition du SNR-12 que dans la condition du SNR-6 (t(63)=-3,526 ; p<0,002). (Voir la Figure 50 pour une représentation graphique des résultats.)

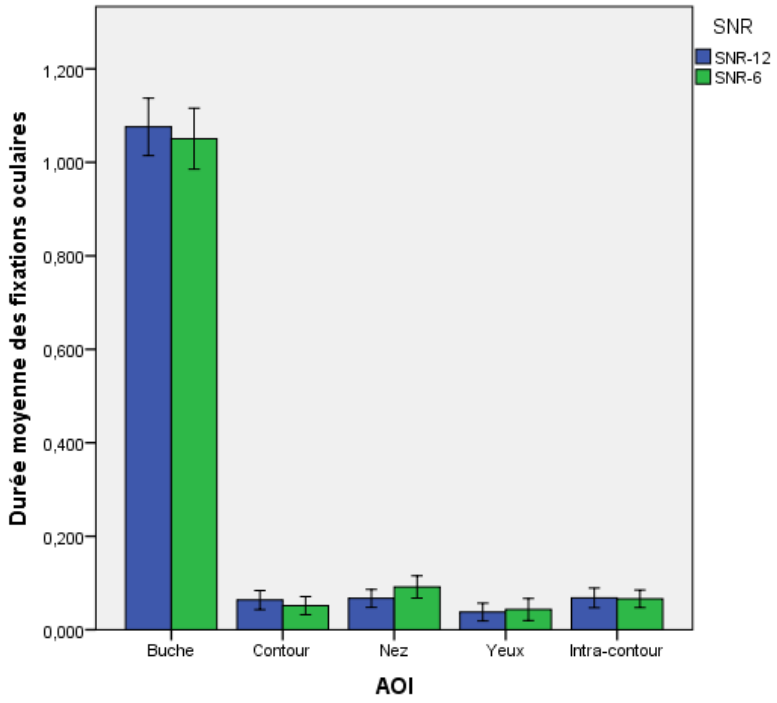

Figure 50. Variations de la durée moyenne (en secondes) des fixations oculaires dans les différentes régions faciales de l'oratrice (AOI) en fonction du degré de dégradation de l'information auditive (SNR).

6.2.3.6 Effet significatif de l'interaction Groupe x Format x AOI

L'effet de l'interaction Groupe x Format x AOI s'est révélé significatif (F(24,1440)=1,711 ; p<0,021 ; ηp^2=0,040). Les comparaisons post-hoc ont été faites en trois temps comme suit : (i) entre les différents groupes pour l'ensemble des conditions combinant un format et une AOI ; (ii) entre les différentes AOIs pour l'ensemble des conditions combinant un groupe d'âge et un format de présentation de l'information visuelle ; (iii) entre les différents formats de présentation de l'information visuelle pour l'ensemble des conditions combinant un

groupe d'âge et une AOI. Les comparaisons post-hoc pour le cas de figure (i) ont été réalisées au moyen du *t*-test pour échantillons indépendants avec l'application de la correction de Tukey HSD pour comparaisons multiples. Les comparaisons post-hoc pour les cas de figure (ii) et (iii) ont été réalisées au moyen du *t*-test pour échantillons appariés. (Voir la Figure 51, la Figure 52 et la Figure 53 pour une représentation graphique des résultats.)

Pour le premier cas de figure, les comparaisons post-hoc ont mis en évidence que la bouche a été fixée plus longtemps dans le groupe « adultes » que dans le groupe « enfants » dans la condition du format AVV ($t(15)=2,781$; $p<0,036$). En outre, dans la condition du format AVV-EBA, l'intra-contour a été fixé plus longtemps dans le groupe « enfants » que dans le groupe « adultes » ($t(15)=3,048$; $p<0,019$).

Pour le deuxième cas de figure, les résultats des comparaisons post-hoc ont mis en évidence que, dans le groupe « adultes », les yeux ont été fixés plus longtemps dans les conditions AVV ($t(15)=2,248$; $p<0,041$) et AVV-BA ($t(15)=2,341$; $p<0,034$) que dans la condition AVV-EBA. Un pattern semblable été observé dans le groupe « enfants » (AVV *vs* AVV-EBA ($t(15)=2,143$; $p<0,050$) et AVV-BA *vs* AVV-EBA ($t(15)=1,968$; $p<0,068$)). Cependant, dans ce dernier groupe, la probabilité que la différence entre les formats AVV-BA et AVV-EBA soit due à une variation aléatoire a été un peu supérieure au seuil critique de signification. Dans le groupe « adolescents », la région buccale a été fixée plus longtemps avec le format AVV qu'avec le format AVV-BA ($t(15)=2,766$; $p<0,015$). Finalement, dans le groupe « pré-adolescents », la bouche a été fixée plus longtemps avec le format AVV-BA qu'avec le format AVV-EBA $t(15)=-2,532$; $p<0,024$).

Pour le troisième cas de figure, les comparaisons post-hoc ont mis en évidence les patterns des résultats significatifs suivants :

(I) Pour le groupe « adultes » : (i) Dans la condition du format AVV, la bouche a été fixée plus longtemps que les autres régions faciales, le nez ($t(15)=25,661$; $p<0,001$), les yeux ($t(15)=22,525$; $p<0,001$), l'intra-contour ($t(15)=23,395$; $p<0,001$) et le contour ($t(15)=24,195$; $p<0,001$). (ii) Le même pattern a été observé dans la condition du format AVV-BA (Bouche *vs* Nez ($t(15)=36,342$; $p<0,001$), Bouche *vs* Yeux ($t(15)=36,651$; $p<0,001$), Bouche *vs* Intra-contour ($t(15)=31,347$; $p<0,001$), Bouche *vs* Contour ($t(15)=24,922$; $p<0,001$)). (iii) Dans la condition du format AVV-EBA, la bouche a également été fixée plus longtemps que le nez ($t(15)=36,769$; $p<0,001$), les yeux ($t(15)=41,741$; $p<0,001$), l'intra-contour ($t(15)=37,667$; $p<0,001$) et le contour

(t(15)=25,899 ; p<0,001). En outre, les yeux ont été fixés moins longtemps que le nez (t(15)=-2,569 ; p<0,022) et le contour (t(15)=2,173 ; p<0,047). La différence entre les yeux et l'intra-contour a également été relativement proche du seuil de signification (t(15)=-1,902 ; p<0,077), les yeux étant fixés moins longtemps que l'intra-contour.

(II) Pour le groupe « adolescents » : (i) Dans la condition du format AVV, la bouche a été fixée plus longtemps que le nez (t(15)=12,680 ; p<0,001), les yeux (t(15)=13,761 ; p<0,001), l'intra-contour (t(15)=10,216 ; p<0,001) et le contour (t(15)=13,690 ; p<0,001). En outre, le nez a été fixé plus longtemps que les yeux (t(15)=2,359 ; p<0,033) et le contour (t(15)=2,944 ; p<0,011). (ii) Dans la condition du format AVV-BA, la bouche a été fixée plus longtemps que le nez (t(15)=11,312 ; p<0,001), les yeux (t(15)=12,241 ; p<0,001), l'intra-contour (t(15)=10,309 ; p<0,001) et le contour (t(15)=11,491 ; p<0,001). De plus, le contour a été fixé moins longtemps que l'intra-contour (t(15)=-3,438 ; p<0,005). La différence entre le contour et le nez a également été relativement proche du seuil de signification (t(15)=-1,937 ; p<0,073), le nez étant fixé plus longtemps que le contour. (iii) Dans la condition du format AVV-EBA, la bouche a été fixée plus longtemps que le nez (t(15)=11,005 ; p<0,001), les yeux (t(15)=11,174 ; p<0,001), l'intra-contour (t(15)=10,381 ; p<0,001) et le contour (t(15)=10,857 ; p<0,001). Par ailleurs, l'intra-contour a été fixé plus longtemps que les yeux (t(15)=2,335 ; p<0,0035) et le contour (t(15)=3,509 ; p<0,004).

(III) Pour le groupe « préadolescents » : (i) Dans la condition du format AVV, la bouche a été fixée plus longtemps que le nez (t(15)=10,292 ; p<0,001), les yeux (t(15)=11,820 ; p<0,001), l'intra-contour (t(15)=11,299 ; p<0,001) et le contour (t(15)=11,391 ; p<0,001). De plus, le nez a été fixé plus longtemps que le contour (t(15)=2,822 ; p<0,014). La différence entre le nez et l'intra-contour a également été très proche du seuil de signification (t(15)=2,108 ; p<0,052), le nez étant fixé plus longtemps que l'intra-contour. (ii) Dans la condition du format AVV-BA, la bouche a été fixée plus longtemps que le nez (t(15)=9,436 ; p<0,001), les yeux (t(15)=11,060 ; p<0,001), l'intra-contour (t(15)=10,243 ; p<0,001) et le contour (t(15)=10,367 ; p<0,001). Le nez a été fixé plus longtemps que les yeux (t(15)=4,084 ; p<0,002) et l'intra-contour (t(15)=2,722 ; p<0,017), et les yeux moins longtemps que l'intra-contour (t(15)=-2,741 ; p<0,016). (iii) Dans la condition du format AVV-EBA, la bouche a été fixée plus longtemps que le nez (t(15)=11,179 ; p<0,001), les yeux (t(15)=12,026 ; p<0,001), l'intra-contour (t(15)=11,533 ; p<0,001) et le contour (t(15)=11,266 ; p<0,001). En

outre, les yeux ont été fixés moins longtemps que le nez (t(15)=-4,328 ; p<0,002) et l'intra-contour (t(15)=-3,735 ; p<0,003).

(IV) Pour le groupe « enfants » : Seul le pattern standard, avec la bouche qui a été fixée plus longtemps que chaque autre région faciale, a été mis en évidence pour les trois formats de présentation de l'information visuelle. (i) Dans la condition du format AVV : Bouche *vs* Nez (t(15)=11,790 ; p<0,001), Bouche *vs* Yeux (t(15)=8,214; p<0,001), Bouche *vs* Intra-contour (t(15)=9,820 ; p<0,001), Bouche *vs* Contour (t(15)=11,104 ; p<0,001). (ii) Dans la condition du format AVV-BA : Bouche *vs* Nez (t(15)=1,169 ; p<0,001), Bouche *vs* Yeux (t(15)=9,275 ; p<0,001), Bouche *vs* Intra-contour (t(15)=11,607 ; p<0,001), Bouche *vs* Contour (t(15)=12,281 ; p<0,001). (iii) Dans la condition du format AVV-EBA : Bouche *vs* Nez (t(15)=13,868 ; p<0,001), Bouche *vs* Yeux (t(15)=15,261 ; p<0,001), Bouche *vs* Intra-contour (t(15)=14,325 ; p<0,001), Bouche *vs* Contour (t(15)=17,838 ; p<0,001).

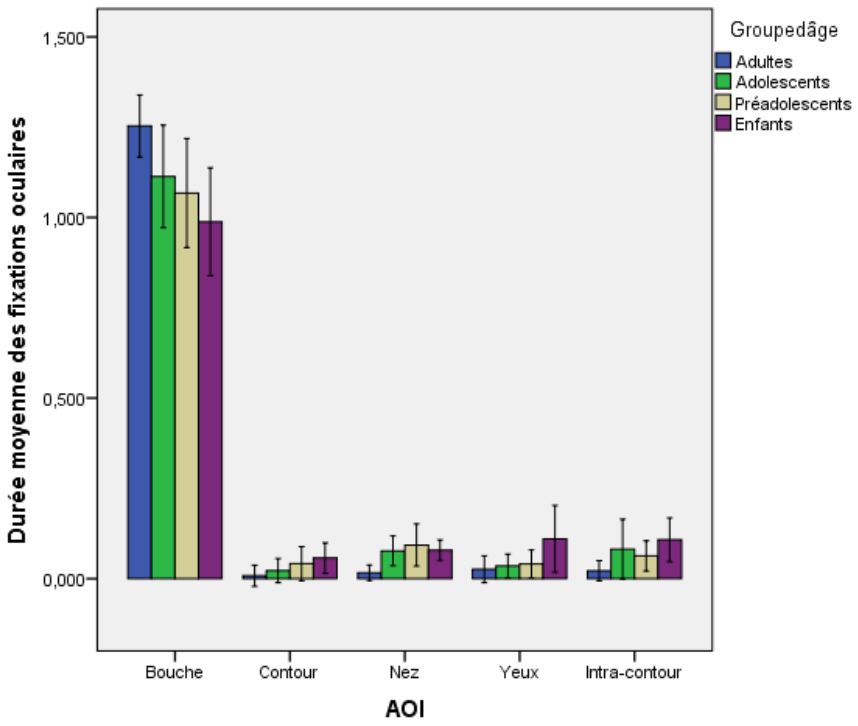

Figure 51. Variations de la durée moyenne(en secondes) des fixations oculaires dans la condition du format AVV en fonction des différentes régions faciales de l'oratrice (AOI) et en fonction du degré de dégradation du groupe d'âge.

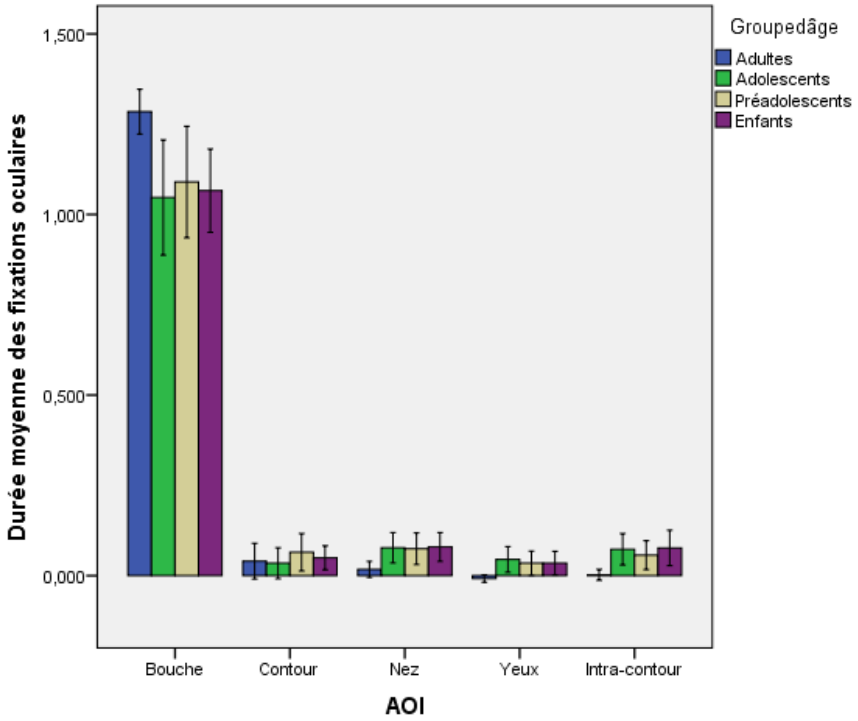

Figure 52. Variations de la durée moyenne (en secondes) des fixations oculaires dans la condition du format AVV-BA en fonction des différentes régions faciales de l'oratrice (AOI) et en fonction du degré de dégradation du groupe d'âge.

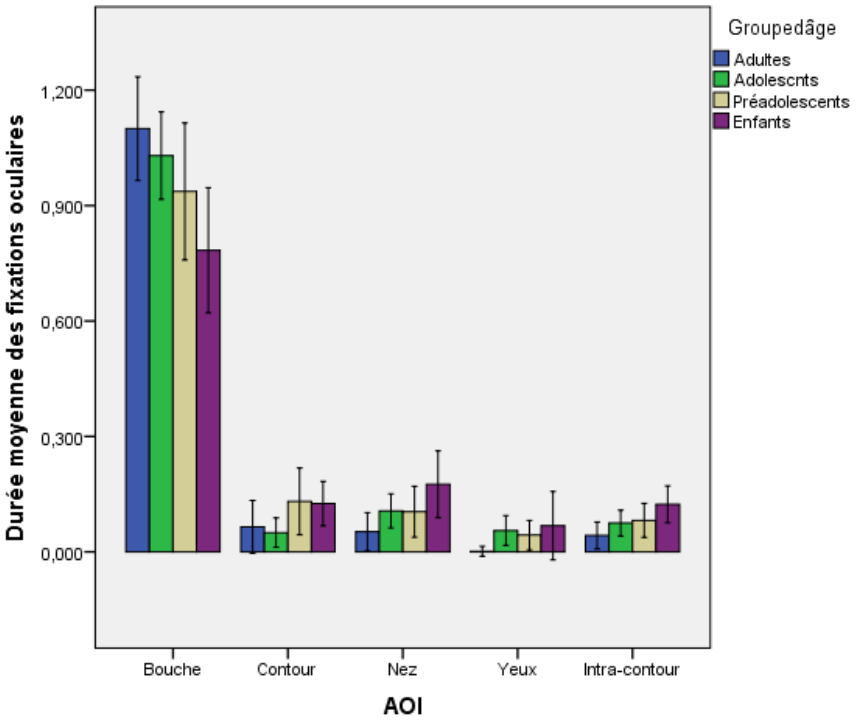

Figure 53. Variations de la durée moyenne (en secondes) des fixations oculaires dans la condition du format AVV-EBA en fonction des différentes régions faciales de l'oratrice (AOI) et en fonction du degré de dégradation du groupe d'âge.

7 Discussion

7.1 Expérience 1

Les résultats de l'expérience 1 nous ont permis de répondre à plusieurs questions de recherche. Plus précisément, il s'agit (i) de la question d'une éventuelle évolution du traitement audio-visuel de la parole de l'enfance tardive à l'âge adulte ; (ii) de la question concernant l'implication du traitement holistique, basé sur la dimension configurale de l'information faciale, dans la perception audio-visuelle de la parole ; (iii) de la question d'une éventuelle évolution du rôle du traitement holistique de l'information faciale dans la perception audio-visuelle de la parole ; (iv) de la question de l'effet des caractéristiques de bas niveau des stimuli visuels, centrés sur la région buccale de l'orateur, sur la perception audio-visuelle de la parole ; (v) de la question du comportement oculaire lors de la perception audio-visuelle de la parole, de son éventuelle évolution lors du développement et de ses éventuelles variations en fonction du format/des caractéristiques des stimuli visuels ; et (vi) d'une éventuelle variation dans les

dimensions évoquées en (i), (ii), (iii), (iv) et (v) en fonction du degré d'incertitude quant à l'input auditif.

7.1.1 Dimension développementale

Au sujet de la dimension développementale de la perception audio-visuelle de la parole, les résultats relatifs à la performance totale (le nombre des répétitions correctes) indiquent une supériorité globale du groupe « adultes » comparativement aux trois autres groupes d'âge. Ceci suggère que le traitement de la parole, notamment de ses aspects acoustiques, connaît une maturation relativement longue, se terminant possiblement à l'adolescence tardive, entre l'âge de 16 et 18 ans. En revanche, l'absence de l'effet du facteur « Groupe » dans les résultats en lien avec le gain AV, suggère que les mécanismes de traitement bimodal de la parole sont déjà pleinement opérationnels à l'enfance tardive. Ce résultat infirme ainsi notre hypothèse de départ selon laquelle au moins les plus jeunes participants devraient présenter un gain AV globalement inférieur à celui du groupe « adultes » et possiblement également à ceux des 2 autres groupes. Nos résultats diffèrent ainsi de ceux rapportés par Ross et *al*. (2011). En effet, ayant utilisé le même paradigme, Ross et *al*. (2011) ont trouvé une différence importante dans le gain AV entre le groupe des participants les plus jeunes (5 à 6 ans) et les participants plus âgés. Le gain AV variait peu entre l'âge de 7 et 11 ans. En revanche, une hausse importante dans le gain AV a été notée par ces auteurs après l'âge de 11 ans. Ces résultats, qui sont en lien avec la durée du processus de maturation des régions périsylviennes (Sowell, Thompson, Leonard, Welcome, Kan, & Toga, 2004), diffèrent ainsi de façon considérable des résultats de notre étude. Plusieurs explications semblent possibles pour une telle différence des résultats. Premièrement, les stimuli utilisés dans l'expérience de Ross et *al*. (2007) consistaient en mots monosyllabiques ayant une haute fréquence en langue anglaise écrite. En revanche, les stimuli de notre expérience consistaient en syllabes de type CV avec une voyelle qui était tenue pour constante, la voyelle /a/, qui est une voyelle très saillante, aussi bien sur le plan auditif que visuel (Benoît, Mohamadi, & Kandel, 1994). Le matériel de démonstration comportait des stimuli monosyllabiques se terminant également par la voyelle /a/. De ce fait, il nous semble possible que les stimuli de notre expérience aient induit un traitement essentiellement centré sur les caractéristiques acoustiques de l'input verbal, notamment phonémiques[44]. Par contraste, les

[44] Notons également qu'aucune réponse donnée par les participants ne comportait une autre voyelle que la voyelle /a/.

stimuli de l'expérience de Ross et *al*. (2007) auraient pu induire, dans une certaine mesure, un traitement lexical dans lequel les participants plus âgés étaient plus performants. Deuxièmement, le protocole de Ross et *al*. (2007) comportait un nombre assez important de stimuli (300 stimuli d'une durée de 1,3s). Il semble ainsi vraisemblable que les demandes attentionnelles imposées aux participants de Ross et *al*. (2011) aient été supérieures à celles de notre protocole qui comportait 132 items (dont 128 étaient des items des essais critiques). Dans la mesure où certains éléments empiriques suggèrent que la perception audio-visuelle de la parole implique une gestion attentionnelle de type top-down (Fernández, Visser, Ventura-Campos, Ávila, Soto-Faraco, 2015), les divergences entre les différents groupes d'âge observées par Ross et *al*. (2011) auraient pu être causées par les différences dans les capacités attentionnelles des participants. Cette explication semble confirmée davantage par le fait que le groupe « enfants » ait produit le nombre le plus important des protocoles individuels non retenus pour l'analyse à la base des critères en lien avec la qualité des données verbales et oculaires. En effet, il est probable que le nombre des protocoles éliminés de l'analyse reflète, au moins partiellement, des capacités attentionnelles moindres de certains participants. Une telle explication est également compatible avec l'hypothèse avancée par Barutchu et *al*. (2010) selon laquelle les performances des enfants en perception audio-visuelle de la parole dépendent de leurs capacités attentionnelles. Finalement, les réponses verbales des participants n'ont été codées que par une seule personne, l'expérimentateur, dans l'étude de Ross et *al*. (2011). Ceci aurait pu induire un biais lié à l'expérimentateur dans le codage des données. Dans le cas de notre étude, en revanche, le codage des réponses a été effectué par des juges aveugles quant aux objectifs de l'étude et quant à la bonne réponse pour chaque essai. Par ailleurs, n'ont été retenus pour l'analyse que les protocoles présentant un accord inter-juges excellent. De ce fait, il est possible que nos données aient été recueillies sur des enfants présentant un niveau relativement élevé de la production de la parole, fonction qui semble être positivement corrélée avec celle de la perception de la parole (Cabbage & Carell, 2014). Globalement, le fait que nos résultats ne montrent pas de différences inter-groupes au niveau du gain AV pourrait signifier soit (i) que les mécanismes de traitement audio-visuel de la parole, essentiellement pour ce qui est de l'aspect purement acoustique, atteignent leur maturité avant l'enfance tardive, soit (ii) que le traitement audio-visuel de la parole est lié au développement dans d'autres domaines cognitifs dans lesquels nos participants les plus jeunes présentaient un profil supérieur à la moyenne.

