Analyse de scènes de parole multisensorielle : Mise en évidence et caractérisation d'un processus de liage audiovisuel préalable à la fusion.

Olha Nahorna

UNIVERSITÉ DE GRENOBLE

THÈSE
Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE
Spécialité : **ISCE / INGENIERIE COGNITION INTERACTION APPRENTISSAGE CREATION**
Arrêté ministériel : 7 août 2006

Présentée par
Olha NAHORNA

Thèse dirigée par **Jean-Luc SCHWARTZ** et **Frédéric BERTHOMMIER**

préparée au sein du **Laboratoire GIPSA-Lab**
dans **l'École Doctorale EDISCE**

Analyse de scènes de parole multisensorielle :

Mise en évidence et caractérisation d'un processus de liage audiovisuel préalable à la fusion

Thèse soutenue publiquement le **2 octobre 2013**,
devant le jury composé de :

Mme Cécile Colin
Assistante professeur à l'Université de Bruxelles (Rapporteur)

Mme Virginie van Wassenhove
Chercheuse CEA, HDR (Rapporteur)

Mme Sonia Kandel
Professeur UPMF (Examinatrice)

Mr Daniel Pressnitzer
DR CNRS Paris (Examinateur)

Mr Jean-Luc Schwartz
DR CNRS Grenoble (Directeur de thèse)

Mr Frédéric Berthommier
CR CNRS Grenoble (Co-directeur de thèse)

Remerciements

Table des matières

Introduction

La parole est un mode de communication propre à l'homme, apparaissant comme une nouvelle étape dans l'évolution. Les mécanismes de la communication parlée sont complexes et partent des propriétés acoustiques du message sonore, puis mettent en jeu des processus de catégorisation des consonnes, voyelles, syllabes, de compréhension des mots, phrases, à travers des processus lexicaux, syntactiques etc. Finalement le processus converge vers la pensée abstraite, qui est le niveau le plus élevé de l'activité intellectuelle propre à Homo-Sapiens. En s'appuyant sur les théories de l'évolution nous pouvons supposer que la parole peut apparaître comme une superstructure s'appuyant sur les propriétés générales des systèmes de traitement et de production des signaux, que nous pouvons retrouver chez les organismes moins évolués, dans de nombreuses espèces animales disposant de capteurs et d'actionneurs similaires :les yeux, les oreilles, le toucher ; et un système de production sonore. Mais par rapport à ces organismes nous sommes capables d'encoder et de traiter des séquences complexes de ces signaux pour référer aux objets, aux actions, à l'espace, au temps, etc. En autres termes nous sommes capables de décoder ces concepts par un symbole ou une étiquette, et puis de les enchainer dans des structures syntaxiques complexes.

Le rôle majeur de la parole est de passer une information ou message et de transmettre une expérience connue. Elle peut être aussi un instrument du travail intellectuel (pensée, réflexions) ou un outil pour exprimer les sentiments, les émotions, l'état intérieur, les idées, voire pour mentir ou tromper. La parole peut être aussi un instrument pour influencer directement ou indirectement son interlocuteur.

Parler c'est produire des gestes et ainsi des sons qui peuvent être décodés et compris par l'interlocuteur, ainsi que savoir décoder le son et les gestes de l'interlocuteur. La production de la parole nécessite des processus d'articulation qui impliquent des organes multiples : cordes vocales, larynx, trachée, langue, mâchoire, lèvres, velum. Elle dépend aussi du cycle respiratoire, grâce à la passation d'air nécessaire à l'émission du son.

Mais les gestes de la parole produisent aussi des signaux visibles, et la parole est un mode de communication non seulement acoustique mais audiovisuel. Dans le cadre du travail de cette thèse nous nous intéressons essentiellement à la parole audiovisuelle, et plus précisément le but de notre étude est de mieux comprendre les mécanismes qui permettent de lier les signaux auditifs et visuels de la parole.

Nous commençons ce manuscrit par une première partie théorique, avec une revue des questions principales de bibliographie parmi lesquelles nous pourrons passer en revue les faits démontrant que la parole est audiovisuelle, puis décrire les principaux modèles d'intégration des inputs auditif et visuels, et les questions sous-jacentes. Puis nous aborderons la question d'existence d'un niveau précoce d'intégration audio-visuelle qui nous conduira à formuler notre hypothèse principale d'existence d'un processus de« liage audiovisuel », qui constituerait une étape préalable à la fusion audio-visuelle.

En nous basant sur cette hypothèse nous aborderons la phase expérimentale de notre travail, organisée en deux parties. D'abord nous entreprendrons de démontrer l'existence du processus du liage en utilisant et en affinant peu à peu un nouveau paradigme expérimental. Puis nous traiterons de la caractérisation de ce processus en mettant en place des expériences qui étudient l'influence de différents paramètres critiques. Nous terminerons ce manuscrit par une discussion sur nos résultats et leur positionnement dans la littérature actuelle.

Partie I
De la parole audiovisuelle à la question du liage : un état de l'art pour une stratégie expérimentale

Chapitre 1. La parole audiovisuelle

1.1 La parole est audiovisuelle

Bien que la lecture labiale ait de tout temps été étudiée et utilisée dans l'éducation et la communication avec les personnes sourdes, les recherches sur la parole ont longtemps été focalisées sur la parole acoustique. Ce n'est que depuis le XXème siècle que sont apparues des mises en évidence incontestables et multiples que la parole est plutôt de nature audiovisuelle.