7.1.2 Degré de dégradation de l'information auditive

Outre l'absence de l'effet du facteur Groupe, les résultats relatifs au gain AV montrent que l'input bimodal facilite la perception de la parole davantage dans les conditions où l'input auditif est moyennement à fortement dégradé, comparativement à la condition de dégradation faible. Ce résultat est ainsi compatible avec les résultats des études précédentes (e.g. Ma et *al*., 2009 ; Ross et *al*., 2007 ; Ross et *al*., 2011). Toutefois, l'interaction SNR x Groupe que nous avons prédite à la base des résultats de Ross et *al*. (2011) n'a pas été significative. Aussi, notre hypothèse selon laquelle le bénéfice tiré de l'input bimodal dans la perception de la parole dégradée par le bruit devrait changer/évoluer avec l'âge en fonction du degré de dégradation de l'information auditive n'a pas été confirmée. Les divergences dans les résultats entre l'étude de Ross et *al*. (2011) et notre étude pourraient s'expliquer par le fait que, dans notre étude, la performance totale des participants dans la condition de présentation auditive des stimuli moyennement à fortement dégradés (ce qui correspondait au SNR-12[45]) a été supérieure à celle classiquement observée pour cette condition. En effet, le pourcentage des items correctement identifiés par les participants adultes dans notre étude dépassait d'environ 10% la performance totale moyenne des adultes, généralement estimée à 20% pour la condition en question. Cette différence entre les résultats des autres études et la nôtre pourrait s'expliquer par les caractéristiques de nos stimuli. En effet, le contexte phonémique (notamment la syllabe /a/) dans lequel ont été présentées les différentes consonnes des syllabes de type CV aurait pu faciliter la perception des consonnes (Benoît et *al*., 1994). Une autre explication possible est relative à nos critères de sélection des protocoles considérés comme valides pour l'analyse statistiques. Tout comme chez les enfants, ces critères aurait pu induire un biais, pourtant plus léger, également dans la sélection des participants adultes. En effet, une élimination des participants présentant des capacités attentionnelles relativement faibles au moment de l'expérience est possible. Toutefois, il nous semble peu probable qu'une telle sélection dans les caractéristiques attentionnelles de l'échantillon, qui, d'un point de vue statistique devrait être assez limitée, s'accompagne d'une différence relativement importante en termes de performance des participants. Il convient ainsi de noter que notre expérience comporte quelques limitations pour l'étude d'éventuelles variations dans l'évolution de la perception audio-visuelle de la parole bruitée induites par le changement dans le degré d'incertitude de l'input auditif.

[45] Rappelons que c'est au SNR-12 que le gain AV le plus élevé est classiquement observé chez les participants adultes (voir Ross et *al*., 2007).

7.1.3 Nature holistique (visage) *vs* localisée (bouche) de l'information visuelle

Au sujet de l'implication éventuelle du traitement holistique de l'information faciale dans la perception bimodale de la parole, le pattern de nos résultats relatifs au gain AV est caractérisé par l'absence de l'effet Format d'une part et par l'effet significatif de l'interaction Groupe x Format d'autre part. Les différences significatives entre les différents formats ont été observées uniquement dans le groupe « adultes ». Elles montrent une meilleure efficacité, des deux formats visuels réduits à la région buccale de l'oratrice comparativement au format présentant le visage entier. Pour le groupe des participants adultes, ce résultat semble ainsi confirmer notre troisième version des hypothèses concernant l'implication du traitement holistique de l'information faciale dans la perception audio-visuelle de la parole et sa pertinence dans ce cadre. Plus précisément, chez les adultes, la présence de l'information faciale semble gêner le traitement audio-visuel de la parole dans ses aspects purement acoustiques qui est, en revanche, facilitée par un input visuel localisé à la bouche seule.

Ce pattern des résultats s'inscrit ainsi dans la lignée des études suggérant que l'extraction des indices acoustiques de l'information visuelle, relative aux mouvements articulatoires, se fait davantage de façon analytique (e.g., Hietanen et *al*., 2001 ; Jordan et *al*., 2014 ; Thomas & Jordan, 2007). Aussi, notre résultat pourrait s'expliquer par le fait que la présence d'un stimulus visage ait automatiquement déclenché le traitement holistique de l'information faciale. Comme nous l'avons souligné dans la dernière partie de la section 3.4 du chapitre 3, le rôle du traitement holistique, même dans un cadre de perception audio-visuelle de la parole, pourrait avoir davantage le rôle de la perception de l'identité et des expressions faciales. En effet, la communication en face à face dans les conditions écologiques dépasse largement les aspects acoustiques de la parole. Il s'agit d'identifier le message et les intentions de notre interlocuteur. Du point de vue de l'évolution, l'automatisation des mécanismes de la perception des visages semblerait ainsi avantageuse pour l'individu. Toutefois, dans le cadre où il s'agit d'identifier les unités phonémiques, à partir d'indices faciaux et d'un input auditif ambigu/dégradé, le traitement holistique du visage semble interférer avec le traitement acoustique pur de la parole. Sur le plan développemental, nos résultats suggèrent que cette possible automatisation du traitement holistique du visage dans la perception de la parole n'est mise en place que vers la fin de l'adolescence, voire même à l'âge adulte. Cet aspect des résultats est concordant avec certains éléments empiriques montrant que la perception de

l'identité et des expressions faciales n'est possiblement pas encore optimale chez les adolescents (Bruce et *al*., 2000 ; Carey et *al*., 1980).

Sur la dimension en question, les résultats de notre étude peuvent également être interprétés en lien avec ceux des études précédentes, ayant mis en évidence certains marqueurs du traitement holistique de l'information faciale dans le cadre de la perception audio-visuelle de la parole. Plus précisément, la tatcherisation du visage de l'orateur est le paradigme qui a fourni les éléments les plus probants quant au fait que la déformation de l'information configurale d'un visage affecte la perception audio-visuelle de la parole. Conformément aux résultats de notre étude, cet effet pourrait être expliqué par le fait que le traitement holistique du visage est perturbé en présence d'un visage tatcherisé. Traiter un visage tatcherisé peut ainsi être relativement coûteux pour l'individu, ce qui laisse moins de ressources cognitives au traitement de la parole qui en est, à terme, perturbé également. Ainsi, même si le traitement holistique semble être impliqué dans le traitement de l'information faciale dans la perception audio-visuelle de la parole, sa fonction ne semble pas être la facilitation de l'extraction des indices acoustiques à partir de l'information visuelle. Dans leur ensemble, les résultats de ces études, ainsi que ceux de la nôtre, pourraient même laisser présager une certaine hiérarchie dans les traitements des différents types d'informations portés par les visages avec les traitements de l'identité et des expressions faciales qui seraient prioritaires par rapport au traitement des indices purement acoustiques. Une telle hiérarchie pourrait être instaurée assez tard dans le développement, notamment dans l'adolescence tardive.

7.1.4 Caractéristiques de l'information visuelle localisée à la bouche

Au sujet des caractéristiques de bas niveaux des stimuli visuels, les résultats relatifs au gain AV montrent que l'efficacité des différents formats AV est globalement la même. Toutefois, en prenant en compte les deux degrés de dégradation de l'information auditive, nos résultats mettent en évidence des variations dans l'efficacité des formats en question. En effet, dans la condition de faible dégradation de l'information auditive (SNR-6), le format AVB-E a été associé à un gain AV plus important que le format AVV. Une tendance, pourtant non significative, est apparue également entre les deux formats comportant la bouche seule de l'oratrice, avec le format AVB-E qui a été plus efficace que le format AVB-M. En revanche, dans la condition d'une dégradation moyenne à forte de l'information auditive, le format AVB-M s'est révélé plus efficace que le format AVB-E. Ce résultat infirme ainsi, au moins partiellement, notre hypothèse selon laquelle le format AVB-M devrait être moins efficace que

le format AVB-E, car les régions noires utilisées pour masquer le visage de l'oratrice, seraient susceptibles d'attirer le regard des participants, le déviant ainsi de la région buccale ce qui perturberait, à terme, le traitement audio-visuel de la parole (Thomas & Jordan, 2004). A ce propos, les résultats relatifs au comportement oculaire des participants, notamment à la durée des fixations oculaires dans la région buccale de l'oratrice, les résultats de notre étude confirment partiellement l'hypothèse de Thomas et Jordan (2004). Globalement, les participants avaient en effet tendance à fixer plus longuement la région buccale avec le format AVB-E qu'avec le format AVB-M. En outre, cette même différence a également été observée pour les formats AVB-E et AVV, avec le premier qui retenait le regard des participants au niveau de la région buccale de l'oratrice plus longuement que le format AVV.

De tels résultats pourraient s'expliquer partiellement par la saillance des objets/des formes visuel(le)s présenté(e)s dans chacun des formats. En effet, le format AVB-E ne comportant qu'un objet, la bouche de l'oratrice, ce dernier avait le rôle du seul objet saillant dans l'image et notre attention visuelle est connue pour être orientée vers et attirée par les éléments visuellement saillants (e.g., Itti & Koch, 2000 ; Reingold & Loschky, 2002). Toutefois, la saillance visuelle de la bouche de l'oratrice entourée de régions créant de forts contrastes lumineux entre la figure et le fond était encore plus marquée que la saillance de la bouche dans le format AVB-E. Il est de ce fait étonnant que le regard des participants n'ait pas été attiré davantage par la région buccale de l'oratrice dans le format AVB-M. En examinant l'effet de l'interaction Groupe x Format, nous constatons que la différence globale dans la durée des fixations oculaires au niveau de la région buccale de l'oratrice entre les formats AVB-E et AVB-M est essentiellement due au pattern de l'inégalité entre ces deux formats observé dans le groupe « adultes ». En effet, les résultats de l'interaction entre les facteurs Groupe et Format, relatifs au comportement oculaire des participants, ont mis en évidence que la différence entre les formats AVB-E et AVB-M n'est apparue que dans le groupe « adultes ». En revanche, les groupes « enfants » et « pré-adolescents », se caractérisent par une différence entre les formats AVB-E et AVV, la bouche de l'oratrice ayant été fixée plus longuement dans la condition AVB-E. Ce même pattern a également été observé chez les participants adultes. Néanmoins, il s'agissait d'une tendance qui n'était pas significative.

Ces résultats pourraient s'expliquer par le fait que le format AVB-M ait induit les participants adultes dans un traitement global des stimuli visuels de ce type, pouvant prendre la forme de suppositions par rapport à ce qui se cache derrière le masque, comme le suggèrent Thomas et Jordan (2004). Une telle explication semble probable, car la perception des patterns

visuels, même incomplets, est marquée par un traitement global chez les adultes. Selon certains auteurs, ce type de traitement pourrait connaître une maturation relativement longue, allant jusqu'à l'adolescence (Hadad, Mauer, & Lewis, 2010 ; Kimchi, Hadad, Behrmann, & Palmer, 2005 ; Sherf, Behrmann, Kimchi, & Luna, 2009). Contrairement aux caractéristiques du format AVB-E, l'application d'un masque noir dans le format AVB-M aurait pu ainsi amplifier la dimension du contexte facial caché et induire une approche globale dans l'inspection de telles images, essentiellement chez les participants adultes chez qui le traitement global des patterns visuels est pleinement fonctionnel. Toutefois, dans le groupe « adultes », le gain AV a été comparable pour les formats AVB-M et AVB-E et supérieur à celui obtenu avec le format AVV. Aussi, il semblerait que le traitement global du format AVB-M ait été peu approfondi, se caractérisant possiblement par une simple stratégie d'inspection de la scène visuelle qui pourrait être (partiellement) automatisée chez les adultes. En revanche, comme nous l'avons exposé plus haut, le format AVV a vraisemblablement déclenché un traitement global de l'information faciale, alors que les indices visuels les plus cruciaux pour la perception des caractéristiques acoustiques des sons produits venaient d'un élément du visage, notamment la bouche. Or, chez les adultes, le traitement visuel d'éléments locaux intégrés dans une scène visuelle plus large est connu pour être sous le contrôle exécutif. En effet, le traitement global du stimulus doit être inhibé (Poirel, Krakowski, Sayah, Pineau, Houdé, & Borst, 2014) ce qui est un processus coûteux sur le plan cognitif. Cette gestion exécutive du traitement global, non pertinent au vu de la tâche de l'expérience 1, a ainsi très probablement interféré avec le traitement bimodal de la parole dans le groupe « adultes ».

7.1.5 Variations de l'influence des différents formats visuels sur la perception bimodale de la parole au cours du développement

Sur le plan développemental, l'interaction Format x Groupe révèle une évolution dans le traitement visuel des différents formats AV. L'absence de différence entre les formats AVB-M et AVB-E, quant à la durée des fixations oculaires dans la région buccale de l'oratrice, dans les groupes « enfants », « pré-adolescents » et « adolescents », pourrait s'expliquer par le processus de maturation du traitement global. Il est en effet possible que, ce type de traitement n'étant pas pleinement fonctionnel avant l'adolescence tardive, son implication dans le traitement du format AVB-M ait été moindre dans les groupes en question. Par ailleurs, le fait qu'aucune différence inter-formats n'ait été observée pour le traitement visuel de la région buccale de l'oratrice chez les adolescents pourrait suggérer que le traitement global des stimuli

visuels, ou au moins la gestion du comportement oculaire qui accompagne un tel traitement, connaît une étape cruciale dans l'organisation fonctionnelle en adolescence. Les éléments empiriques produits dans le cadre des études précédemment évoquées (Hadad et *al.*, 2010 ; Kimchi et *al*., 2005 ; Sherf et *al*., 2009) semblent en effet appuyer cette hypothèse.

Globalement, les résultats relatifs à la durée des fixations oculaires dans la région buccale de l'oratrice pour les formats AV de l'expérience 1 suggèrent que le comportement oculaire qui sous-tend le traitement des éléments locaux d'une scène visuelle, en l'occurrence de la bouche d'un visage, connaît une certaine évolution au cours du développement. Avant l'adolescence, le facteur de saillance visuelle, telle que décrite pour le format AVB-E, semble être la façon la plus efficace pour orienter le regard vers la cible. Ceci est probablement dû au fait que les mécanismes visuo-attentionnels impliqués dans le traitement des caractéristiques de bas niveau des stimuli visuel de ce type sont déjà pleinement fonctionnels en enfance (pour une revue sur la question, voir Braddick & Atkinson, 2011). En revanche, chez les adultes, les deux variantes des stimuli localisés à la région buccale de l'oratrice, celle du format AVB-E et AVB-M, affectent le comportement visuo-attentionnel, chacune de façon différente. Le traitement du format AVB-M s'est probablement accompagné d'une inspection visuelle allant davantage vers le masque qui occultait le visage, ce qui est vraisemblablement une stratégie propre au traitement global des patterns visuels. Ce type de traitement semble connaître une maturation longue, se terminant possiblement à l'adolescence tardive.

7.1.6 Comportement oculomoteur et performances en perception bimodale de la parole

Dans le cadre de l'interprétation des résultats de l'expérience 1, il est également important de prendre en compte aussi bien les résultats relatifs au gain AV que ceux reflétant le comportement oculaire des participants lors du traitement visuel des stimuli AV. Ainsi, nous pouvons constater que le format AVB-E qui a, globalement, attiré le regard des participants le plus longtemps au niveau de la région buccale, n'a pas été systématiquement associé au gain AV le plus élevé. En effet, dans la condition de dégradation faible de l'information auditive (SNR-6), le gain AV obtenu avec le format AVB-E a été plus élevé que celui obtenu avec le format AVV (et présentait une tendance vers ce pattern pour le format AVB-M). En revanche, dans la condition de dégradation moyenne/forte de l'information auditive (SNR-12), le gain AV a été plus élevé pour le format AVB-M comparativement au format AVB-E. Il semblerait ainsi que, dans les conditions où un degré faible d'incertitude est associé à l'input auditif, la perception bimodale de la parole est facilitée par une orientation du regard vers la région

buccale. Ceci pourrait possiblement faciliter le traitement des indices acoustiques apportés par l'input visuel. En revanche, dans les conditions où l'input auditif est marqué d'une ambiguïté assez importante, les conditions de présentation de la région buccale propres au format AVB-M semblent faciliter le traitement des indices visuels relatifs aux sons produits. Notre interprétation de ces résultats est que le fort contraste lumineux entre la région buccale et le masque, ainsi que le fait de traiter l'information visuelle en vision périphérique (pattern propre au groupe « adultes ») sont deux facteurs qui ont pu faciliter la détection et la perception du mouvement. En effet, la vision périphérique est caractérisée par une bonne capacité à détecter le mouvement (e.g., Berg, Berglund, Strang, & Baum, 2007). En outre, le traitement visuel local du mouvement semble être amélioré par des stimuli présentant un haut contraste lumineux par rapport au reste de la scène visuelle (e.g., Huerta, Amato, Roca, & Gonzàlez, 2013). Aussi, il semblerait que la perception bimodale de la parole profite davantage des indices visuels relatifs à la dimension temporelle (mouvements préparatoires) de la production verbale orale dans les conditions où l'intelligibilité de l'information auditive est fortement perturbée. Une telle explication surligne ainsi l'importance des mécanismes attentionnels et leur modulation par les caractéristiques de bas niveau de l'input visuel pour la perception bimodale de la parole dans les conditions de haute incertitude quant à l'input auditif.

Globalement, il semble ainsi que le format comportant l'information faciale localisée à la seule région buccale de l'oratrice peut être avantageux pour la perception des aspects purement acoustiques de la parole. Cet avantage semble au moins partiellement lié aux différences dans le traitement visuo-attentionnel de l'information visuelle présentée. Notons toutefois que ce n'est que dans le groupe « adultes » que les deux formats AV localisés à la région buccale ont été associés à un gain AV supérieur à celui observé avec le format AVV. L'absence de l'interaction Groupe x Format x SNR laisse supposer que les différences inter-formats observées dans le groupe « adultes » ne dépendaient pas simplement d'une influence des caractéristiques de bas niveau des stimuli visuels sur le traitement visuo-attentionnel (revoir la section 7.1.4 du présent chapitre pour l'interprétation des différences inter-formats dans le groupe « adultes »).

7.1.7 Comportement oculomoteur

Les données relatives à la durée des fixations dans les différentes régions faciales de l'oratrice dans la condition AVV nous ont permis d'étudier également la façon de collecter l'information visuelle lors de la perception audio-visuelle de la parole et son éventuelle

variation en fonction de l'âge et du degré de dégradation de l'information auditive. Pour ce type de données, les résultats ont montré que la région buccale de l'oratrice a été la principale cible visuelle. Elle a été fixée plus longuement que chaque autre région faciale, notamment le nez, les yeux, l'intra-contour et le contour. Une autre cible visuelle importante dans la perception audio-visuelle de la parole semble être le nez. En effet, le nez de l'oratrice a été fixé plus longuement que les yeux, le contour et l'extra-contour. Contrairement à ce qui a été observé pour la perception audio-visuelle de la parole dans les conditions où l'input auditif est normalement audible (e.g., Everdell et *al*., 2007 ; Lansing & McConkie, 2003 ; Paré et *al*., 2003), le regard des participants n'a pas été particulièrement attiré par les yeux de l'oratrice dans notre expérience. De plus, l'effet de l'interaction entre les facteurs AOI et SNR a révélé que le pattern de la durée des fixations oculaires variait en fonction du degré de l'intelligibilité de l'input auditif. Ces variations concernaient les régions de la bouche, des yeux et du contour facial. En effet, dans la condition de dégradation moyenne à forte de l'information auditive, la bouche a été fixée plus longuement que dans la condition de dégradation faible. En revanche, la durée des fixations oculaires des yeux et du contour facial de l'oratrice a été moins longue dans la condition de dégradation forte que dans la condition de dégradation faible de l'information auditive.

Quant à l'augmentation de la durée des fixations oculaires dans la région buccale dans la condition de forte dégradation de l'information auditive, les résultats confirment notre hypothèse et concordent avec ce qui a été observé par Vatikotis-Bateson et *al*. (1998) dans la mesure où l'augmentation dans la dégradation de l'information auditive s'accompagnait d'un attrait plus important du regard des participants dans la région buccale de l'oratrice. En revanche, le pattern des variations du comportement oculaire en fonction de l'intelligibilité de l'information auditive lors de la perception de la parole rapporté par Buchan et *al*. (2008) n'a pas été observé dans notre expérience. Plus précisément, Buchan et *al*. (2008) ont rapporté que la dégradation de l'input auditif s'accompagnait des fixations plus longues au niveau du nez, alors que la durée des fixations au niveau des yeux diminuait. Dans notre expérience toutefois, la durée des fixations oculaires du nez de l'oratrice ne variait pas entre les deux conditions de dégradation de l'information auditive. Aussi, cette dimension des résultats infirme la partie de notre hypothèse qui y était relative. Plusieurs raisons potentiellement explicatives des différences des résultats entre notre étude et celle de Buchan et *al*. (2008) peuvent être envisagées. Premièrement, notre expérience ne comportait pas de condition avec des stimuli auditifs normalement audibles. Il est probable que dans de telles conditions la durée des

fixations au niveau des yeux de l'oratrice aurait été plus longue et la durée des fixations du nez plus courte que dans les conditions avec un input auditif dégradé. Deuxièmement, la non variation de la durée des fixations du nez de l'oratrice entre les deux conditions de dégradation de l'information auditive pourrait être relative au type de tâche ainsi qu'aux caractéristiques des stimuli visuels utilisés dans notre étude. En effet, le traitement bimodal de la parole dans notre expérience impliquait uniquement l'aspect acoustique des syllabes, particulièrement de la consonne de chaque syllabe. Une telle tâche a pu induire une stratégie de traitement visuel particulière. Plus précisément, il semble que les participants se soient focalisés sur les éléments faciaux leur permettant une extraction optimale des indices visuels relatifs à la parole de l'information faciale déjà dans la condition de dégradation faible de l'information auditive. Une augmentation de dégradation de l'input auditif a pu induire une stratégie de traitement visuel impliquant un retrait des fixations oculaires des régions les moins pertinentes, les yeux et le contour facial, pour augmenter la durée des fixations oculaires au niveau de la région faciale la plus informative quant à l'input auditif, notamment la bouche. Ce type de stratégie de l'adaptation du traitement visuel de l'information faciale en fonction du degré d'intelligibilité de l'input auditif a, par ailleurs, pu être facilité par le fait qu'aucun mouvement et aucune expression faciale particulière n'apparaissaient dans la partie supérieure du visage de l'oratrice.[46]

Finalement, notons que l'interaction attendue entre les facteurs AOI et Groupe n'a pas été significative. Contrairement à notre hypothèse, le pattern du comportement oculaire ne variait pas en fonction du groupe d'âge. Aussi, il semblerait que les mécanismes visuo-attentionnels impliqués dans le traitement de l'information faciale lors de la perception audio-visuelle de la parole sont pleinement fonctionnels à l'enfance tardive. Toutefois, la validité d'un tel résultat est possiblement restreinte aux conditions de notre expérience où (i) les participants avaient pour tâche de traiter la parole dans son aspect purement acoustique, (ii) l'information visuelle relative à la parole a été apportée par un visage relativement statique et inexpressif, et (iii) l'input auditif a été dégradé par le bruit. En outre, les profils cognitifs de nos participants

[46] Dans ce contexte, il convient de noter qu'un pattern du comportement oculaire marqué d'une forte tendance de traitement d'un élément majeur de la scène visuelle ne signifie pas que l'information des autres parties du visage n'a pas pu être extraite et traitée, au moins dans une certaine mesure (pour plus de détails, voir Calvo, Fernández -Martin, & Nummenmaa, 2014). Au niveau de la section 7.1.2, nous avons notamment émis l'hypothèse que les adultes auraient traitée l'information relative à l'organisation globale des composantes faciales. Cette hypothèse n'est pas contradictoire avec l'interprétation des résultats relatifs au comportement oculaire dans l'expérience 1.

du groupe « enfants » auraient pu être (légèrement) supérieurs à la moyenne (revoir la section 7.1.1 du présent chapitre pour plus de détails sur ce sujet). Finalement, contrairement au postulat longtemps tenu pour valide dans la recherche sur le comportement oculaire lors du traitement d'une scène visuelle, l'extraction de l'information d'un stimulus visuel n'est pas limité à l'endroit d'une fixation oculaire donnée. En effet, l'information visuelle peut être extraite également en vision périphérique (e.g., Rayner, 1975 ; Reingold, Charness, Pomplun, & Stampe, 2011). Notre étude met ainsi l'accent sur l'importance de combiner les données relatives au comportement oculaire avec d'autres types de données (comportementales ou autres) afin de cerner le type d'information qui a pu être extrait des stimuli visuels et la nature de traitement que cette information a reçu. (Dans ce contexte, revoir la section 7.1.3 pour l'hypothèse du traitement holistique possiblement impliqué dans le traitement de l'information faciale lors de la perception audio-visuelle de la parole chez les adultes, mais pas dans les autres groupes d'âge.) Pour les raisons évoquée, il est difficile d'évaluer la validité externe de ce résultat ainsi que sa portée/signification dans le domaine de la perception audio-visuelle de la parole. D'autres études sont nécessaires pour éclairer la problématique relative aux stratégies spatiales de l'extraction de l'information visuelle, faciale, et leur éventuelle évolution au cours du développement.

7.2 Expérience 2

Les données recueillies dans le cadre de l'expérience 2 nous ont permis essentiellement d'étudier (i) la contribution des mouvements dans les régions extra-buccales à la perception bimodale de la parole ; (ii) le comportement oculaire lors de la perception audio-visuelle de la parole et ses éventuelles variations en fonction du type d'information apportée par le visage de l'oratrice ; (iii) une éventuelle évolution dans les aspects (i) et (ii) au cours du développement allant de l'enfance tardive à l'âge adulte ; (iv) un éventuel impact du degré de l'intelligibilité de l'input auditif sur les aspects (i), (ii) et (iii).

7.2.1 Contribution des mouvements dans les régions extra-buccales à la perception de la parole

Les résultats relatifs au gain AV ont mis en évidence l'effet significatif du facteur Format. En effet, pour les formats AVV et AVV-BA, le gain AV s'est révélé être comparable et supérieur à celui obtenu avec le format AVV-EBA. Un tel résultat suggère que les indices

visuels restreints aux mouvements articulatoires dans les seules régions extra-buccales sont bien moins informatifs/utiles pour la perception bimodale de la parole que ceux provenant de la bouche seule. En effet, le fait que les formats AVV et AVV-BA aient produit un gain AV comparable chez les participants indique que les indices visuels de la parole localisés à la seule région buccale de l'oratrice sont également informatifs pour la perception bimodale de la parole que les indices provenant, de façon conjointe, aussi bien de la région buccale et que des mouvements dans les régions buccales.

L'inspection des résultats relatifs à la performance totale des participants de l'expérience 2 nous informe que l'interaction entre les facteurs Format et SNR a été significative. Elle met en évidence que, au niveau du SNR-12, la performance totale des participants a été supérieure pour toutes les conditions AV comparativement à la condition AU seule. En outre, à ce SNR, la performance totale pour les formats AVV et AVV-BA a été comparable et supérieure à celle observée avec le format AVV-EBA. Ce pattern des différences inter-conditions était différent au niveau du SNR-6. En effet, à ce niveau de dégradation de l'information auditive, on constate que la performance totale des participants ne différait pas entre les conditions AVV-EBA et AU. Par ailleurs, le gain AV obtenu avec le format AVV-EBA au SNR-6 était nul. De tels résultats suggèrent ainsi que l'information visuelle relative aux mouvements dans les régions extra-buccales de l'oratrice n'est utile à la perception bimodale de la parole que dans le cas de haute incertitude quant à l'input auditif. Dans la mesure où le format AVV-EBA n'apportait aucune information visuelle quant aux caractéristiques acoustiques des sons produits (les mouvements de la bouche n'étant pas présentés), les indices visuels relatifs aux mouvements dans les régions extra-buccales ne pouvaient apporter que des informations relatives aux aspects temporels de la production verbale orale. Il semble ainsi que, seul, ce type d'indices contribue à une meilleure perception des unités de la parole dans les conditions de haute dégradation de l'information auditive, vraisemblablement en baissant le seuil de détection des sons à traiter (voir Grant & Seitz, 2000). Cet aspect des résultats semble ainsi appuyer l'interprétation des résultats de l'expérience 1 relatifs à la hausse du gain AV au niveau du SNR-12 pour le format AVB-M. En effet, l'interprétation proposée de ce résultat, selon laquelle ce type de format a pu faciliter la détection du mouvement dans la région buccale de l'oratrice et, à terme, la détection des indices visuels relatifs aux aspects temporels de la production de la parole semble cohérente avec les résultats de l'expérience 2.

Les indices visuels restreints aux seuls mouvements des régions extra-buccales semblent ainsi pouvoir affecter la perception bimodale de la parole. Toutefois, leur apport particulier à

ce type de perception apparaît comme limité aussi bien au niveau des caractéristiques de l'input auditif qu'au niveau des caractéristiques de l'input visuel. En effet, sur ce dernier point, les résultats de notre expérience montrent que le gain AV, ainsi que la performance totale, ont été comparables pour les formats AVV et AVV-BA aux deux niveaux de dégradation de l'information auditive. Aussi, il semble que l'ajout des indices visuels provenant des mouvements dans les régions extra-buccales aux indices visuels relatifs aux mouvements articulatoires dans la région buccale de l'oratrice n'améliore pas le traitement bimodal de la parole.

Cette dernière dimension des résultats semble ainsi cohérente avec ce qui a été observé au niveau de l'expérience 1 en lien avec la tendance significative vers un gain AV plus élevé pour les formats AV présentant l'information faciale de façon localisée à la région buccale seule. Dans ce contexte, l'absence de différence dans le gain AV entre les formats AVV et AVV-BA (le gain AV était légèrement, mais non significativement, inférieur dans la condition AVV-BA) et la non variation de ce pattern en fonction du groupe d'âge et en fonction du SNR suggère que l'information sur les mouvements articulatoires, apportée par le visage de l'oratrice, a reçu un même traitement/un traitement comparable, que les mouvements dans les régions extra-buccales soient présents ou pas. Aussi, il semblerait que ce soit bien le contexte facial qui influence le traitement de l'information visuelle dans le cadre de la perception bimodale de la parole, au moins en ce qui concerne le traitement des aspects purement acoustiques de la production verbale orale. Par ailleurs, l'influence de l'information faciale, holistique/configurale, ne semble pas être influencé par des variations, pourtant restreintes, dans la nature écologique du stimulus. (En effet, le format AVV-BA comportait un aspect purement artificiel relatif à la nature statique du contexte facial lors de la production d la parole.)

Notons finalement que nos résultats relatifs à l'apport de l'information visuelle restreinte aux mouvements dans les régions extra-buccales de l'oratrice diffèrent de ceux obtenus par Thomas et Jordan (2004). En effet, ces auteurs ont observé que l'apport de ce type de l'information visuelle était faible mais significatif dans la condition de dégradation faible de l'information auditive. Cette différence dans les résultats de deux études est probablement relative aux stimuli utilisés. En effet, Thomas et Jordan (2004) ont opté pour des mots. Dans ce cas, les connaissances lexicales auraient pu amplifier l'effet de l'apprentissage, propre au design expérimental à mesures répétées, qui aurait pu, à terme, avoir une influence sur les réponses des participants. Ceci est d'autant plus probable que le nombre des stimuli a été peu élevé (6 mots mono syllabiques (*bay*, *beer*, *gay*, *gear*, *map*, et *nap*)) et que les résultats dans de

nombreuses conditions audio-visuelles étaient proches du pattern de l'effet plafond[47]. Par contraste, dans notre étude, la tâche des participants était vraisemblablement limitée au traitement phonologique de la parole. Il s'agissait, plus précisément, d'identifier la première consonne de chaque syllabe.

7.2.2 Dimension développementale

Sur le plan développemental, les résultats de l'expérience 2 sont cohérents avec ceux de l'expérience 1. En effet, d'une part l'effet du facteur Groupe a été observé pour les données de la performance totale. L'effet en question présentait le même pattern que celui de l'expérience 1. Plus précisément, la performance totale des adultes a été supérieure à celle observée dans les autres groupes d'âge qui ne différaient pas entre eux à ce niveau. Ce résultat confirme ainsi la conclusion proposée au niveau de la discussion des résultats de l'expérience 1. Il semblerait notamment que les adultes ont de meilleures capacités de traitement de la parole que les adolescents, les pré-adolescents et les enfants. Ceci indique une maturation possiblement très longue des mécanismes cognitifs impliqués dans le traitement des aspects purement acoustiques. Pour d'autres explications possibles, relatives aux modalités méthodologiques de notre expérience, revoir la section 7.1.1 du présent chapitre. D'autre part, aucune différence inter-groupes n'a été observée au niveau du gain AV. L'absence du facteur Groupe pour le gain AV est également conforme à ce qui a été observé dans l'expérience 1. Un tel résultat appuie davantage la conclusion selon laquelle les mécanismes spécifiquement impliqués dans le traitement bimodal de la parole atteignent leur maturité avant l'enfance tardive. (Toutefois, revoir de nouveau la section 7.1.1 du présent chapitre pour les limites de cette conclusion.) Dans ce contexte il convient de noter que l'interaction Groupe x Format n'a pas été significative. Ceci suggère ainsi que les enfants traitent l'information faciale relative à la production de la parole aussi bien que les adultes, même quand cette dernière est limitée aux seuls mouvements dans les régions extra-buccales de l'oratrice.

[47] Pour cette raison, la validité des résultats quant aux questions étudiées par Thomas et Jordan (2004) est relativement limitée.

7.2.3 Comportement oculomoteur

Quant aux résultats relatifs à la durée des fixations oculaires dans les différentes régions faciales de l'oratrice, les résultats de l'expérience 2 mettent en évidence des différences dans cette dimension du comportement oculaire par rapport à l'expérience 1. En effet, l'effet du facteur AOI a révélé la présence du pattern standard quant aux différences dans la durée des fixations oculaires entre la bouche et chaque autre région faciale, avec la bouche qui a attiré le regard des participants plus longtemps. Toutefois, sur l'ensemble des formats, le nez a été fixé plus longtemps qu'une seule autre région faciale, notamment la région des yeux. Par ailleurs, l'intra-contour, comportant les régions extra-buccales de l'oratrice, présentait également une tendance, pourtant non significative, vers une durée plus longue des fixations oculaires dans cette région comparativement aux yeux de l'oratrice. Ce pattern global présentait des variations aussi bien en fonction du groupe d'âge qu'en fonction du format de présentation de l'information visuelle. Finalement, le pattern de l'interaction entre les facteurs AOI et Groupe variait également en fonction du degré de dégradation de l'information auditive.

Afin de comparer le pattern du comportement oculaire des participants quant à la durée des fixations oculaires dans les différentes régions faciales de l'oratrice obtenu avec le format AVV entre les expériences 1 et 2, nous devons nous intéresser de plus près à l'effet de l'interaction entre les facteurs AOI et Format de l'expérience 2. Ce type de résultats nous révèle que, pour l'ensemble des groupes d'âge, la durée des fixations oculaires présentait le pattern caractéristique de l'effet du facteur AOI, présenté ci-dessus. Comparativement aux résultats de l'expérience 1, nous constatons ainsi que, dans le format AVV, le nez n'a pas attiré significativement plus le regard que l'intra-contour et que le contour. En revanche, l'intra-contour a été fixé plus longtemps que la région des yeux avec le format en question, ce qui n'était également pas le cas dans l'expérience 1. Dans la mesure où les stimuli de la condition de présentation AVV ont été exactement les mêmes dans les deux expériences, il semblerait que le comportement oculaire des participants ait pu être influencé par les autres conditions de présentation AV. Dans l'expérience 1, les deux formats visuels présentant l'information faciale restreinte à la bouche seule extraite du contexte facial auraient ainsi pu induire une stratégie d'exploration visuelle du format AVV centrée davantage sur la bouche. Par contraste, les formats AVV-BA et AVV-EBA semblent avoir induit une stratégie d'exploration visuelle du format AVV plus dispersée, privilégiant davantage les régions de l'intra-contour et du contour au détriment de la région du nez.

Les résultats de l'expérience 2 révèlent également que le pattern des fixations oculaires dans les différentes régions faciales de l'oratrice dans la condition AVV variait en fonction du groupe d'âge, ce qui n'a pas été le cas dans l'expérience 1. Cette observation s'appliquait d'ailleurs à l'ensemble des formats AV de l'expérience 2. Dans leur ensemble, les résultats de l'interaction Groupe x Format x AOI révèlent plusieurs caractéristiques générales quant à la façon de traiter visuellement les stimuli des conditions AVV, AVV-BA et AVV-EBA. Premièrement, le pattern du comportement oculaire des participants d'un groupe d'âge donné était le même ou très similaire pour les formats AVV et AVV-BA. Ceci suggère que, indépendamment de la nature intégrale (mouvements de la bouche et mouvements dans les régions extra-buccales) ou partielle (mouvements de la bouche) de l'information visuelle relative à la parole, l'information faciale, quand elle est complète (comportant toutes les composantes faciales) provoque la même stratégie visuo-attentionnelle et reçoit vraisemblablement le même traitement quant à l'extraction de l'information visuelle. Deuxièmement, pour l'ensemble des groupes d'âge, trois patterns du comportement oculaire dans les conditions AVV et AVV-BA ont été observés. Le premier consistait en une durée des fixations plus élevée au niveau de la région buccale comparativement à chaque autre région faciale. Il caractérisait le groupe « adultes » et le groupe « enfants ». Les groupes « adolescents » et « pré-adolescents » présentaient certaines similitudes dans le pattern oculaire associé aux formats AVV et AVV-BA, mais également certaines différences. Les similitudes concernaient (i) la différence dans la durée des fixations oculaires entre la région buccale et chaque autre région faciale, et (ii) une durée plus longue des fixations au niveau du nez par rapport à la région des yeux. Quant aux différences, dans le groupe « adolescents », le nez a été fixé plus longuement que le contour (il s'agit cependant d'une tendance non significative), alors que le groupe « pré-adolescents », le nez a été fixé plus longtemps que l'intra-contour (cette tendance n'a pas été significative pour le format AVV). Dans leur ensemble, les différences inter-groupes quant au pattern du comportement oculaire lors de la perception audio-visuelle de la parole pourraient être consécutives à une éventuelle réorganisation fonctionnelle des mécanismes de traitement de l'information faciale se terminant possiblement à l'adolescence tardive. Une telle explication est cohérente avec ce qui a été proposé pour expliquer les différences dans l'efficacité des formats AVV, AVB-E et AVB-M et son variation inter-groupes observées dans l'expérience 1 (revoir la section 7.1.2 du présent chapitre).

Finalement, les résultats de l'interaction Groupe x Format x AOI nous révèlent également que le pattern du comportement oculaire associé au format AVV-EBA différait du

pattern associé à deux autres formats AV à l'intérieur de chaque groupe. Cette dimension des résultats indique ainsi que ce type de format a déclenché des stratégies de traitement visuo-spatial différentes que les formats AVV et AVV-BA. Cela pourrait s'expliquer par le fait que les indices visuels utiles pour la perception de la parole se trouvaient uniquement dans les régions extra-buccales dans ce type de format et que ces régions ont pu, par conséquent, attirer le regard des participants davantage que cela n'a été le cas avec les formats AVV et AVV-BA. Certains aspects des résultats de l'interaction Groupe x Format x AOI pourraient confirmer cette hypothèse, en mettant en évidence une différence au profit du contour et de l'intra-contour dans le pattern oculaire associé au format AVV-EBA qui n'étaient pas présentes pour les formats AVV et AVV-EBA. Par exemple, le format AVV-EBA a induit chez les adultes une durée des fixations plus longue aussi bien du contour que de l'intra-contour comparativement aux yeux. En revanche, le fait que la durée des fixations des yeux soit équivalente à celle de la région du nez pour le format en question dans le groupe des adolescents n'est pas conforme à l'explication proposée. En effet, une telle explication semble peu probable, car une stratégie optimale pour extraire et par la suite traiter les indices visuels des mouvements dans les régions extra-buccales serait de fixer davantage la bouche et éventuellement le nez. Les indices des mouvements dans les régions extra-buccales pourraient ainsi être extraits par vision périphérique avec la même efficacité pour la partie gauche et droite du visage.

Aussi, il semble possible que le comportement oculaire des participants observé avec le format AVV-EBA ait été davantage influencé par l'absence d'une composante clef du visage, la bouche. L'effet du facteur Format, montrant qu'avec le format AVV-EBA la durée des fixations oculaires au niveau de l'oratrice a été moindre que celle obtenue avec le format AVV serait susceptible de confirmer cette explication. En effet, globalement, le regard des participants semble avoir été moins attiré par le visage présentant un biais vraisemblablement assez important quant à la version écologique du stimulus, en l'occurrence par un visage dépourvu d'une de ses composantes essentielles et visuellement saillantes, la bouche. Finalement, l'attention visuelle des participants et son orientation vers le visage et notamment la région buccale de l'oratrice aurait également pu être amoindrie à cause du caractère non informatif ou faiblement informatif des indices visuels par rapport aux sons produits. Notre étude ne nous permet pas de trancher entre les deux explications, d'autres études sont nécessaires pour éclairer ce sujet.

7.2.4 Degré de dégradation de l'information auditive

Un dernier aspect des résultats de l'expérience 2 concerne l'effet du degré de dégradation de l'information auditive sur le traitement bimodal de la parole et sur le comportement oculaire dans ce cadre. Quant au traitement audio-visuel de la parole, les résultats relatifs au gain AV a révélé l'effet significatif du facteur SNR à ce niveau. Conformément aux résultats des études précédentes et ceux de l'expérience 1, le gain AV a été plus élevé dans la condition de dégradation forte que dans la condition de dégradation faible de l'information auditive. Ce pattern semble être stable de l'enfance tardive à l'âge adulte, suggérant davantage une relative maturité des mécanismes de traitement bimodal de la parole à l'enfance tardive. Quant à l'effet du degré de dégradation de l'information auditive par le bruit sur le comportement oculaire des participants lors de la perception audio-visuelle de la parole, es résultats ont mis en évidence l'effet significatif de l'interaction entre les facteurs SNR et AOI. Le pattern des différences inter-conditions impliquant les modalités de ces deux facteurs montrent certaines modifications du pattern caractérisant l'effet du facteur AOI précédemment décrit. Plus précisément, pour les deux niveaux de dégradation de l'information auditive, le pattern relatif à l'effet du facteur AOI a été observé. Au niveau du SNR-6, on constate également que la durée des fixations oculaires dans la région du nez a été plus longue que celle mesurée au niveau du contour et de l'intra-contour. De manière surprenante, pour le SNR-12, le contour de l'oratrice a été fixé plus longuement que les yeux. En outre, pour l'ensemble des formats AV, (i) le nez a été fixé moins longtemps dans la condition de dégradation forte que dans la condition de dégradation faible et (ii) aucune variation n'a été observée pour la région buccale. Ces résultats sont surprenants et vont à l'encontre de notre hypothèse ainsi que des résultats des études précédentes (Buchan et *al*., 2008 ; Vatikotis-Bateson et *al*., 1998). Il est possible que des stratégies visuo-attentionnelles impliquées dans le traitement des formats AV de l'expérience 2 aient été influencées par la présence d'un format à caractère (fortement) non écologique, le format AVV-EBA, et possiblement aussi par le caractère plus légèrement non écologique du format AVV-BA. Si tel est le cas, les mécanismes gérant le comportement oculaire impliqué dans le traitement audio-visuel de la parole semblent hautement sensibles aux caractéristiques de l'information faciale et notamment à tout biais dans son aspect écologique, qui pourrait avoir un effet dominant sur la gestion visuo-attentionnelle au détriment d'autres facteurs pourtant extrêmement importants pour la perception de la parole, tel que l'intelligibilité de l'information auditive. (En effet, l'interaction entre les facteurs SNR et Format n'a pas été significative.) Finalement, on constate également l'absence de l'interaction SNR x Groupe. Cette dimension des résultats est conforme à ce qui a été observé dans l'expérience 1. Elle

suggère ainsi que les mécanismes de l'adaptation du comportement oculomoteur en fonction du degré de l'intelligibilité de l'input auditif dans le cadre de la perception bimodale de la parole atteignent leur maturité avant l'enfance tardive.

8 Conclusion et ouvertures pour la recherche future

8.1 Conclusion générale

Dans leur ensemble, les résultats des expériences 1 et 2 nous ont permis de répondre aux questions de l'étude initialement posées, de considérer la validité de telles conclusions pour les populations, les mécanismes cognitifs et les phénomènes étudiés et de proposer des pistes pour la recherche future sur le plan de la problématique et de l'approche méthodologique pouvant être pertinente.

Les conclusions majeures de notre étude concernent d'une part, les caractéristiques de la perception audio-visuelle de la parole et du comportement oculaire sous-jacent à l'extraction de l'information visuelle dans ce cadre et, d'autre part, leur dépendance/leurs variations en fonction (i) des caractéristiques de l'information visuelle, faciale, (ii) du moment développemental se situant sur la trajectoire allant de l'enfance tardive à l'âge adulte et (iii) des caractéristiques de l'input auditif, notamment de son degré de l'intelligibilité.

Globalement, les résultats de notre étude montrent que le traitement de l'information faciale lors de la perception bimodale de la parole implique le traitement holistique du visage de l'orateur. Ce phénomène n'étant présent que chez les participants adultes de notre étude, les mécanismes cognitifs le sous-tendant semblent connaître un processus de maturation relativement longue, se terminant possiblement à l'adolescence tardive. Le traitement holistique de l'information faciale dans le contexte où l'objectif de l'individu adulte est de traiter les aspects acoustiques de la parole perturbe la perception audio-visuelle de la parole. Il est vraisemblable que le traitement holistique soit déclenché automatiquement chez les adultes en présence de l'information faciale et qu'il soit activement inhibé au profit d'une approche analytique permettant le traitement optimal des indices visuels les plus informatifs quant aux sons produits, qui sont apportés par la région buccale de l'orateur.

Les mouvements articulatoires provenant uniquement des régions extra-buccales de l'orateur semblent peu informatifs pour la reconnaissance des unités phonémiques de la production verbale orale. Etant dépourvu des informations en lien avec les articulateurs les plus

cruciaux quant aux caractéristiques acoustiques des sons (lèvres, langue, dents), ce type de mouvements, visuellement bien moins saillant que les lèvres de l'orateur, n'est informatif que sur les aspects temporels de la production de la parole et semble faciliter la perception de la parole dans les conditions où l'intelligibilité de l'input auditif est moyennement à fortement atteinte. Dans de telles conditions, l'information sur la dimension temporelle de la production verbale orale semble jouer un rôle important. Pour les stimuli présentant la région buccale de l'orateur, elle semble pouvoir être améliorée par l'application d'un masque qui introduit un contexte statique et de forts contrastes lumineux entre la région exposée (la bouche) et la région masquée (le reste du visage).

Les résultats relatifs à la dimension développementale des mécanismes du traitement bimodal de la parole suggèrent que ces derniers atteignent leur maturité avant l'enfance tardive. La seule variabilité dans le traitement bimodal de la parole observée sur le plan développemental était relative à la nature localisée (bouche) *vs* holistique (visage) de l'information visuelle, exposée précédemment. Elle est vraisemblablement liée au processus de maturation et d'automatisation des mécanismes impliqués dans le traitement holistique des visages.

Quant au comportement oculaire lors de la perception audio-visuelle de la parole bruitée, la bouche est apparue comme cible majeure du regard, le nez étant également une cible importante dans ce cadre. Pour l'information faciale dont les caractéristiques respectent celles des visages rencontrés dans les conditions écologiques, une augmentation de l'incertitude quant à l'input auditif provoque essentiellement une diminution dans la durée des fixations de la région des yeux au profit de la région buccale. Le mode de présentation de l'information faciale attirant le plus fortement le regard vers la région buccale de l'orateur est celui qui est réduit à la région en question et ne comporte pas de contrastes visuels entre la région buccale et l'arrière-plan. L'attrait du regard sur la région buccale de l'orateur semble être une stratégie efficace pour le traitement bimodal des aspects acoustiques de la parole, mais uniquement dans les cas où l'input auditif reste peu dégradé. Chez les adultes, l'information faciale relative à la production de la parole ainsi que d'autres types de l'information faciale possiblement relatifs à la configuration des composantes faciales semblent être bien encodés et ensuite traités également en vision périphérique.

Finalement, le comportement oculaire dans la perception audio-visuelle de la parole et la gestion visuo-attentionnelle le sous-tendant ne présentent pas de variations sur la trajectoire développementale allant de l'enfance tardive à l'âge adulte pour les stimuli visuels comportant

le visage entier de l'orateur dont les caractéristiques sont conformes à ce type des stimuli rencontrés dans les conditions écologiques de la production de la parole. En revanche, l'atteinte au caractère écologique du visage de l'orateur semble provoquer des variations développementales essentiellement au niveau de la région du nez, des régions extra-buccales et du contour facial. La gestion du comportement oculomoteur lors de la perception audio-visuelle de la parole pour un type de stimuli visuels donné pourrait être sensible aux caractéristiques d'autres types de stimuli visuels présentés dans l'expérience.

8.2 Considérations méthodologiques

Les expériences menées dans le cadre du présent travail de thèse ont contribué à l'enrichissement des connaissances sur la perception audio-visuelle de la parole en s'intéressant aux caractéristiques de l'input visuel et à l'aspect développemental dans ce domaine. L'aspect développemental couvrant la période allant de l'enfance à l'âge adulte ayant été très peu étudié au moyen du paradigme de dégradation de l'information auditive par le bruit, notre étude a mis en évidence certaines contraintes méthodologiques qui peuvent servir de guide pour la recherche future. Premièrement, nous avons constaté que la production verbale des enfants (répétitions des items perçus) a été bien moins intelligible et facile à catégorisée que celle des participants plus âgés. Il semble ainsi important d'introduire le principe de codage des réponses par des juges indépendants et aveugles quant aux objectifs et hypothèses de l'étude afin d'éviter tout biais lié à l'expérimentateur. Deuxièmement, le paradigme de la dégradation de l'information auditive par le bruit est souvent utilisé dans le cadre d'un design expérimental à mesures répétées. Dans ce type de design, le protocole expérimental est souvent long et peut éventuellement poser un problème de gestion attentionnelle pour les enfants. Dans les expériences où les précautions sont faites pour s'assurer de la qualité des données recueillies, comme cela a été le cas dans notre étude, les critères méthodologiques définissant la qualité acceptable des données peuvent induire un biais dans la constitution de l'échantillon des participants jeunes. En effet, ces derniers pourraient éventuellement présenter des capacités attentionnelles plus élevées que la moyenne de la population étudiée et, par conséquent, possiblement induire un biais dans les résultats en atteignant leur validité externe. Nous préconisons ainsi la mise en place de protocoles plus courts pour ce type de population.

Quant à la dimension relative aux caractéristiques des stimuli visuels et leur éventuelle influence sur la perception bimodale de la parole et le comportement oculomoteur dans ce

cadre, les résultats de notre étude suggèrent que la nature localisée *vs* holistique de l'information faciale pourraient influencer le traitement bimodal de la parole chez les adultes. Par ailleurs les caractéristiques visuelles de bas niveau du format de présentation de l'information localisée semblent affecter le traitement audio-visuel de la parole également chez les participants plus jeunes. Le pattern de ces influences, tel que mis en évidence dans notre étude, est ainsi à prendre en compte dans la mise en place des stimuli visuels en fonction des questions d'étude. Finalement, le contexte, en termes des différentes conditions de présentation de l'information visuelle, semble impacter le comportement oculomoteur lors de la perception audio-visuelle de la parole. Il nous semble ainsi pertinent d'éviter le design expérimental à mesures répétées dans les études portant sur l'effet des caractéristiques des stimuli visuels sur le comportement oculomoteur dans le cadre en question.

8.3 Ouvertures pour la recherche future

Le présent travail de thèse offre différentes pistes pour la recherche future. Premièrement, la problématique relative à la nature holistique (visage entier) *vs* localisée (bouche) de l'information faciale et son impact sur la perception audio-visuelle de la parole pourrait être étudiée avec d'autres types de population. Deux types de populations nous semblent particulièrement intéressantes, notamment les personnes autistes et les personnes dyslexiques. Les personnes autistes sont considérées comme présentant des spécificités dans le traitement des visages avec, comme caractéristique principale, l'évitement de fixer la région des yeux (pour une revue, voir Guillon, Hadjikhani, Baudel, & Roré, 2014). Une des hypothèses expliquant ce phénomène est celle de l'anxiété sociale (Takana & Tsung, 2013) selon laquelle la région des yeux serait évitée car son traitement provoque un état d'anxiété chez la personne autiste. Par ailleurs, les personnes autistes présenteraient également des déficits dans la perception bimodale de la parole (Woynaroski, Kwakye, Foss-Feig, Stevenson, Stone, & Wallace, 2014). Quant aux personnes dyslexiques, elles se caractérisent par un déficit de la mise en place et de l'utilisation des représentations phonologiques (Vellutino, Fletcher, Snowling, & Scanlon, 2004). Les personnes dyslexiques présenteraient également un déficit au niveau de la perception bimodale (Hahn, Foxe, & Molholm, 2014). Il serait ainsi intéressant d'explorer (i) si et dans quelle mesure des variations dans la nature holistique de l'information faciale et les variations dans les caractéristiques de bas niveau des stimuli visuels réduits à la région buccale de l'orateur (qui semblent affecter le traitement visuo-attentionnel de l'information faciale ainsi établie) influence la perception bimodale de la parole dans ces

populations et (ii) quelles sont les éventuelles variations développementales du traitement des différents formats de présentation de l'information visuelle chez ces populations.

Les résultats de notre étude ayant mis en évidence une certaine stabilité du comportement oculomoteur dans la perception de la parole pour la période allant de l'enfance tardive à l'âge adulte. Il serait intéressant d'étudier cet aspect comportemental également avec des participants plus jeunes. Par ailleurs, la nature de notre tâche était telle que le traitement de la parole a été restreint à ses aspects purement acoustiques pour lequel la région buccale semble en effet porteuse des indices les plus utiles/informatifs. En proposant un autre type de tâche, impliquant le traitement sémantique du message verbal oralement produit, en prenant en compte également l'aspect prosodique de ce message, les indices visuels d'autres régions faciales seraient certainement informatifs pour la perception et la compréhension de la parole. Dans une telle situation, les caractéristiques du comportement oculomoteur risqueraient de différer grandement de celles observées dans notre étude. Il est également probable que l'on aurait observé une évolution au cours du développement pour la période prise en compte dans le présent travail de thèse.

La dimension des résultats relative à l'importance de la détection du mouvement (notamment des mouvements préparatoires) pour la perception bimodale de la parole fortement bruitée pourrait également être étudiée avec une approche différente, impliquant la méthode d'enregistrement des potentiels locaux au moyen de l'EEG. En effet, si l'interprétation de nos résultats est juste, c'est-à-dire si le format AVB-M a réellement aidé la perception de la parole car il a facilité la détection des mouvements articulatoires, ce qui, à termes, a rehaussé la sensitivité perceptive au signal auditif, cela a dû affecter le mécanisme d'ajustement de phase dans l'activité oscillatoire entre le V1 et le A1.

Une autre dimension des résultats de notre étude qui offre une ouverture pour la recherche future concerne l'effet du format AVB-E sur les mécanismes visuo-attentionnels des participants. En effet, dans la mesure où le format AVB-E a globalement attiré le mieux l'attention visuelle des participants à la région buccale de l'oratrice, il serait également intéressant de tester l'efficacité de ce type de présentation de l'information visuelle dans le cadre d'entraînements phonologiques pour les enfants préscolaires, les enfants présentant des difficultés de traitement phonologique ou pour les adultes apprenant une langue seconde. De tels entraînements pourraient intégrer le mode AVB-E de présentation visuelle en supplément de la présentation AVV, la présentation AVB-E se faisant, par exemple, sous forme de feed-back.

Finalement, la généralisation des résultats de notre étude semble problématique pour la population des enfants âgés de 7 à 9 ans en raison du caractère possiblement non représentatif de notre échantillon quant aux domaines de l'attention et de la production verbale orale. Il serait de ce fait intéressant d'étudier un éventuel lien entre les deux domaines en question et la perception audio-visuelle de la parole et d'identifier les différents patterns de trajectoire développementale pouvant possiblement être établis à la base des trois dimensions en question.

9 Références bibliographiques

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

Adank, P., Rueschemeyer, S., & Bekkering, H. (2013). The role of accent imitation in sensorimotor integration during processing of intelligible speech. *Frontiers in Human Neuroscience, 7*, 634.

Aloufy, S., Lapidot, M., & Myslobodsky, M. (1996). Differences in susceptibility to the "blending illusion" among native Hebrew and English speakers. *Brain and Language, 53*, 51–57.

Alsius, A., Navarra, J., & Soto-Faraco, S. (2007). Attention to touch reduces audiovisual speech integration. *Experimental Brain Research, 183*, 399–404.

Alsius, A., Möttönen, R., Sams, M. E., Soto-Faraco, S., & Tiippana, K. (2014). Effect of attentional load on audiovisual speech perception: Evidence from ERPs. *Frontiers in Psychology, 5*, 2–9.

Altieri, N. (2010). *Toward a unified theory of audiovisual integration in speech perception*. Bloomington, IN: Indiana University.

Altieri, N., Pisoni, D. B., & Townsend, J. T. (2011). Some behavioral and neurobiological constraints on theories of audiovisual speech integration: A review and suggestions for new directions. *Seeing Perceiving, 24*, 513–539.

Altieri, N., & Townsend, J. T. (2011). An assessment of behavioral dynamic information processing measures in audiovisual speech perception. *Frontiers in Psychology, 2*, 238.

Arlinger, S. (1991). Results of visual information processing tests in elderly people with presbycusis. *Acta Oto-Laryngologica, 111*, 143–148.

Arnal L. H., Morillon B., Kell C. A., & Giraud A.-L. (2009) Dual neural routing of visual facilitation in speech processing. *The Journal of Neuroscience, 29*, 13445–13453.

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

Auer, E. T., Jr. (2002). The influence of the lexicon on speech read word recognition: Contrasting segmental and lexical distinctiveness. *Psychonomic Bulletin and Review, 9*, 341–347.

Baart, M., Stekelenburg, J. J., & Vroomen, J. (2014). Electrophysiological evidence for speech-specific audiovisual integration. *Neuropsychologia, 53*, 115–121.

Badin, P., Tarabaka, Y., Elisei, F., & Bailly, G. (2010). Can you "read" tongue movements? Evaluation of the contribution of tongue display to speech understanding. *Speech Communication, 52*, 493–503.

Bahrick, L. E., & Lickliter, R. (2000). Intersensory redundancy guides attentional selectivity and perceptual learning in infancy. *Developmental Psychology, 36*, 190–201.

Bahrick, L. E., & Lickliter, R. (2002). Intersensory redundancy guides early perceptual and cognitive development. In R. Kail (Ed.), *Advances in child development and behavior* (Vol. 30, pp. 153–187). New York, NY: Academic Press.

Bahrick, L. E., & Lickliter, R. (2009). Perceptual development: Intermodal perception. In B. Goldstein (Ed.), *Encyclopedia of Perception, Vol. 2*, (pp. 753-756). Newbury Park, CA: Sage Publishers.

Bahrick, L. E., & Lickliter, R. (2012). The role of intersensory redundancy in early perceptual, cognitive, and social development. In A. Bremner, D. J. Lewkowicz & C. Spence (Eds.), *Multisensory development* (pp. 183–206). Oxford, England: Oxford University Press.

Bahrick, L. E., & Lickliter, R. (2014). Learning to attend selectively. The dual role of intersensory redundancy. *Current Directions in Psychological Science, 23*, 414–420.

Bartlett, J. C., & Searcy, J. (1993). Inversion and configuration of faces. *Cognitive Psychology, 25*, 281–316

Barutchu, A., Crewther, D. P., & Crewther, S. G. (2009). The race that precedes coactivation: Development of multisensory facilitation in children. *Developmental Science, 12*, 464–473.

Barutchu, A., Danaher, J., Crewther, S. G., Innes-Brown, H., Shivdasani, M. N., & Paolini, A. G. (2010). Audiovisual integration in noise by children and adults. *Journal of Experimental Child Psychology, 105*, 38–50.

Baum, S. H., Martin, R. C., Hamilton, A. C., & Beauchamp, M. S. (2012). Multisensory speech perception without the left superior temporal sulcus. *NeuroImage, 62*, 1825–1832.

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*, 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*, 809–823.

Bebko, J. M., Schroeder, J. H., & Weiss, J. A. (2014). The McGurk effect in children with autism and Asperger syndrome. *Autism Research, 7*, 50–59.

Benevento L. A., Fallon J. H., Davis B., & Rezak, M. (1977). Auditory-visual interaction in single cells in the cortex of the superior temporal sulcus and the orbital frontal cortex of the macaque monkey. *Experimental Neurology, 57*, 849–872.

Benjamin, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B (Methodological), 57*, 289–300.

Benoït, C., Guiard-Marigny, B., Le Goff, B., & Adjoudani, A. (1996). Which components of the face do humans and machines best speech read? In D. G. Stork & M. E. Hennecke (Eds.), *Speechreading by humans and machines: Models, systems, and applications* (pp. 315–328). Berlin, Germany: Springer-Verlag.

Benoît, C., Mohamadi, T., & Kandel, S. (1994). Effects of phonetic context on audio-visual intelligibility of French. *Journal of Speech and Hearing Research, 37*, 1195–1203.

Berg, W. P., Berglund, E. D., Strang, A. J., & Baum, M. J. (2007). Attention-capturing properties of high frequency luminance flicker: Implications for brake light conspicuity. *Transportation Research Part F, 10*, 22–32.

Berger, K. W., Garner, M., & Sudman, J. (1971). The effect of degree of facial exposure and the vertical angle of vision on speechreading performance. *Teacher of the Deaf, 69*, 322–326.

Bernstein, L. E. (2005). Phonetic processing by the speech perceiving brain. In D. B. Pisoni & R. E. Remez (Eds.), *Handbook of speech perception* (pp. 79–98). Malden, MA: Blackwell.

Bernstein, L. E., Auer, E. T., & Moore, J. K. (2004). Audiovisual speech binding: cConvergence or association? In G. A. Calvert, C. Spence & B. E. Stein (Eds.), *The handbook of multisensory perception* (pp. 203–224). Cambridge, MA: MIT Press.

Bernstein, L. E., Auer, E. T., & Takayanagi, S. (2004). Auditory speech detection in noise enhanced by lipreading. *Speech and Communication, 44*, 5–18.

Berryhill, M., Kveraga, K., Webb, L., & Hughes, H. C. (2007).Multimodal access to verbal name codes. *Perception & Psychophysics, 69*, 628–640.

Besle, J., Fort, A., Delpuech, C., & Giard, M.-H. (2004). Bimodal speech: Early visual effect in the human auditory cortex. *European Journal of Neuroscience, 20*, 2225–2234.

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding.

Psychological Review, 94, 115–147.

Binnie, C. A., Montgomery, A. A., & Jackson, P. L. (1974). Auditory and visual contribution to the perception of consonants. *Journal of Speech and Hearing Research. 17*, 619–630.

Birch, H. G., & Lefford, A. (1963). Intersensory development in children. *Monographs of the Society for Research in Child Development, 25*, 1–48.

Bishop, C. W., & Miller, L.M. (2009). A multisensory cortical network for understanding speech in noise. *Journal of Cognitive Neuroscience, 21*, 1790–1805.

Blamey, P., Cowan, R., Alcantara, J., Whitford, L., & Clark, G. (1989). Speech perception using combinations of auditory, visual, and tactile information. *Journal of Rehabilitation Research and Development, 26*, 15–24.

Borrie, S. A. (2015). Visual speech information: A help or hindrance in perceptual processing of dysarthric speech. *The Journal of the Acoustical Society of America, 137*, 1473–1480.

Bosch, L., & Sebastian-Galles, N. (2001). Early language differentiation in bilingual infants. In J. Cenoz & F. Genesee (Eds.), *Trends in bilingual acquisition* (pp. 71–93). Amsterdam: John Benjamins Publishing Company.

Bower, T. G. R. (1974). *Development in infancy*. San Francisco: W.H. Freeman.

Braddick, O., & Atkinson, J. (2011). Development of human visual function. *Vision Research, 51*, 1588–1609.

Braida, L. D. (1991). Crossmodal integration in the identification of consonant segments. *Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology, 43*, 647–677.

Bremner, A., Lewkowicz, D. J., & Spence, C. (2012). Multisensory development. New York, NY: Oxford University Press.

Brandwein, A. B., Foxe, J. J., Russo, N. N., Altschuler, T. S., Gomes, H., & Molholm, S. (2011). The development of audiovisual multisensory integration across childhood and early adolescence: A high-density electrical mapping study. *Cerebral Cortex, 21*, 1042–1055.

Brenna, V., Nava, E., Turati, C., Montirosso, R., Cavallini, A., & Borgatti, R. (2015). Intersensory redundancy promotes visual rhythm discrimination in visually impaired infants. *Infant Behavior & Development, 39*, 92–97.

Bristow, D., Dehaene-Lambertz, G., Mattout, J., Soares, C., Gliga1, T., Baillet, S., & Mangin, J.-F. (2009). Hearing faces: How the infant brain matches the face it sees with the speech it hears. *Journal of Cognitive Neuroscience, 21*, Pages 905–921.

Bronson, G. W. (1994). Infants' transitions toward adult-like scanning. *Child Development, 65*, 1243–1261.

Bruce, V., Campbell, R. N., Doherty-Sneddon, G., Import, A., Langton, S., McAuley, S., & Wright, R. (2000). Testing face processing skills in children. *British Journal of Developmental Psychology, 18*, 319–333.

Bruce, C., Desimone, R., & Gross, C. G. (1981). Visual properties of neurons in a polysensory area in superior temporal sulcus of the macaque. *Journal of Neurophysiology, 46*, 369–384.

Bruce, V., & Young, A.W. (2012). *Face perception*. Psychology Press, London; New York.

Buchan, J., N., Paré, M., & Munhall, K. G. (2008). The effect of varying talker identity and listening conditions on gaze behavior during audiovisual speech perception. *Brain Research, 1242*, 162–171.

Buchel, C., & Friston, K. (2001). Interactions among neuronal systems assessed with functional neuroimaging. *Revue Neurologique (Société Neurologique de Paris), 157*, 807–815.

Burke, S. N., & Barnes, C. A. (2006). Neural plasticity in the aging brain. Nature Reviews Neuroscience, *7*, 30–40.

Burnham, D., & Dodd, B. (1998). Familiarity and novelty in infant cross- language studies: Factors, problems, and a possible solution. In C. Rovee-Collier & H. Hayne (Eds.), *Advances in infancy research* (pp. 170–187). New York, NY: Greenwood Publishing Group.

Burnham, D., & Dodd. B. (2004). Auditory-visual speech integration by prelinguistic infants: perception of an emergent consonant in the McGurk effect. *Developmental Psychobiology, 45*, 204–220.

Buzsáki, G., & Draguhn, A. (2004). Neuronal oscillations in cortical networks. *Science, 304*, 1926–1929.

Cabbage, K., & Carell, T. (2014). The relationship between speech perception and production: Evidence from children with speech production errors. *Journal of Acoustical Society of America, 135*, 2420.

Calderone, D. J., Lakatos, P., Butler, P. D., & Castellanos, F. X. (2014). Entrainment of neural oscillations as a modifiable substrate of attention. *Trends in Cognitive Science, 18,* 300–309.

Callan, D., Callan, A., Gamez, M., Sato, M., & Kawato, M. (2010). Premotor cortex mediates perceptual performance. *Neuroimage, 51,* 844–858.

Callan, D., Callan, A., & Jones, J. A. (2014). Speech motor brain regions are differentially recruited during perception of native and foreign-accented phonemes for first and second language listeners. *Frontiers in Neuroscience, 8*, 275.

Callan, D. E., Jones, J. A., Callan, A. M., & Akahane-Yamada, R. (2004). Phonetic perceptual identification by native- and second-language speakers differentially activates brain regions involved with acoustic phonetic processing and those involved with articulatory auditory/orosensory internal models. *NeuroImage, 22*, 1182–1194.

Callan, D. E., Jones, J. A., Munhall, K., Callan, A. M., Kroos, C. & Vatikiotis-Bateson, E. (2003). Neural processes underlying perceptual enhancement by visual speech gestures. *Neuroreport, 14*, 2213–2218.

Calvert, G. A., Bullmore, E., Brammer, M. J., Campbell, R., Iversen, S. D., Woodruff, P., … David, A.S. (1997). Silent lipreading activates the auditory cortex. *Science, 276*, 593-596.

Calvert, G. A., Bullmore, E. T., Brammer, M., Campbell, R., Williams, S., McGuire, P. K., ... David, A. S. (1997). Activation of auditory cortex during silent lipreading. *Science, 276,* 593–596.

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.

Calvo, M. G., Fernández -Martin, A., & Nummenmaa, L. (2014). Recognition of facial expressions in peripheral versus central vision: Role of the eyes and the mouth. *Psychological Research, 78*, 180–195.

Campbell, R. (1992). The neuropsychology of lipreading. *Philosophical Transactions of the Royal Society of London, B 335*, 39–45.

Campbell, R., Dodd, B., & Burnham, D. (Eds.). (1998). *Hearing by eye II: Advances in the psychology of speechreading and auditory-visual speech*. Hove, England: Psychology Press.

Campbell, R., Landis, T., Regard, M., 1986. Face recognition and lipreading: A neurological dissociation. *Brain, 109*, 509–521.

Campi, K. L., Bales, K. L., Grunewald, R., & Krubitzer, L. (2010). Connections of auditory and visual cortex in the Prairie Vole (*Microtus ochrogaster*): Evidence for multisensory processing in primary sensory areas. *Cerebral Cortex, 20*, 89–108.

Capek, C. M., Bavelier, D., Corina, D., Newman, A. J., Jezzard, P., & Neville, H. J. (2004). The cortical organization of audio-visual sentence comprehension: an fMRI study at 4 Tesla. *Cognitive Brain Research, 20*, 111–119.

Cappe, C., & Barone, P. (2005). Heteromodal connections supporting multisensory integration at low levels of cortical processing in the monkey. *European Journal of Neuroscience, 22*, 2886–2902.

Carbon, C.-C., Schweinberger, S. R., Kaufmann, J. M., & Leder, H. (2005). The Thatcher illusion seen by the brain: An event-related brain potentials study. *Cognitive Brain Research, 24*, 544–555.

Carey, S., Diamond, R., & Woods, B. (1980). Development of face recognition: A maturational component? *Developmental Psychology, 16*, 257–269.

Cavé, C., Stroumza, A., & Bastien-Toniazzo, M. (2007). The McGurk effect in dyslexic and normal-reading children: An experimental study. In *Proceedings of The Auditory-Visual Speech Processing 2007, (AVSP 2007*), 31 August - 3 September (pp. 54–58). Hilvarenbeek, The Netherlands.

Cavedon, A. (1980). Contorno e disparazione retinica come determinanti della localizzazione in profondita le condizioni della percezione di un foro [Contour and retinal displacement as determinants of localization in depth of the conditions of the perception of a hole]. *Universita di Padova Instituto di Psicologia Report, 12*, 1–21.

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., & Ghazanfar, A. A. (2009). The natural statistics of audiovisual speech. *PLoS Computational Biology, 5*, e1000436.

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *The Journal of the Acoustical Society of America, 25*, 975–979.

Cienkowski, K. M., & Carney, A. E. (2002). Auditory–visual speech perception and aging. *Ear and Hearing, 23*, 439–49.

Clark, J., Yallop, C., & Fletcher, J. (2007). *An introduction to phonetics and phonology* (3rd ed). Hoboken, NJ: Wiley-Blackwell.

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, & Behavioral Neuroscience, 4*, 117–126.

Conrey, B., & Pisoni, D.B. (2006). Auditory–visual speech perception and synchrony detection for speech and nonspeech signals. *Journal of Acoustical Society of America, 119*, 4065–4073.

Cornett, O. (1967). Cued speech. *American Annals of the Deaf, 112*, 3–13.

Cravo, A. M., Rohenkohl, G., Wyart, V., & Nobre, A. C. (2013). Temporal expectation enhances contrast sensitivity by phase entrainment of low-frequency oscillations in visual cortex. *The Journal of Neuroscience, 33*, 4002–4010.

Davis, C., & Kim, J. (2004). Audio-visual interactions with intact clearly audible speech. *Quarterly Journal of Experimental Psychology, 57*, 1103–1121.

Davis, E., & Kim, J. (2006). Audio-visual speech perception off the top of the head. *Cognition, 100*, B21–B31.

de Gelder, B., & Bertelson, P. (2003). Multisensory integration, perception and ecological validity. *Trends in Cognitive Sciences, 7*, 460–467.

de Gelder, B., & Vroomer, J. (1998). Impairment of speech-reading in prosopagnosia. *Speech Communication, 26*, 89–96.

de Marco, G., Vrignaud, P., Destrieux, C., de Marco, D., Testelin, S., Devauchelle, B., & Berquin, P. (2009). Principle of structural equation modeling for exploring functional interactivity within a putative network of interconnected brain areas. *Magnetic Resonance Imaging, 27*, 1–12.

Denes, P. B., & Pinson, E. N. (1993). *The speech chain: The physics and biology of spoken language* (2nd edition). Oxford: W. H. Freeman and Company.

Desjardins, R. N. & Werker, J. F. (1996). 4-month-old female infants are influenced by visible speech. Poster presented at *Xth Biennial International Conference of Infant Studies*, 18-21 April, Providence, RI.

Desjardins, R. N., Rogers, J., & Werker, J. F. (1997). An exploration of why preschoolers perform differently than do adults in audiovisual speech perception tasks. *Journal of Experimental Child Psychology, 66*, 85–110.

Desjardins, R. N., & Werker, J. W. (2004). Is the integration of heard and seen speech mandatory for infants? *Developmental Psychobiology, 45*, 187–203.

Diamond, R., & Carey, S. (1986). Why faces are and are not special: An effect of expertise. *Journal of Experimental Psychology: General, 115*, 107–117.

Dick, A. S., Solodkin, A., & Small, S. L. (2009). Neural development of networks for audiovisual speech comprehension. *Brain & Language, 114*, 101–114.

Di Pellegrino, G., Fadiga, L., Fogassi, L., Gallese, V., & Rizzolatti, G (1992). Understanding motor events: A neurophysiological study. *Experimental Brain Research, 91*, 176–180.

Dodd, B. (1979) Lip reading in infants - Attention to speech presented in- synchrony and out-of-synchrony. *Cognitive Psychology, 11*, 478–484.

Dohen, M. (2009). Speech Through the ear, the eye, the mouth and the hand. In Esposito A., A. Hussain & M. Marinaro (Eds), *Multimodal signals: Cognitive and algorithmic issues* (pp. 24-39). Berlin/Heidelberg: Springer.

Donnelly, N., Humphreys, G. W., & Sawyer, J. (1994). Stimulus factors affecting the categorization of faces and scrambled faces. *Acta Psychologica, 85*, 219–234.

Dubois, J., Dehaene-Lambertz, G., Perrin, M., Mangin, J. F., Cointepas, Y., Duchesnay, E., … Hertz-Pannier, L. (2008). Asynchrony of the early maturation of white matter bundles in healthy infants: Quantitative landmarks revealed noninvasively by diffusion tensor imaging. *Human Brain Mapping, 29*, 14–27.

Dupont, S., Aubin, J., & Ménard, L. (2005). A study of the McGurk effect in 4 and 5-year-old French Canadian children. *ZAS Papers in Linguistics 40*, 1–17.

Dupont, S., & Luettin, J. (2000). Audio-visual speech modeling for continuous speech recognition. *IEEE Transactions on Multimedia, 2*, 141–151.

Eckert, M. A., Kamdar, N. V., Chang, C. E., Beckmann, C. F., Greicius, M. D., & Menon, V. (2008). A cross-modal system inking primary auditory and visual cortices: Evidence from intrinsic fMRI connectivity analysis. *Human Brain Mapping, 29*, 848–857.

Elliott, L. L. (1979). Performance of children aged 9 to 17 years on a test of speech intelligibility in noise using sentence material with controlled word predictability. *The Journal of the Acoustical Society of America, 66*, 651–653.

Eramudugolla, R., Hendrson, R., & Matingley, J. B. (2010). Effects of audio–visual integration on the detection of masked speech and non-speech sounds. *Brain and Cognition, 75*, 60–66.

Erber, N. (1969). Interaction of audition and vision in the recognition of oral speech stimuli. *Journal of Speech and Hearing Research, 12*, 423–425.

Erber, N. P. (1971) Auditory and audiovisual reception of words in low frequency noise by children with normal hearing and by children with impaired hearing. *Journal of Speech, Language and Hearing Research, 143*, 496–512.

Erber, N. P. (1975) Auditory–visual perception in speech. *Journal of Speech and Hearing Disorders, 40*, 481–492.

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

Eskelund, K., MacDonald, E. N., & Andersen, T. S. (2015). Face configuration affects speech perception: Evidence from a McGurk mismatch negativity study. *Neuropsychologia, 66*, 48–54.

Everdell, I. T., Marsh, H., Yurick, M., Munhall, K. G., & Paré, M. (2007). Gaze behavior in audiovisual speech perception: Asymmetrical distribution of face-directed fixations. *Perception, 36*, 1535–1545.

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

Feld, J., & Sommers, M. S. (2011). There goes the neighborhood: Lipreading and the structure of the mental lexicon. *Speech Communication, 53*, 220–228.

Fernández, L. M., Visser, V., Ventura-Campos, N., Ávila, C., & Soto-Faraco, S. (2015). Top-down attention regulates the neural expression of audiovisual integration, *NeuroImage, 119*, 272–285.

Field, J., Muir, D., Pilon, R., Sinclair, M., & Dodwell, P. (1980). Infants' orientation to lateral sounds from birth to three months. *Child Development, 51*, 295–298.

Flom, R., & Bahrick, L. E. (2007). The development of infant discrimination of affect in multimodal and unimodal stimulation: The role of intersensory redundancy. *Developmental Psychology, 43*, 238–252.

Fowler, C. A., & Dekle, D. J. (1991). Listening with eye and hand: Cross-modal contributions to speech perception. *Journal of Experimental Psychology: Human Perception and Performance, 17*, 816–828.

Freiherr, J., Lundström, J. N., Habel, U., & Reetz, K. (2013). Multisensory integration mechanisms during aging. *Frontiers in Human Neuroscience, 7*, 863.

Fridriksson, J., Moss, J., Davis, B., Baylis, G. C., Bonilha, L., & Rorden, C. (2008). Motor speech perception modulates the cortical language areas. *NeuroImage, 41*, 605–613.

Furl, N., Garrido, L., Dolan, R. J., Driver, J., & Duchain, B. (2011). Fusiform gyrus face selectivity relates to individual differences in facial recognition ability. *Journal of Cognitive Neuroscience, 23*, 1723–1740.

Galantucci, B., Fowler, C. A., & Turvey, M. T. (2006). The motor theory of speech perception reviewed. *Psychonomic Bulletin & Review, 13*, 361–377.

Gallese, V., Fadiga, L., Fogassi, L., & Rizzolatti, G. (1996). Action recognition in the premotor cortex. *Brain, 119*, 593–609.

George, N., Evans, J., Fiori, N., Davidoff, J., & Renault, B. (1996). Brain events related to normal and moderately scrambled faces. *Cognitive Brain Research, 4*, 65–76.

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*, 5004–5012.

Ghitza, O. (2011). Linking speech perception and neurophysiology: Speech decoding guided by cascaded oscillators locked to the input rhythm. *Frontiers in Psychology, 2*, 130.

Ghitza, O., Giraud, A.-L., & Poeppel, D. (2013). Neuronal oscillations and speech perception: Critical-band temporal envelopes are the essence. *Frontiers in Human Neuroscience, 6*, 340.

Gibson, E. J. (1969). *Principles of perceptual learning and development*. New York, NY: Appleton.

Gibson, E. J. (1984). Perceptual development from the ecological approach. In M. E. Lamb, A. L. Brown & B. Rogoff (Eds.), *Advances in developmental psychology* (pp. 243–86). Hillsdale New Jersey: Erlbaum.

Gick, B., Wilson, I., & Derrick, D. (2013). *Articulatory phonetics*. Hoboken, NJ: Wiley-Blackwell.

Giraud, A.-L., & Poeppel, D. (2012). Cortical oscillations and speech processing: Emerging computational principles and operations. *Nature Neuroscience, 15*, 511–517.

Glisky, E. L. (2007). Changes in cognitive function in human aging. In D. R. Riddle (Ed.), *Brain aging: Models, methods and mechanisms* (pp. 4–20). Boca Raton, FL: CRC Press.

Goldstein, A. G. (1965). Learning of inverted and normally oriented faces in children and adults. *Psychonomic Science, 3,* 447–448.

Gordon, M., & Allen, S. (2009). Audiovisual speech in older and younger adults: Integrating a distorted visual signal with speech in noise. *Experimental Aging Research, 35,* 202–219.

Gosselin, P. A., & Gagné, J.-P. (2011). Older adults expend more listening effort than younger adults recognizing audiovisual speech in noise. *International Journal of Audiology, 50,* 786–792.

Grant, K. W. (2001). The effect of speechreading on masked detection thresholds for filtered speech. *The Journal of the Acoustical Society of America, 109*, 2272–2275.

Grant, K. W. (2002). Measures of auditory-visual integration for speech understanding: A theoretical perspective. *The Journal of Acoustical Society of America, 112*, 30–33.

Grant, K. W., & Seitz, P. F. (1998). Measures of auditory-visual integration in nonsense syllables and sentences. *Journal of the Acoustical Society of America, 104*, 2438–2450.

Grant, K. W., & Seitz, P. F. (2000). The use of visible cues for improving auditory detection of spoken sentences. *Journal of the Acoustical Society of America, 108,* 1197–1208.

Grant, K. W., Walden, B. E., & Seitz, P. F. (1998). Auditory-visual speech recognition by hearing-impaired subjects: consonant recognition, sentence recognition, and auditory-visual integration. *Journal of the Acoustical Society of America, 103*, 2677–2690.

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 & Psychophysics, 50,* 524–536.

Greenberg, H. J., & Bode, D. L. (1968). Visual discrimination of consonants. *Journal of Speech and Hearing Research, 11,* 869–874.

Greenwood, P. M., & Parasuraman, R. (2010). Neuronal and cognitive plasticity: A neurocognitive framework for ameliorating cognitive aging. *Frontiers in Aging Neuroscience, 2*, 150.

Guillon, Q., Hadjikhani, N., Baudel, S., & Roré, B. (2014). Visual social attention in autism spectrum disorder: Insights from eye tracking studies. *Neuroscience and Biobehavioral Reviews, 42*, 279–297.

Hadad, B. S., Maurer, D., & Lewis, T. L. (2010). The development of contour interpolation: Evidence from subjective contours. *Journal of Experimental Child Psychology, 106*, 163–176.

Hall, D. A., Fussell, C., & Summerfield, A. Q. (2005). Reading fluent speech from talking faces: Typical brain networks and individual differences. *Journal of Cognitive Neuroscience, 17*, 939–953.

Hahn, N., Foxe, J. J., & Molholm, S. (2014). Impairments of multisensory integration and cross-sensory learning as pathways to dyslexia. *Neuroscience and Biobehavioral Reviews, 47*, 384–392.

Hayashi, Y., & Sekiyama, K. (1998). Native-foreign language effect in the McGurk effect: A test with Chinese and Japanese. Poster presented at *Xth Biennial Auditory-Visual Speech Processing (AVSP'98)*, 4-6 December (pp. 61–66). Sydney, Australia.

Hazan, V., Kim, J., & Chen, Y. (2010). Audiovisual perception in adverse conditions: Language, speaker and listener effects. *Speech Communication, 52*, 996–1009.

Hauser, M. D. (1996). *The evolution of communication*. Cambridge, Mass: MIT Press.

Henry, M. J., & Herrmann, B. (2014). Low-frequency neural oscillations support dynamic attending in temporal context. *Timing & Time Perception, 2*, 62–86.

Henry, M. J., & Herrmann, B., & Obleser, J. (2014). Entrained neural oscillations in multiple frequency bands comodulate behavior. *Proceedings of the National Academy of Science, 111*, 14935–14940.

Henry, M. J., & Obleser, J. (2012). Frequency modulation entrains slow neural oscillations and optimizes human listening behavior. *Proceedings of the National Academy of Sciences of the United States of America, 109*, 20095–20100.

Hetu, R., Truchon-Gagnon, C. & Bilodeau, S. (1990) Problems of noise in school settings: a review of the literature and the results of an explorative study. *International Journal of Speech* & *Language Pathology* and *Audiology, 14*, 31–39.

Hickok, G. (2010). The role of mirror neurons in speech and language processing. *Brain and Language, 112*, 1–2.

Hietanen, J. K., Manninen, P., Sams, M., & Surakka, V. (2001). Does audiovisual speech perception use information about facial configuration? *European Journal of Cognitive Psychology, 13*, 395–407.

Hikosaka, K., Iwai, E., Saito, H., & Tanaka, K. (1988). Polysensory properties of neurons in the anterior bank of the caudal superior temporal sulcus of the macaque monkey. *Journal of Neurophysiology, 60*, 1615–1637.

Hillock-Dunn, A., & Wallace, M. T. (2012). Developmental changes in the multisensory temporal binding window persist into adolescence. *Developmental Science, 15*, 688–696.

Hockley N. S., & Polka, L. (1994). A developmental study of audiovisual speech perception using the McGurk paradigm, *Journal of the Acoustical Society of America, 96*, 3309.

Hof, P. R., & Mobbs, C. V. (2009). *Handbook of the neuroscience of aging*. Waltham, MA: Elsevier.

Huerta, I., Amato, A., Roca, X., & Gonzàlez, J. (2013). Exploiting multiple cues in motion segmentation based on background subtraction. *Neurocomputing, 100*, 183–190.

Huyse, A., Berthomier, F., & Leybaert, J. (2013). Degradation of labial information modifies audiovisual speech perception in cochlear-implanted children. *Ear and Hearing, 34*, 110–121.

Hyde, D. C., Jones, B. L., Flom, R., & Porter, C. L. (2011). Neural signatures of face–voice synchrony in 5-month-old human infants. *Developmental Psychobiology, 53*, 359–3570.

IJsseldijk, F. J. (1992). Speechreading performance under different conditions of video image, repetition, and speech rate. *Journal of Speech and Hearing Research, 35*, 466–471.

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research, 40*, 1489–1506.

Jiang, J. T., Alwan, A., Keating, P. A., Auer, E. T., & Bernstein, L. E. (2002). On the relationship between face movements, tongue movements, and speech acoustics. *EURASIP Journal on Advances in Signal Processing, 2002*, 1174–1188.

Jordan, T. R., & Bevan, K. M. (1997). Seeing and hearing rotated faces: Influences of facial orientation on visual and audiovisual speech recognition. *Journal of Experimental Psychology: Human Perception and Performance, 23*, 388–403.

Jordan, T. R., McCotter, M. V., & Thomas, S. M. (2000). Visual and audiovisual speech perception with color and gray scale facial images. *Perception & Psychophysics, 62*, 1394–1404.

Jordan, T. R., & Sergeant, P. C. (1998). Effects of facial image size on visual and audiovisual speech recognition. In R. Campbell, B. Dodd & D. Burnham (Eds.), *Hearing by eye II: Advances in the psychology of speechreading and auditory-visual speech* (pp. 155–176). London, UK: Psychology Press.

Jordan, T. R., & Sergeant, P. C. (2000). Effects of distance on visual and audiovisual speech recognition. *Language and Speech, 43,* 107–124.

Jordan, T. R., Sergeant, P. C., Martin, C., Thomas, S. M., & Thow, P. (1997). Effects of horizontal viewing angle on visual and audiovisual speech perception. In *Computational cybernetics and simulation: 1997 IEEE International Conference on Systems, Man, and Cybernetics* (Vol. 2, pp. 1626–1631). Piscataway, NJ: Institute of Electrical and Electronics Engineers.

Jordan, T. R., Sheen, M., Abedipour, L., & Paterson, K. B. (2014). Visual speech perception in foveal and extrafoveal vision: Further implications for divisions in hemispheric projections. *PLoS ONE, 9,* e98273.

Jordan, T. R., & Thomas, S. M. (2001). Effects of horizontal viewing angle on visual and audiovisual speech recognition. *Journal of Experimental Psychology: Human Perception and Performance, 6,* 1386–1403.

Jordan, T. R., & Thomas, S. M. (2007). Hemiface contributions to hemispheric dominance in visual speech perception. *Neuropsychology, 21,* 721–731.

Jordan, T. R., & Thomas, S. M. (2011). When half a face is as good as a whole: Effects of simple substantial occlusion on visual and audiovisual speech perception. *Attention, Perception, Psychophysiology, 73,* 2270–2285.

Jusczyk, P. W., & Aslin, R. N. (1995). Infants' detection of the sound patterns of words in fluent speech. *Cognitive Psychology, 29,* 1–23.

Kaganovich, N., & Schumaker, J. (2014). Audiovisual integration for speech during mid-childhood: Electrophysiological evidence. *Brain & Language, 139,* 36–48.

Kanwisher, N., McDermont, J., & Chun, M. M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. *The Journal of Neuroscience, 17,* 4302–4311.

Kanwisher, N., & Yovel, G. (2006). The fusiform face area: a cortical region specialized for the perception of faces. Philosophical transactions of the Royal Society. *Biological sciences, 361,* 2109–2128.

Kato, M., & Konishi, Y. (2013). Where and how infants look: The development of scan paths and fixations in face perception. *Infant Behavior & Development, 36,* 32–41.

Kawase, T., Yamaguchi, K., Ogawa, T., Suzuki, K.-I., Suzuki, M., Itoh, M., … Fujii, T. (2005). Recruitment of fusiform face area associated with listening to degraded speech sounds in auditory–visual speech perception: A PET study. *Neuroscience Letters, 382,* 254–258.

Kayser, C., Petkov, C. I., & Logothetis, N. K. (2008). Visual modulation of neurons in auditory cortex. *Cerebral Cortex, 18*, 1560–1574.

Kim, J., & Davis, C. (2003). Testing the cueing hypothesis for the AV speech detection. In: *Proceedings of the Auditory-Visual Speech Processing (AVSP'03)*, 5-8 September (pp. 9–12). Saint Jorioz, France.

Kim, J., & Davis, C. (2004). Investigating the audio–visual speech detection advantage. *Speech Communication, 44*, 19–30.

Kim, J., & Davis, C. (2011). Testing audio-visual familiarity effects on speech perception in noise. In: *Proceedings of the 17th International Congress of Phonetic Sciences (ICPhS XVII),* 17-21 August (pp. 17–21). Hong Kong, China.

Kimchi, R., Hadad, B., Behrmann, M., & Palmer, S. E. (2005). Microgenesis and ontogenesis of perceptual organization: Evidence from global and local processing of hierarchical patterns. *Psychological Science, 16*, 282–290.

King, A. J. (2005). Multisensory integration: Strategies for synchronization. *Current Biology, 15*, R339–R341.

Kislyuk, D. S., Möttönen, R., & Sams, M. (2008). Visual processing affects the neural basis of auditory discrimination. *Journal of Cognitive Neuroscience, 20*, 2175–2184.

Knowland, V. C. P., Mercure, E., Karmiloff-Smith, A., Dick, F., & Thomas, M. S. C. (2014). Audio-visual speech perception: A developmental ERP investigation. *Developmental Science, 17*, 110–124.

Kryter, K. D. (1996). *Handbook of hearing and the effects of noise*. New York: Academic Press.

Kubicek, C., de Boisferon A. H., Dupierrix, E., Pascalis, O., Loevenbruck, H., Gerain, J., & Schwarzer, G. (2014). Cross-odal matching of audio-visual German and French fluent speech in infancy. *PLoS ONE, 9*, e89275.

Kuhl, P. K., & Meltzoff, A. N. (1982). The bimodal perception of speech in infancy. *Science, 218*, 1138–1141.

Kuhl, P. K., & Meltzoff, A. N. (1984). The intermodal representation of speech in infants. *Infant Behavior & Development, 7*, 361–381.

Kuhl, P. K., & Meltzoff, A. N. (1988). Speech as an intermodal object of perception. In A. Yonas (Ed.), *Perceptual development in infancy: The Minnesota Symposia on Child Psychology* (pp. 235–266). Hillsdale, NJ: Erlbaum.

Kuhl, P. K., Stevens, E., Hayashi, A., Deguchi, T., Kiritani, S., & Iverson, P. (2006). Infants show a facilitation effect for native language phonetic perception between 6 and 12 months. *Developmental Science, 9*, F13–21.

Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., & Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science, 255,* 606–608.

Kuhl, P. K., Williams, K. A., & Meltzoff, A. N. (1991). Cross-modal speech perception in adults and infants using nonspeech auditory stimuli. *Journal of Experimental Psychology: Human Perception and Performance, 17,* 829–840.

Kushnerenko, E., Teinonen, T., Volein, A., & Csibra, G. (2008). Electrophysiological evidence of illusory audiovisual speech percept in human infants. *Proceedings of the National Academy of Science, 105*, 11442–11445.

Lakatos, P., Karmos, G., Mehta, A. D., Ulbert, I., & Schroeder, C. E. (2008). Entrainment of neuronal oscillations as a mechanism of attentional selection. *Science, 320,* 110–113.

Lansing, C. R., & McConkie, G. W. (2003). Word identification and eye fixation locations in visual and visual-plus-auditory presentations of spoken sentences. *Perception & Psychophysics, 65*, 536–552.

Laurienti, P., Burdette, J., Maldjian, J., & Wallace, M. (2006). Enhanced multisensory integration in older adults. *Neurobiology of Aging, 27*, 1155–1163.

Leder, H., & Carbon, C. C. (2006). Face-specific configural processing of relational information. *British Journal of Psychology, 97*, 19–29.

Lewkowicz, D. J. (1996). Perception of auditory–visual temporal synchrony in human infants. *Journal of Experimental Psychology : Human Perception and Performance, 22*, 1094–1106.

Lewkowicz, D. J. (2000). The development of intersensory temporal perception: An epigenetic systems/limitations view. *Psychological Bulletin, 126*, 281–308.

Lewkowicz, D. J. (2002). Heterogeneity and heterochrony in the development of intersensory perception. *Cognitive Brain Research, 14*, 41–63.

Lewkowicz, D. J. (2010). Infant perception of audio-visual speech synchrony. *Developmental Psychology, 46*, 66–77.

Lewkowicz, D. J., & Ghazanfar, A. A. (2009). The emergence of multisensory systems through perceptual narrowing. *Trends in Cognive Sciences, 13*, 470–78.

Lewkowicz, D. J., Minar, N. J., Tift, A. H., & Brandon, M. (2015). Perception of the multisensory coherence of fluent audiovisual speech in infancy: Its emergence and the role of experience. *Journal of Experimental Child Psychology, 130*, 147–162.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review, 74*, 431–461.

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition, 21,* 1–36.

Liversedge, S., Gilchrist, I., & Everling, S. (2013). *The Oxford handbook of eye movements* (1st edition). Oxford, UK: Oxford University Press.

Liu, X.Z., &Yan, D. (2007). Ageing and hearing loss. *The Journal of Pathology, 211*, 188–197.

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing, 19*, 1–36.

Luo, H., Liu, Z., & Poeppel, D. (2010). Auditory cortex tracks both auditory and visual stimulus dynamics using low-frequency neuronal phase modulation. *PLoS Biology, 8*, e1000445.

Ma, W. J., Zhou, X., Ross, L. A., Foxe, J. J., & Parra, L. C. (2009). Lip-reading aids word recognition most in moderate noise: A Bayesian explanation using high-dimensional feature space. *PLoS One, 4*, e4638.

MacKain, K., Studdert-Kennedy, M., Spieker, S., & Stern, D. (1983). Infant intermodal speech perception is a left-hemisphere function. *Science, 219,* 1347–1349.

Marassa, L. K., & Lansing, C. R. (1995). Visual word recognition in 2 facial motion conditions: Full face versus lips-plus-mandible. *Journal of Speech and Hearing Research, 38,* 1387–1394.

Marslen-Wilson, W., & Tyler, L. (1980). The temporal structure of spoken language understanding. *Cognition, 8*, 1–71.

Massaro, D. W. (1984). Children's perception of visual and auditory speech. *Child development, 55*, 1777–1788.

Massaro, D. W. (1987). *Speech perception by ear and eye: A paradigm for psychological inquiry*. Hillsdale, NJ: Erlbaum Associates.

Massaro, D. W. (1998). *Perceiving talking faces: From speech perception to a behavioral principle*. Cambridge, MA: The MIT Press.

Massaro, D. W. (2004). From multisensory integration to talking heads and language learning. In G. A. Calvert, C. Spence & B. E. Stein (Eds.), *The handbook of multisensory processes* (pp. 153–176). Cambridge, MA: The MIT Press.

Massaro, D. W., & Cohen, M. M. (1996). Perceiving speech from inverted faces. *Perception & Psychophysics, 58,* 1047–1065.

Massaro, D. W., Thompson, L. A., Barron, B., & Lauren, E. (1986). Developmental changes in visual and auditory contributions to speech perception. *Journal of Experimental Child Psychology, 41*, 93–113.

Mattys, S. L., Bernstein, L. E., & Auer, E. T., Jr. (2002). Stimulus-based lexical distinctiveness as a general word-recognition mechanism. *Perception and Psychophysics, 64,* 667–679.

Matchin, W., Groulx, K., & Hickok, G. (2014). Audiovisual speech integration does not rely on the motor system: Evidence from articulatory suppression, the McGurk effect and fMRI. *Journal of Cognitive Neuroscience, 26,* 606–620.

Maye, J., Werker, J. F., & Gerken, L. A. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition, 82,* B101–B111.

McCormick, B. (1979). Audio-visual discrimination of speech. *Clinical Otolaryngology & Allied Sciences, 45,* 355–361.

McCotter, M. V., & Jordan, T. R. (2003). The role of facial colour and luminance in visual and audio-visual speech perception. *Perception, 32*, 921–936.

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

McIntosh, A. R., & Gonzalez-Lima, F. (1994). Structural equation modeling and its application to network analysis in functional brain imaging. *Human Brain Mapping, 2*, 2–22.

Meredith, M. A. (2002). On the neuronal basis for multisensory convergence: A brief overview. *Cognitive Brain Research, 14,* 31–40.

Meredith, M. A., Nemitz, J. W., & Stein, B. E. (1987). Determinants of multisensory integration in superior colliculus neurons. I. Temporal factors. *Journal of Neuroscience, 7,* 3215–3229.

Meredith, M. A., & Stein, B. E. (1986). Spatial factors determine the activity of multisensory neurons in cat superior colliculus. *Brain Research,* 365, 350–354.

Miller, J. (1986). Time course of coactivation in bimodal divided attention. *Perception & Psychophysics, 40,* 331–343.

Miller, L. M., & D'Esposito, M. (2005). Perceptual fusion and stimulus coincidence in the cross-modal integration of speech. *The Journal of Neuroscience, 25,* 5884–5893.

Mills, J. H., Schmiedt, R. A., & Dubno, J. R. (2006). Older and wiser, but losing hearing nonetheless. *Hearing Health, Summer*, 12–19.

Mondloch, C. J., Le Grand, R., & Mauer, D. (2002). Configural face processing develops more slowly than featural face processing. *Perception, 31*, 553–566.

Moulin-Frier, C., & Arbib, M. (2013). Recognizing speech in a novel accent: The motor theory of speech perception reframed. *Biological Cybernetics, 107,* 421–447.

Muir, D., Clifton, R., & Clarkson, M. G. (1989). The development of a human auditory localization response: A U-shaped function. *Canadian Journal of Psychology, 43*, 199–216.

Munhall, K. G., Gribble, P., Sacco, L., & Ward, M. (1996). Temporal constraints on the McGurk effect. *Perception & Psychophysics, 58*, 351–362.

Munhall, K. G., Kroos, C., Jozan, G., & Vatikiotis-Bateson, E. (2004). Spatial frequency requirements for audiovisual speech perception. *Perception & Psychophysics, 66*, 574–583.

Munhall, K. G., & Vatikiotis-Bateson, E. (1998). The moving face during speech communication. In R. Campbell, B. Dodd & D. Burnham (Eds.), *Hearing by eye II: Advances in the psychology of speechreading and auditory-visual speech* (pp. 123–139). London: Psychology Press.

Munhall, K., & Vatikiotis-Bateson, E. (2004). Spatial and temporal constraints on audiovisual speech perception. In G. A. Calvert, C. Spence & B. E. Stein (Eds.), *The handbook of multisensory processing* (pp. 177–188). Cambridge, MA: The MIT Press.

Näätänen, R., Paavilainen, P., Titinen, H., Jiang, D., & Alho, D. (1993). Attention and mismatch negativity. *Psychophysiology, 30*, 436–450.

Narinesingh, C., Goltz, H. C., Raashid, R. A., & Wong, A. M. F. (2015). Developmental trajectory of McGurk effect susceptibility 7 in children and adults with amblyopia. *Investigative Ophthalmology & Visual Science, 56*, 2107–2113.

Nath, A. R., & Beauchamp, M. S. (2011). Dynamic changes in superior temporal sulcus connectivity during perception of noisy audiovisual speech. *Journal of Neuroscience, 31*, 1704–1714.

Nath, A. R., Fava, E. E., & Beauchamp, M. S. (2011). Neural correlates of interindividual differences in children's audiovisual speech perception. *The Journal of Neuroscience, 31*, 13963–13971.

Navarra, J., & Soto-Faraco, S. (2007). Hearing lips in a second language: Visual articulatory information enables the perception of second language sounds. *Psychological Research, 71*, 4–12.

Narayan, C., Werker, J. F., & Beddor, P. (2010). The interaction between acoustic salience and language experience in developmental speech perception: evidence from nasal place discrimination. *Developmental Science, 13*, 407–20.

Nazzi, T., & Ramus, F. (2003). Perception and acquisition of linguistic rhythm by infants. *Speech Communication, 41*, 233–244.

Nelson, R., & Palmer, S. E. (2001). Of holes and wholes: The perception of surrounded regions. *Perception, 30,* 1213–1226.

O'Neill, J. J. (1954). Contributions of the visual component of oral symbols to speech comprehension. *Journal of Speech and Hearing Disorders, 19*, 429.

Paré, M., 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 & Psychophysics, 65*, 533–567.

Patterson, M. L., & Werker, J. F. (2003). Two-month-old infants match phonetic information in lips and voice. *Developmental Science, 6*, 191–196.

Peelle, J. E., & Sommers, M. S. (in press). Prediction and constraint in audiovisual speech perception. *Cortex*.

Poeppel, D., Idsardi, W. J., & van Wassenhove, V. (2008). Speech perception at the interface of neurobiology and linguistics. *Philosophical Transaction of the Royal Society of London: Series B, Biological Science, 363,* 1071–1086.

Poirel, N., Krakowski, C. S., Sayah, S., Pineau, A., Houdé, O. & Borst, G. (2014). Inhibitory control during local processing: A negative priming study of local-global processing. *Experimental Psychology, 61*, 205–214.

Perrodin, C., Kayser, C., Logothetis, N. K., & Petkov, C. I. (2015). Natural asynchronies in audiovisual communication signals regulate neuronal multisensory interactions in voice-sensitive cortex. *Proceedings of the National Academy of Science, 112*, 273–278.

Piepers, D. W., & Robins, R. A. (2012). A Review and clarification of the terms "holistic," "configural," and "relational" in the face perception literature. *Frontiers in Psychology, 3*, 559.

Pons, F., Lewkowicz, D. J., Soto-Faraco, S., & Sebastián-Gallés, N. (2009). Narrowing of intersensory speech perception in infancy. *Proceedings of the National Academy of Sciences of the United States of America, 106*, 10598–10602.

Preminger, J. E., Lin, H. B., Payen, M., & Levitt, H. (1998). Selective visual masking in speechreading. *Journal of Speech, Language, and Hearing Research, 41,* 564–575.

Prkachin, G. C. (2003). The effects of orientation on detection and identification of facial expressions of emotion. *British Journal of Psychology, 94,* 45–62.

Raichle, M. E. (2010). Two views of brain function. *Trends in Cognitive Science, 14,* 180–190.

Rakover, S. S., & Teucher, B. (1997). Facial inversion effects: Parts and whole relationship. *Perception & Psychophysics, 59,* 752–761.

Ramsey, P. H. (1980). Exact type 1 error rates for robustness of Student's t-test with unequal variances. *Journal of Educational and Behavioral Statistics, 5,* 337–349.

Rauschecker, J. (2011). An expanded role for the dorsal auditory pathway in sensorimotor control and integration. *Hearing Research, 271,* 16–25.

Rayner, K. (1975). The perceptual span and peripheral cues in reading. *Cognitive Psychology, 7,* 65–81.

Reingold, E. M., Charness, N., Pomplun, M., & Stampe, D. M., (2001). Visual span in expert chess players: Evidence from eye movements. *Psychological Science, 12,* 48–55.

Reingold, E. M., & Loschky, L. C. (2002). Saliency of peripheral targets in gaze-contingent multiresolutional displays. *Behavior Research Methods, Instruments & Computers, 34,* 491–499.

Reisberg, D., McLean, J., & Goldfield, A. (1987). Easy to hear but hard to understand: a lip-reading advantage with intact auditory stimuli. In B. Dodd & R. Campbell (Eds.), *Hearing by eye: the psychology of lip-reading* (pp. 97–113). Hillsdale, NJ: Lawrence Erlbaum Associates.

Reynolds, G. D., Bahrick, L. E., Lickliter, R., & Guy, M. W. (2014). Neural correlates of intersensory processing in 5-month-old Infants. *Developmental Psychobiology, 56,* 355–372.

Rizzolatti, G., Fadiga, L., Gallese, V., & Fogassi, L. (1996). Premotor cortex and the recognition of motor actions. *Brain Research: Cognitive Brain Research, 3,* 131–141.

Robbins, R., & McKone, E. (2007). No face-like processing for objects-of- expertise in three behavioural tasks. *Cognition, 103,* 34–79.

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 vowels in noise. *Journal of Acoustical Society of America, 103,* 3677–3689.

Rockland, K. S., & Ojima, H. (2003). Multisensory convergence in calcarine visual areas in macaque monkey. *International Journal of Psychophysiology, 50,* 19–26.

Rockland, K. S., & Van Hoesen, G. W. (1994). Direct temporal-occipital feedback connections to striate cortex (V1) in the macaque monkey. *Cerebral Cortex, 4*, 300–313.

Rosenblum, L. D. (2005). Primacy of multimodal speech perception. In: D. B. Pisoni & R. E. Remez (Eds.), *Handbook of Speech Perception* (pp. 51–78). Malden, MA: Blackwell Publishing

Rosenblum, L. D. (2008). Speech perception as a multimodal phenomenon. *Current Directions in Psychological Science, 12*, 405–409.

Rosenblum, L. D., Johnson, J. A., & Saldaña, H. M. (1996). Visual kinematic information for embellishing speech in noise. *Journal of Speech and Hearing Research, 39*, 1159–1170.

Rosenblum, L. D., Schmuckler, M. A., & Johnson, J. A. (1997). The McGurk effect in infants. *Perception & Psychophysics, 59*, 347–357.

Rosenblum, L. D., Yakel, D. A., & Green, K. P. (2000). Face and mouth inversion effects on visual and audiovisual speech perception. *Journal of Experimental Psychology: Human Perception and Performance, 26*, 806–819.

Ross, L. A., Saint-Amour, D., Leavitt, V. M., Javitt, D. C., & Foxe, J. (2007). Do you see what I am saying? Exploring visual enhancement of speech in noisy environments. *Cerebral Cortex, 17*, 1147–1153.

Ross, L. A., Molholm, S., Blanco, D., Gomez-Ramirez, M., Saint-Amour, D., & Foxe, J. J. (2011). The development of multisensory speech perception continues into the late childhood years. *European Journal of Neuroscience, 33*, 2329–2337.

Rouger, J., Fraysse, D., Deguine, O., & Barone, P. (2008). McGurk effects in cochlear-implanted deaf subjects. *Brain Research, 10*, 87–99.

Saffran, J. R., Werker, J. & Werner, L. (2006). The infant's auditory world: Hearing, speech, and the beginnings of language. In Siegler, R. & Kuhn, D. (Eds), *Handbook of child development* (pp. 58–108). New York, NY: Wiley.

Saleem, K. S., Suzuki, W., Tanaka, K., & Hashikawa, T. (2000). Connections between anterior inferotemporal cortex and superior temporal sulcus regions in the macaque monkey. *The Journal of Neuroscience, 20*, 5083–5101.

Sato, W., Kochiyama, T., & Yoshikawa, S. (2011). The inversion effect for neutral and emotional facial expressions on amygdala activity. *Brain Research, 10*, 84–90.

Scheinberg, J. C. (1980). Analysis of speechreading cues using an interleaved technique. *Journal of Communication Disorders, 13*, 489–492.

Schmider, E., Ziegler, M., Danay, E., Meyer, L., & Bühner, M. (2010). Is it really robust? Reinvestigating the robustness of ANOVA against the violations of normal distribution

assumption. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 6*, 147–151.

Schroeder, C. E., & Foxe, J. (2005). Multisensory contributions to low-level, "unisensory" processing. *Current Opinion in Neurobiology, 15*, 454–458.

Schroeder, C. E., & Lakatos, P. (2009). Low-frequency neuronal oscillations as instruments of sensory selection. *Trends in Neuroscience, 32*, 9–18.

Schroeder, C. E., Lakatos, P., Kajikawa, Y., Partan, S., & Puce, A. (2008). Neuronal oscillations and visual amplification of speech. *Trends in Cognitive Science, 12*, 106–113.

Schroeder, M. M., Lipton, R. B., Ritter, W., Giesser, B. S., & Vaughan, H. G. Jr. (1995). Event-related potential correlates of early processing in normal aging. *International Journal of Neuroscience, 80*, 371–382.

Schroeder, C. E., Wilson, D. A., Radman, T., Scharfman, H., & Lakatos, P. (2010). Dynamics of active sensing and perceptual selection. *Current Opinion in Neurobiology, 20*, 172–176.

Schneider, G. E. (2014). *Brain structure and its origins: In development and in evolution of behavior and the mind*. Cambridge, MA: MIT Press.

Schwartz, J.-L. (2004). La parole multisensorielle: Plaidoyer, problèmes, perspective. In: *Actes des XXVes Journées d'Etude sur la Parole JEP 2004*, 19-22 April (pp. 11–18). Paris, France.

Schwartz, J., Basirat, A., Menard, L., & Sato, M. (2012). The perception-for-action-control theory (PACT): A perceptuo-motor theory of speech perception. *Journal of Neurolinguist*ics, *25*, 336–354.

Schwartz, J.-L., Berthommier, F., & Savariaux, C. (2004). Seeing to hear better: Evidence for early audio-visual interactions in speech identification. *Cognition, 93*, 69–78.

Schwartz, J.-L., & Savariaux, C. (2014). No, there is no 150 ms lead of visual speech on auditory speech, but a range of audiovisual asynchronies, varying from small audio lead to large audio lag. *Computational Biology, 10*, e1003743.

Sekiyama, K. (1997). Cultural and linguistic factors in audiovisual speech processing: The McGurk effect in Chinese subjects. *Perception & Psychophysics, 59*, 73–80.

Sekiyama, K., & Burnham, D. (2004). Issues in the development of auditory-visual speech perception: Adults, infants, and children. In *Interspeech 2004 - ICSLP, 8th International Conference on Spoken Language Processing*, 4-8 October (pp. 821–825). Jeju Island, Korea.

Sekiyama, K., & Burnham, D. (2008). Impact of language on development of auditory-visual speech perception. *Developmental Science, 11*, 306–20.

Sekiyama, K., Kanno, I., Miura, S., & Sugita, Y. (2003). Auditoryvisual speech perception examined by fMRI and PET. *Neuroscience Research, 47*, 277–287.

Sekiyama, K., Soshi, T., & Sakamoto, S. (2014). Enhanced audiovisual integration with aging in speech perception: A heightened McGurk effect in older adults. *Frontiers in Psychology, 5*, 323.

Sekiyama, K. & Tohkura, Y. (1991). McGurk effect in non-English listeners: Few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility. *Journal of Acoustical Society of America, 90,* 1797–1805.

Sekiyama, K., & Tohkura, Y. (1993). Inter-language differences in the influence of visual cues in speech perception. *Journal of Phonetics, 21*, 427–444.

Selzer, B., & Pandya, D. N. (1994). Parietal, temporal, and occipital projections to cortex of the superior temporal sulcus in the rhesus monkey: a retrograde tracer study. *Journal of Computational Neuroscience, 343*, 445–463.

Shaw, P., Kabani, N. J., Lerch, J. P., Eckstrand, K., Lenroot, R., Gogtay, N., … Wise, S. P. (2008). Neurodevelopmental trajectories of the human cerebral cortex. *The Journal of Neuroscience, 28*, 3586–3594.

Sherf, K. S., Behrmann, M., Kimchi, R., & Luna, B. (2009). Emergence of global shape processing continues through adolescence. *Child Development, 80*, 162–177.

Skipper, J. I., Nusbaum, H., & Small, S. L. (2005). Listening to talking faces: motor cortical activation during speech perception. *NeuroImage, 25*, 76–89.

Skipper, J. I., Nusbaum, H. C., & Small, S. L. (2006). Lending a helping hand to hearing: Another motor theory of speech perception. In M. A. Arbib (Ed.), *Action to language via the mirror neuron system* (pp. 250–285). Cambridge, UK: Cambridge University Press.

Skipper, J. I., van Wassenhove, V., Nusbaum, H. C., & Small, S. L. (2007). Hearing lips and seeing voices: How cortical areas supporting speech production mediate audiovisual speech perception. *Cerebral Cortex, 17*, 2387–2399.

Sommers, M. S., Tye-Murray, N., & Spehar, B. (2005). Auditory-visual speech perception and auditory-visual enhancement in normal-hearing younger and older adults. *Ear and Hearing, 26*, 263–275.

Sowell, E. R., Thompson, P. M., Leonard, C. M., Welcome, S. E., Kan, E. & Toga, A. W. (2004). Longitudinal mapping of cortical thickness and brain growth in normal children. *The Journal of Neuroscience, 22*, 8223–8231.

Soto-Faraco, S., & Alsius, A. (2009). Deconstructing the McGurk-MacDonald illusion. *Journal of Experimental Psychology: Human Perception and Performance, 35*, 580–587.

Soto-Faraco, S., Calabresi, M., Navarra, J., Werker, J. F., & Lewkowicz, D. J. (2012). Development of audiovisual speech perception. In A. Bremner, D. Lewkowicz & C. Spence (Eds). *Multisensory Development* (pp. 435–452). Oxford, UK: Oxford University Press.

Spear, P. D. (1993).Neural bases of visual deficits during aging. *Vision Research, 33*, 2589–2609.

Spence, C., & Squire, S. (2003). Multisensory integration: Maintaining the perception of synchrony. *Current Biology, 13*, R519–R521.

Sporns, O. (2011). *Networks of the brain*. Cambridge, MA: The MIT Press.

Stein, B. E., & Meredith M. A. (1993). *The merging of the senses*. Cambridge, MA: The MIT Press.

Stein, B. E., Stanford, T. R., & Rowland, B. A. (2009). The neural basis of multisensory integration in the midbrain: Its organization and maturation. *Hearing Research, 258*, 4–15.

Stekelenburg, J. J., & Vroomen, J. (2007). Neural correlates of multisensory integration of ecologically valid audiovisual events. *Journal of Cognitive Neuroscience, 19*, 1964–1973.

Stevenson, R. A., & James, T. W. (2009). Audiovisual integration in human superior temporal sulcus: Inverse effectiveness and the neural processing of speech and object recognition. *NeuroImage, 44*, 1210–1223.

Stevenson, R. A., Neums, C. E., Baum, S.H., Zurkovsky, L., Barense, M. D., Newhouse, P. A., & Wallace, M. T. (2015). Deficits in audiovisual speech perception in normal aging emerge at the level of whole-word recognition. *Neurobiology of Aging, 36*, 283–291.

Stevenson, R. A., VanDerKolk, R. M., Pisoni, D. B., & James, T. W. (2010). Discrete neural substrates underlie complementary audiovisual speech integration processes. *NeuroImage, 55*, 1339–1345.

Stone, L. (1957). *Facial cues of context in lip reading.* Los Angeles: John Tracy Clinic.

Stone, J. L., & Hughes, J. R. (2013). Early history of electroencephalography and establishment of the American Clinical Neurophysiology Society. *Journal of Clinical Neurophysiology, 30*, 28–44.

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America, 26*, 212–215.

Summerfield, A. Q. (1979). Use of visual information for phonetic perception. *Phonetica, 36*, 314–331.

Summerfield, A. Q. (1987). Some preliminaries to a comprehensive account of audio-visual speech perception. In B. Dodd & R. Campbell (Eds.), *Hearing by eye* (pp. 3–51). London: Erlbaum Associates.

Summerfield, Q., & McGrath, M. (1984). Detection and resolution of audio-visual incompatibility in the perception of vowels. *Quarterly Journal of Experimental Psychology, 36*, 51–74.

Summerfield, A. Q., MacLeod, A., McGrath, M., & Brooke, M. (1989). Lips, teeth and the benefits of lipreading. In A. W. Young & H. D. Ellis (Eds.), *Handbook of research on face processing* (pp. 223–233). Amsterdam: North-Holland.

Tanaka, J. W., & Farah, M. J. (1993). Parts and wholes in face recognition. *Quarterly Journal of Experimental Psychology, 46A*, 225–245.

Tanaka, J. W., & Farah, M. J. (2003). The holistic representation of faces. In G. Rhodes & M. A. Peterson (Eds.), Analytic and Holistic Processes in Perception of Faces, Objects and Scenes (pp. 53–74). New York: Oxford University Press.

Tanaka, J. W., Farah, M. J., & Drain, H. M. (1995). What causes the face inversion effect ? *Journal of Experimental Psychology: Human Perception and Performance, 21*, 628–634.

Tanaka, J. W., & Sung, A. (2013). The "eye avoidance" hypothesis of autism face processing. *Journal of Autism and Developmental Disorders, 21*, 1–15.

Taubert, J., Apthorp, D., Aagten-Murphy, D., & Alais, D. (2011). The role of holistic processing in face perception: Evidence from the face inversion effect. *Vision Research, 51*, 1273–1278.

Teinonen, T., Aslin, R. N., Alku, P., & Csibra, G. (2008). Visual speech contributes to phonetic learning in 6-month-old infants. *Cognition, 105*, 850–855.

Thomas, S. M., & Jordan, T. R. (2002). Determining the influence of Gaussian blurring on inversion effects with talking faces. *Perception & Psychophysics, 64*, 932–944.

Thomas, S. M., & Jordan, T. R. (2004). Contributions of oral and extraoral facial movement to visual and audiovisual speech perception. *Journal of Experimental Psychology: Human Perception and Performance, 30*, 873–888.

Thompson, P. (1980). Margaret Thatcher: A new illusion. *Perception, 9*, 483–484.

Thut, G., Miniussi, C., & Gross, J. (2012). The functional importance of rhythmic activity in the brain. *Current Biology, 22*, R658–R663.

Tiipana, K. (2014). What is the McGurk effect? *Frontiers in Psychology, 5*, 725.

Tiippana, K., Tiainen, M., Vainio, L., & Vainio, M. (2013). Acoustic and visual phonetic features in the McGurk effect. In: *Proceedings of Interspeech 2013, August 25-29* (pp. 1634–1638). Lyon, France.

Tremblay, C., Champoux, F., Voss, P., Bacon, B. A., Lepore, F., & Théoret, H. (2007). Speech and non-speech audio-visual illusions: A developmental study. *PLoS ONE, 8*, e742.

Tremblay, K., & Ross, B. (2007). Effects of age and age-related hearing loss on the brain. *Journal of Communication Disorders, 40*, 305–312.

Tye-Murray, N., Sommers, M. S., & Spehar, B. (2007a). Audiovisual integration and lipreading abilities of older adults with normal and impaired hearing. *Ear and Hearing, 28*, 656–668.

Tye-Murray, N., Sommers, M. S., & Spehar, B. (2007b). Auditory and visual lexical neighborhoods in audiovisual speech perception. *Trends in Amplification, 11*, 233–241.

Tye-Murray, N., Sommers, M. S., Spehar, B., Myerson, J., & Hale, S. (2010). Aging, audiovisual integration, and the principle of inverse effectiveness. *Ear and Hearing, 31*, 636–644.

Tye-Murray, N., Spehar, B., Myerson, J., Sommers, M. S., & Hale, S. (2011). Crossmodal enhancement of speech detection in young and older adults: Does signal content matter? *Ear and Hearing, 32*, 650–655.

Valentine, T. (1988). Upside-down faces: A review of the effect of inversion upon face recognition. *British Journal of Psychology, 79*, 471–491.

Van Belle, G., De Graef, P., Verfaillie, K., Robinson, B., & Lefèvre, P. (2010). Face inversion impairs holistic perception: Evidence from gaze-contingent stimulation. *Journal of Vision, 10*, 10, 1–13.

Vander Wyk, B. C., Ramsay, G. J., Hudac, C. M., Jones, W., Lin, D., Klin, A. … Pephly, K. A. (2010). Cortical integration of audio–visual speech and non-speech stimuli. *Brain and Cognition, 74*, 97–106.

Vanrullen, R., Busch, N. A., Drewes, J., & Dubois, J. (2011). Ongoing EEG phase as a trial-by-trial predictor of perceptual and attentional variability. *Frontiers in Psychology, 2*, 60.

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 Sciences of the United States of America, 102*, 1181–1186.

van Wassenhove, V., Grant, K.W., & Poeppel, D. (2007). Temporal window of integration in auditory–visual speech perception. *Neuropsychologia, 45*, 598–607.

Vatakis, A., & Spence, C. (2008). Investigating the effects of inversion on configural processing with an audiovisual temporal-order judgment task. *Perception, 37*, 143–160.

Vatikiotis-Bateson, E., Eigsti, I. M., Yano, S., Munhall, K. G. (1998). Eye movement of perceivers during audiovisual speech perception. *Perception & Psychophysics, 60*, 926–940.

Vatikiotis-Bateson, E., Munhall, K. G., Hirayama, M., Lee, Y. V., & Terzopoulos, D. (1996). Dynamics of facial motion in speech: Kinematic and electromyographic studies of orofacial structures. In D. G. Stork & M. Hennecke (Eds.), *Speechreading by humans and machines: Models, systems, and applications* (pp. 221–232). Berlin, Germany: Springer-Verlag.

Vellutino, F. R., Fletcher, J. M., Snowling, M. J., & Scanlon, D. M. (2004). Specific reading disability (dyslexia): what have we learned in the past four decades? *Journal of Child Psychology & Psychiatry, 45*, 2–40.

Vroomen, J., & Stekelenburg, J (2011). Perception of intersensory synchrony in audiovisual speech processing: Not that special. *Cognition, 118*, 75–83.

Walton, G. E., & Bower, T. G. (1993). Amodal representations of speech in infants. *Infant Behavior & Development, 16*, 233–243.

Wayne, R. V., & Johnsrude, I. S. (2012). The role of visual speech information in supporting perceptual learning of degraded speech. *Journal of Experimental Psychology: Applied, 18*, 419–435.

Weikum, W. M., Vouloumanos, A., Navarra, J., Soto-Faraco, S., Sebastián-Gallés, N., & Werker, J. F. (2007). Visual language discrimination in infancy. *Science, 316*, 1159.

Werker J. F., & Tees, R. C. (1984). Cross-language speech perception: evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development, 7*, 49–63.

Wightman, F., Kistler, D., & Brungart, D. (2006). Informational masking of speech in children: Auditory–visual integration. *Journal of the Acoustical Society of America, 119*, 3940–3949.

Wilson, S. M., & Iacoboni, M. (2006). Neural responses to non-native phonemes varying in producibility: Evidence for the sensorimotor nature of speech perception. *NeuroImage, 33*, 316–325.

Winneke, A. H., & Phillips, N. A. (2011). Does audiovisual speech offer a fountain of youth for old ears? An event-related brain potential study of age differences in audiovisual speech perception. *Psychology and Aging, 26*, 427–438.

Woynaroski, T. G., Kwakye, L. D., Foss-Feig, J. H., Stevenson, R. A., Stone, W. L., & Wallace, M. T. (2014). Multisensory speech perception in children with autism spectrum disorder. *Journal of Autism and Developmental Disorders, 43,* 2891–2901.

Yakovlev, P., & Lecours, A. R. (1967). The myelogenetic cycles of regional maturation of the brain. In A. Minkovski (Ed.), *Regional development of the brain in early life* (pp. 3–69). Oxford: Blackwell.

Yao, Y. (2011). *Effects of neighborhood density on pronunciation variation* (Thèse non-publiée). University of California, Barley.

Yehia, H. C., Kuratate, T., & Vatikiotis-Bateson, E. (2002). Linking facial animation, head motion and speech acoustics. *Journal of the International Phonetic Association, 30*, 555–568.

Yeung, H. H., & Werker, J. F. (2009). Learning words' sounds before learning how words sound: 9-month-olds use distinct objects as cues to categorize speech information. *Cognition, 113*, 234–243.

Yi, H. G., Phelps, J. E., Smiljanic, R., & Chandrasekaran, B. (2013). Reduced efficiency of audiovisual integration for nonnative speech. *The Journal of Acoustical Society of America, 134*, 387–393.

Yi, H.-G., Smiljanic, R., & Chandrasekaran, B. (2014). The neural processing of foreign-accented speech and its relationship to listener bias. *Frontiers in Human Neuroscience, 8*, 768.

Yin, R. K. (1969). Looking at upside-down faces. *Journal of Experimental Psychology, 81*, 141–145.

Zmigrod, S., & Hommel, B. (2013). Feature integration across multimodal perception and action. *Multisensory Research, 26,* 143–157.

Table des matières

Documents annexes

Annexe 1. Formulaire de consentement pour les participants majeurs

Annexe 2. Formulaire de consentement parental pour les participants mineurs

Annexe 3. Formulaire d'autorisation de droit à l'image

Annexe 4. Représentation schématique de l'organisation des conditions expérimentales en fonction du mode de présentation audio(-visuelle) et en fonction du degré de dégradation de l'information auditive

Annexe 5. Exemple de l'ordre de la présentation des stimuli expérimentaux

Annexe 6. Exemple de la feuille de codage des réponses des participants

Annexe 7. Tableaux comportant le stimulus acoustiquement proche et les stimuli acoustiquement éloignés pour chaque item expérimental ayant servi à la mise en place des feuilles de codage des réponses des participants

Annexe 1

Formulaire de consentement libre et éclairé

Cette recherche est réalisée dans le cadre de la thèse doctorale de Grozdana Erjavec (grozdana.erjavec@etud.univ-paris8.fr) sous la direction de Denis Legros, professeur émérite. Grozdana Erjavec est affiliée à l'école doctorale Cognition, Langage, Interaction (Université de Paris VIII) et au groupe Lutin (http://www.lutin-userlab.fr/accueil/) du laboratoire CHART.

Avant d'accepter de participer à ce projet de recherche, veuillez prendre le temps de lire et de comprendre les renseignements qui suivent. Ce document vous explique le but de ce projet de recherche, ses procédures, avantages, risques et inconvénients. Nous vous invitons à poser toutes les questions que vous jugerez utiles à la personne qui vous présente ce document.

La recherche a pour but d'étudier l'apport de l'information visuelle, relative aux mouvements articulatoires, à la perception de la parole.

Aucune technique invasive n'est utilisée dans la recherche. La recherche ne comporte pas de risques ou d'inconvénients notables.

Votre participation à cette recherche consiste à :

- Suivre la présentation du matériel (5 min) ;
- Visualiser une série de vidéos comportant des syllabes prononcées par une personne française de sexe féminin et dégradées par du bruit dans le but de les répéter à haute voix (10 min).

Vous êtes libre de participer à ce projet de recherche. Vous pouvez aussi mettre fin à votre participation sans conséquence négative ou préjudice et sans avoir à justifier votre décision. Si vous décidez de mettre fin à votre participation, il est important de le signaler à l'expérimentateur. Tous les renseignements personnels vous concernant seront alors détruits.

Les mesures suivantes seront appliquées pour assurer la confidentialité des renseignements fournis par les participants:

- les noms des participants ne paraîtront dans aucun rapport;
- les divers documents de la recherche seront codifiés ; seul le chercheur aura accès à la liste des noms et des codes;
- les résultats individuels des participants ne seront jamais communiqués;
- les matériaux de la recherche, incluant les données et les enregistrements, seront conservés sur ordinateur ou sur un autre support ; les fichiers seront protégés par un mot de passe.

Je soussigné(e) _____

demeurant à _____

consens librement à participer à la recherche intitulée : « Apport de l'information visuelle à la perception de la parole». J'ai pris connaissance du formulaire et j'ai compris le but, la nature, les avantages, les risques et les inconvénients du projet de recherche. Je suis satisfait(e) des explications, précisions et réponses que le chercheur m'a fournies, le cas échéant, quant à ma participation à ce projet.

Signature du participant, de la participante Date

J'ai expliqué le but, la nature, les avantages, les risques et les inconvénients du projet de recherche au participant. J'ai répondu au meilleur de ma connaissance aux questions posées et j'ai vérifié la compréhension du participant.

Signature du chercheur Date

Annexe 2

Formulaire de consentement parental libre et éclairé

Cette recherche est réalisée dans le cadre de la thèse doctorale de Grozdana Erjavec (grozdana.erjavec@etud.univ-paris8.fr) sous la direction de Denis Legros, professeur émérite. Grozdana Erjavec est affiliée à l'école doctorale Cognition, Langage, Interaction (Université de Paris VIII) et au groupe Lutin (http://www.lutin-userlab.fr/accueil/) du laboratoire CHART.

Avant d'accepter d'autoriser votre enfant à participer à ce projet de recherche, veuillez prendre le temps de lire et de comprendre les renseignements qui suivent. Ce document vous explique le but de ce projet de recherche, ses procédures, avantages, risques et inconvénients. Nous vous invitons à poser toutes les questions que vous jugerez utiles à la personne qui vous présente ce document.

La recherche a pour but d'étudier l'apport de l'information visuelle, relative aux mouvements articulatoires, à la perception de la parole.

Aucune technique invasive n'est utilisée dans la recherche. La recherche ne comporte pas de risques ou d'inconvénients notables.

La participation de votre enfant à cette recherche consiste à :

- Suivre la présentation du matériel (5 min) ;
- Visualiser une série de vidéos comportant des syllabes prononcées par une personne française de sexe féminin et dégradées par du bruit dans le but de les répéter à haute voix (10 min).

Votre enfant est libre de participer à ce projet de recherche. Il/elle peut aussi mettre fin à sa participation sans conséquence négative ou préjudice et sans avoir à justifier sa décision. Si votre enfant décide de mettre fin à votre participation, il est important de le signaler à l'expérimentateur. Tous les renseignements personnels concernant votre enfant seront alors détruits.

Les mesures suivantes seront appliquées pour assurer la confidentialité des renseignements fournis par les participants:

- les noms des participants ne paraîtront dans aucun rapport;
- les divers documents de la recherche seront codifiés ; seul le chercheur aura accès à la liste des noms et des codes;
- les résultats individuels des participants ne seront jamais communiqués;

- les matériaux de la recherche, incluant les données et les enregistrements, seront conservés sur ordinateur ou sur un autre support ; les fichiers seront protégés par un mot de passe.

Je soussigné(e) _____

demeurant à _____

ayant l'autorité parentale sur _____

autorise mon enfant à participer à la recherche intitulée : « Apport de l'information visuelle à la perception de la parole». J'ai pris connaissance du formulaire et j'ai compris le but, la nature, les avantages, les risques et les inconvénients du projet de recherche. Je suis satisfait(e) des explications, précisions et réponses que le chercheur m'a fournies, le cas échéant, quant à ma participation à ce projet.

Signature du participant, de la participante Date

J'ai expliqué le but, la nature, les avantages, les risques et les inconvénients du projet de recherche au participant. J'ai répondu au meilleur de ma connaissance aux questions posées et j'ai vérifié la compréhension du participant.

Signature du chercheur Date

Annexe 3

AUTORISATION DE REPRODUCTION ET DE REPRESENTATION DE PHOTOGRAPHIES

Je soussigné……

Demeurant……..

Autorise :……..,

à me photographier et à me filmer,

le ……

A :……

Et à utiliser mon image ;

En conséquence de quoi et conformément aux dispositions relatives au droit à l'image et au droit au nom, **j'autorise ……à fixer, reproduire et communiquer au public les vidéos enregistrées et les photographies prises dans le cadre de la présente**.

Les photographies pourront être exploitées et utilisées directement par Mademoiselle Grozdana Erjavec ou être cédées à des tiers, sous toute forme et tous supports connus et inconnus à ce jour, dans le monde entier, sans limitation de durée, intégralement ou par extraits et notamment :

- Presse,
- Livre,
- Carte postale,
- Exposition,
- Publicité,
- Projection publique,
- Concours,
- Expositions ou conférences scientifiques.

Le bénéficiaire de l'autorisation s'interdit expressément de procéder à une exploitation des photographies susceptible de porter atteinte à la vie privée ou à la réputation, ni d'utiliser les photographies de la présente, dans tout support à caractère pornographique, raciste, xénophobe ou toute autre exploitation préjudiciable.

Il s'efforcera dans la mesure du possible, de tenir à disposition un justificatif de chaque parution des photographies sur simple demande. Il encouragera ses partenaires à faire de même et mettra en œuvre tous les moyens nécessaires à la réalisation de cet objectif.

Je me reconnais être entièrement rempli de mes droits et je ne pourrai prétendre à aucune rémunération pour l'exploitation des droits visés aux présentes.

Je garantis que je ne suis pas lié par un contrat exclusif relatif à l'utilisation de mon image ou de mon nom.

Election de domicile est faite par chacune des parties à l'adresse précisée aux présentes.

Pour tout litige né de l'interprétation ou de l'exécution des présentes, il est fait attribution expresse de juridiction aux tribunaux compétents statuant en droit français.

Fait à ……., le……en deux exemplaires et de bonne foi.

Signatures précédées du nom et prénom :

Annexe 4

Blocs Ordre 1	**AVV**		**AU**		**AVB-E**		**AVB-M**	
Sous-blocs Ordre 1.1	SNR-6	SNR-12	SNR-6	SNR-12	SNR-6	SNR-12	SNR-6	SNR-12
Sous-blocs Ordre 1.2	SNR-12	SNR-6	SNR-12	SNR-6	SNR-12	SNR-6	SNR-12	SNR-6
Blocs Ordre 2	**AU**		**AVB-M**		**AVV**		**AVB-E**	
Sous-blocs Ordre 2.1	SNR-6	SNR-12	SNR-6	SNR-12	SNR-6	SNR-12	SNR-6	SNR-12
Sous-blocs Ordre 2.2	SNR-12	SNR-6	SNR-12	SNR-6	SNR-12	SNR-6	SNR-12	SNR-6
Blocs Ordre 3	**AVB-M**		**AVB-E**		**AU**		**AVV**	
Sous-blocs Ordre 3.1	SNR-6	SNR-12	SNR-6	SNR-12	SNR-6	SNR-12	SNR-6	SNR-12
Sous-blocs Ordre 3.2	SNR-12	SNR-6	SNR-12	SNR-6	SNR-12	SNR-6	SNR-12	SNR-6
Blocs Ordre 4	**AVB-E**		**AVV**		**AU**		**AVB-M**	
Sous-blocs Ordre 4.1	SNR-6	SNR-12	SNR-6	SNR-12	SNR-6	SNR-12	SNR-6	SNR-12
Sous-blocs Ordre 4.2	SNR-12	SNR-6	SNR-12	SNR-6	SNR-12	SNR-6	SNR-12	SNR-6

Annexe 5

	AVV SNR-12	AVV SNR-6	AU SNR-12	AU SNR-6	AVB-E SNR-12	AVB-E SNR-6	AVB-M SNR-12	AVB-M SNR-6
Item de l'illustration	**WA**		**GNA**		**WA**		**GNA**	
Items des essais critiques	JA	ZA	JA	GA	TA	LA	JA	GA
	TA	LA	LA	BA	JA	JA	BA	NA
	BA	DA	MA	LA	VA	VA	FA	KA
	RA	TA	FA	NA	KA	SA	SA	TA
	NA	SA	CHA	MA	RA	DA	PA	VA
	PA	RA	NA	RA	MA	PA	RA	MA
	VA	NA	DA	ZA	GA	TA	LA	JA
	MA	CHA	VA	DA	DA	FA	MA	BA
	CHA	KA	KA	KA	CHA	GA	VA	DA
	KA	JA	GA	SA	BA	BA	CHA	PA
	ZA	GA	ZA	FA	PA	ZA	DA	FA
	SA	MA	SA	PA	ZA	CHA	GA	ZA
	FA	FA	BA	JA	NA	NA	ZA	LA
	GA	BA	TA	VA	LA	KA	KA	SA
	DA	PA	RA	TA	SA	RA	TA	RA
	LA	VA	PA	CHA	FA	MA	NA	CHA

Annexe 6

Item 1	**Wa**			
Item 2	Cha	Fa	Ja	Autre
Item 3	Na	Da	Ta	Autre
Item 4	Ba	Ga	Pa	Autre
Item 5	Ja	La	Ra	Autre
Item 6	Da	Na	Ma	Autre
Item 7	Pa	Ba	Cha	Autre
Item 8	Fa	Va	Ka	Autre
Item 9	Na	Ma	Za	Autre
Item 10	Ja	Ba	Sha	Autre
Item 11	Ka	Ga	La	Autre
Item 12	Sa	Va	Za	Autre
Item 13	Sa	Za	La	Autre
Item 14	Va	Na	Fa	Autre
Item 15	Ka	Sa	Ga	Autre
Item 16	Pa	Da	Ta	Autre
Item 17	Ja	Na	La	Autre
Item18	Sa	Za	Ka	Autre
Item 19	La	Va	Na	Autre
Item 20	Fa	Da	Ta	Autre
Item 21	Ta	Ma	Da	Autre
Item 22	Za	Na	Sa	Autre
Item 23	Ra	La	Cha	Autre
Item 24	Ma	Na	Za	Autre

Item 25	Ta	Cha	Ja	Autre
Item 26	Fa	Ka	Ga	Autre
Item 27	Ja	Pa	Cha	Autre
Item 28	Ka	Ga	Ra	Autre
Item 29	Na	Ka	Ma	Autre
Item 30	Fa	Ta	Va	Autre
Item 31	Pa	Sa	Ba	Autre
Item 32	Ka	Ba	Pa	Autre
Item 33	Va	La	Fa	Autre
Item 34	**Gna**			
Item 35	Ja	Ra	Cha	Autre
Item 36	Va	Na	La	Autre
Item 37	Ja	Na	Ma	Autre
Item 38	Fa	Ta	Va	Autre
Item 39	Ja	Cha	Ba	Autre
Item 40	Na	Sa	Ma	Autre
Item 41	Ta	La	Da	Autre
Item 42	Va	Fa	Pa	Autre
Item 43	Ga	Na	Ka	Autre
Item 44	Ka	Ga	Fa	Autre
Item 45	Da	Sa	Za	Autre
Item 46	Sa	Ma	Za	Autre
Item 47	Ba	Ja	Pa	Autre
Item 48	Fa	Da	Ta	Autre

Item 49	La	Ra	Sa	Autre
Item 50	Cha	Pa	Ba	Autre
Item 51	Ka	Da	Ga	Autre
Item 52	Ba	Ja	Pa	Autre
Item 53	Na	Va	La	Autre
Item 54	Ma	Na	Ga	Autre
Item 55	Na	Sa	Ma	Autre
Item 56	Ra	Za	La	Autre
Item 57	Za	Sa	Va	Autre
Item 58	Ta	Ba	Da	Autre
Item 59	Ja	Ka	Ga	Autre
Item 60	Za	La	Sa	Autre
Item 61	Va	Fa	Ma	Autre
Item 62	Pa	Ta	Ba	Autre
Item 63	Ka	Ja	Cha	Autre
Item 64	Va	Fa	Na	Autre
Item 65	Da	Ra	Ta	Autre
Item 66	Ga	Cha	Ja	Autre
Item 67	**Wa**			
Item 68	Ta	Pa	Da	Autre
Item 69	Ma	Cha	Ja	Autre
Item 70	Cha	Va	Fa	Autre
Item 71	Ga	La	Ka	Autre
Item 72	Ra	Fa	La	Autre

Item 73	Va	Na	Ma	Autre
Item 74	Ga	Ma	Ka	Autre
Item 75	Ta	Da	Ja	Autre
Item 76	Ma	Cha	Ja	Autre
Item 77	Ba	Fa	Pa	Autre
Item 78	Pa	Ba	Ra	Autre
Item 79	Ta	Sa	Za	Autre
Item 80	Na	Ba	Ma	Autre
Item 81	Na	La	Ja	Autre
Item 82	Pa	Za	Sa	Autre
Item 83	Fa	Ta	Va	Autre
Item 84	Ma	Ra	La	Autre
Item 85	Ka	Cha	Ja	Autre
Item 86	Ta	Va	Fa	Autre
Item 87	Za	Ma	Sa	Autre
Item 88	Da	Ta	Cha	Autre
Item 89	Pa	Ga	Ba	Autre
Item 90	Fa	Da	Ta	Autre
Item 91	Va	Fa	Na	Autre
Item 92	Ka	Ga	Ja	Autre
Item 93	La	Ba	Pa	Autre
Item 94	Za	Va	Sa	Autre
Item 95	Ja	Ta	Cha	Autre
Item 96	Ma	Za	Na	Autre

Item 97	Ka	Ga	Sa	Autre
Item 98	La	Ra	Pa	Autre
Item 99	Ja	Ma	Na	Autre
Item 100	**Gna**			
Item 101	Va	Cha	Ja	Autre
Item 102	Pa	Ga	Ba	Autre
Item 103	Fa	Ta	Va	Autre
Item 104	Ma	Sa	Za	Autre
Item 105	Pa	Ba	Cha	Autre
Item 106	Ra	Fa	La	Autre
Item 107	Na	Za	La	Autre
Item 108	Ta	Na	Ma	Autre
Item 109	Va	Fa	Ka	Autre
Item 110	Na	Cha	Ja	Autre
Item 111	Da	Va	Ta	Autre
Item 112	Ka	Ga	Ja	Autre
Item 113	Ra	Sa	Za	Autre
Item 114	Ga	Ka	Na	Autre
Item 115	Ta	Za	Da	Autre
Item 116	Sa	Na	Ma	Autre
Item 117	Ka	Ga	Cha	Autre
Item 118	Na	Fa	Ma	Autre
Item 119	Ga	La	Ka	Autre
Item 120	Ta	Sa	Da	Autre

Item 121	Fa	Va	Ra	Autre
Item 122	Na	Ma	Ja	Autre
Item 123	Sa	Pa	Ba	Autre
Item 124	Ja	Ga	Cha	Autre
Item 125	Ta	Ra	Da	Autre
Item 126	Pa	Ba	Za	Autre
Item 127	Va	Na	Fa	Autre
Item 128	Pa	Sa	Za	Autre
Item 129	Cha	La	Na	Autre
Item 130	Sa	Za	Ga	Autre
Item 131	La	Ta	Ra	Autre
Item 132	Pa	Ja	Cha	Autre

Annexe 7

Stimulus de l'essai critique	**Stimulus acoustiquement proche**	**Stimuli acoustiquement éloignés**
BA /ba/	PA /pa/	FA /fa/ ; GA /ga/ ; KA /ka/, LA /la/ ; NA /na/ ; RA /ʁa/ ; SA /sa/ ; CHA /ʃa/ ; TA /ta/ ; ZA /za/ ; JA /ʒa/
DA /da/	TA /ta/	FA /fa/ ; GA /ga/ ; KA /ka/, LA /la/ ; MA /ma/ ; PA /pa/ ; RA /ʁa/ ; SA /sa/ ; CHA /ʃa/ ; VA /va/ ; ZA /za/ ; JA /ʒa/
FA /fa/	VA /va/	BA /ba/ ; DA /da/ ; GA /ga/ ; KA /ka/, LA /la/ ; MA /ma/ ; NA /na/ ; PA /pa/ ; RA /ʁa/ ; SA /sa/ ; CHA /ʃa/ ; TA /ta/ ; ZA /za/ ; JA /ʒa/
GA /ga/	KA /ka/	BA /ba/ ; DA /da/ ; FA /fa/ ; LA /la/ ; MA /ma/, NA /na/ ; PA /pa/ ; RA /ʁa/ ; SA /sa/ ; CHA /ʃa/ ; TA /ta/ ; VA /va/ ; ZA /za/ ; JA /ʒa/
KA /ka/	GA /ga/	BA /ba/ ; DA /da/ ; FA /fa/ ; LA /la/ ; MA /ma/ ; NA /na/ ; PA /pa/ ; RA /ʁa/ ; SA /sa/ ; CHA /ʃa/ ; TA /ta/ ; VA /va/ ; ZA /za/ ; JA /ʒa/
LA /la/	NA /na/	BA /ba/ ; DA /da/ ; FA /fa/ ; GA /ga/ ; KA /ka/ ; MA /ma/ ; PA /pa/ ; SA /sa/ ; CHA /ʃa/ ; TA /ta/ ; VA /va/ ; ZA /za/ ; JA /ʒa/
MA /ma/	NA /na/	DA /da/ ; FA /fa/ ; GA /ga/ ; KA /ka/, LA /la/ ; RA /ʁa/ ; SA /sa/ ; CHA /ʃa/ ; TA /ta/ ; VA /va/ ; ZA /za/ ; JA /ʒa/
NA /na/	MA /ma/	BA /ba/ ; DA /da/ ; FA /fa/ ; GA /ga/ ; KA /ka/, PA /pa/ ; RA /ʁa/ ; SA /sa/ ; CHA /ʃa/ ; TA /ta/ ; VA /va/ ; ZA /za/ ; JA /ʒa/
PA /pa/	BA /ba/	DA /da/ ; FA /fa/ ; GA /ga/ ; KA /ka/, LA /la/ ; NA /na/ ; RA /ʁa/ ; SA /sa/ ; CHA /ʃa/ ; TA /ta/ ; VA /va/ ; ZA /za/ ; JA /ʒa/
RA /ʁa/	LA /la/	BA /ba/ ; DA /da/ ; FA /fa/ ; GA /ga/ ; KA /ka/ ; MA /ma/ ; NA /na/ ; PA /pa/ ; SA /sa/ ; CHA /ʃa/ ; TA /ta/ ; VA /va/ ; ZA /za/ ; JA /ʒa/
SA /sa/	ZA /za/	BA /ba/ ; DA /da/ ; FA /fa/ ; GA /ga/ ; KA /ka/ ; LA /la/ ; MA /ma/ ; NA /na/ ; PA /pa/ ; RA /ʁa/ ; TA /ta/ ; VA /va/ ; JA /ʒa/
CHA /ʃa/	JA /ʒa/	BA /ba/ ; DA /da/ ; FA /fa/ ; GA /ga/ ; KA /ka/ ; LA /la/ ; MA /ma/ ; NA /na/ ; PA /pa/ ; RA /ʁa/ ; SA /sa/ ; TA /ta/ ; VA /va/
TA /ta/	DA /da/	BA /ba/ ; FA /fa/ ; GA /ga/ ; KA /ka/ ; LA /la/ ; MA /ma/ ; NA /na/ ; PA /pa/ ; RA /ʁa/ ; SA /sa/ ; CHA /ʃa/ ; VA /va/ ; ZA /za/ ; JA /ʒa/
VA /va/	FA /fa/	BA /ba/ ; DA /da/ ; GA /ga/ ; KA /ka/ ; LA /la/ ; MA /ma/ ; NA /na/ ; PA /pa/ ; RA /ʁa/ ; SA /sa/ ; CHA /ʃa/ ; TA /ta/ ; ZA /za/ ; JA /ʒa/
ZA /za/	SA /sa/	BA /ba/ ; DA /da/ ; FA /fa/ ; GA /ga/ ; KA /ka/ ; LA /la/ ; MA /ma/ ; NA /na/ ; PA /pa/ ; RA /ʁa/ ; CHA /ʃa/ ; TA /ta/ ; VA /va/
JA /ʒa/	CHA /ʃa/	BA /ba/ ; DA /da/ ; FA /fa/ ; GA /ga/ ; KA /ka/ ; LA /la/ ; MA /ma/ ; NA /na/ ; PA /pa/ ; RA /ʁa/ ; TA /ta/ ; VA /va/ ; ZA /za/