Dans la suite de ce chapitre, nous ne détaillerons pas les données sur les sujets malentendants, pour lesquels le rôle de la modalité visuelle est une évidence et une nécessité. Chacun sait bien que les sourds ou les malentendants peuvent « lire sur les lèvres » pour compenser leur handicap (Bernstein et al, 2000), et on peut également compenser l'information visuelle incomplète en utilisant le « Langage parlé Complété » (Attina, 2005). Mais nous allons montrer que la lecture labiale est impliquée dans le traitement de la parole chez tous les sujets, dans différents types de conditions et de paradigmes qui font l'objet de cette première section.

Par exemple, c'est le cas de l'effet "Cocktail party", mentionné par (Cherry, 1953), qui vient d'une situation naturelle que nous pouvons tous observer au cours de notre vie. Lors d'une soirée, où il y a des nombreuses personnes qui parlent simultanément, en créant un bruit environnemental assez fort, nous arrivons à poursuivre la communication et comprendre le message de notre interlocuteur, malgré le bruit. En même temps nous restons sensibles aux signaux extérieurs de notre conversation, par exemple, si quelqu'un prononce notre prénom, nous réagissons très rapidement à ce signal. Parmi des explications possibles à ce phénomène, la lecture labiale semble avoir une place importante.

Le paradigme de mise en évidence est bien illustré par les travaux de Sumby et Pollack, sur l'anglais américain (Sumby & Pollack, 1954). Des vocabulaires de 8, 16, 32, 64, 128 mots bisyllabiques spondaïques[1] étaient présentés à 6 sujets sous diverses conditions de bruit. Dans ces résultats l'intelligibilité de la parole diminue avec la diminution du rapport parole à bruit et avec l'augmentation de la taille du vocabulaire (Figure 1, à gauche). Dans la condition audiovisuelle les relations entre la taille du vocabulaire et le rapport parole à bruit sont similaires à celles de la condition auditive, mais la résistance au bruit est plus forte (Figure 1, à droite). Ainsi, la lecture labiale produit un gain d'intelligibilité dans le bruit ou quand la parole

[1] Les mots spondaïques sont les mots composés de deux syllabes longues, par exemple «cupcake », « baseball »

auditive est peu fiable. Plus tard, Erber (Erber, 1969) a observé des résultats semblables dans le même type de paradigme. De plus, Erber a mis en évidence une importante variabilité inter-sujets dans la condition audiovisuelle, qu'il explique par la variation des compétences de lecture labiale parmi les sujets naïfs.

Plus récemment, Benoit et al. (Benoit et al, 1994) ont confirmé en français l'impact de la lecture labiale sur l'intelligibilité dans la condition bruitée, en utilisant cette fois des non-mots VCVCV. Les résultats obtenus (Figure 2) sont similaires à ceux décrits par Sumby et Pollack (Sumby & Pollack, 1954) et Erber (Erber, 1969).

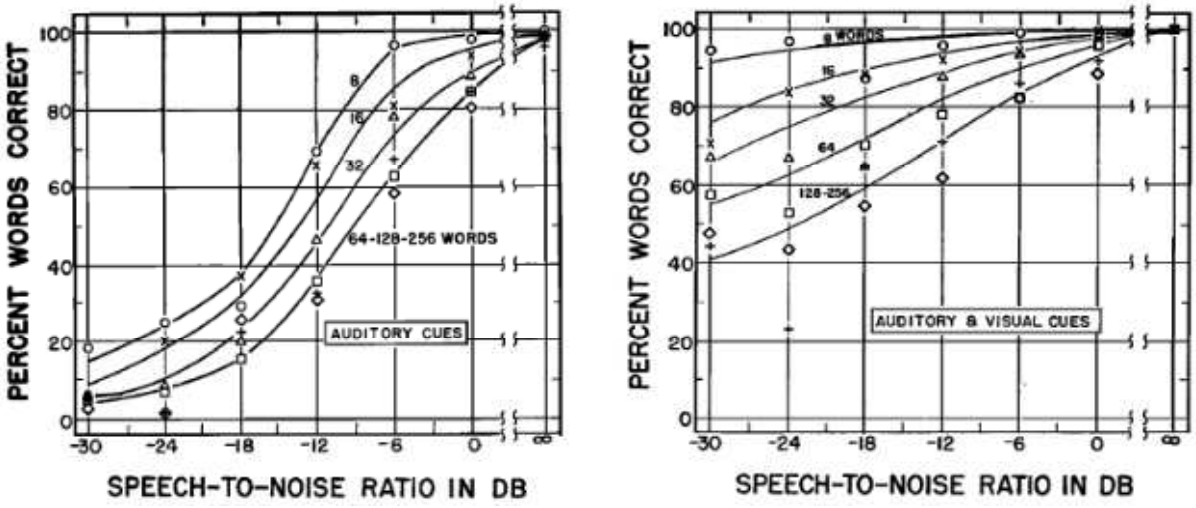


Figure 1 – Intelligibilité de la parole en fonction du rapport parole à bruit : à gauche dans la condition audio pur, à droite dans la condition audiovisuelle. Les courbes sont paramétrées par la taille du vocabulaire composé de mots spondaïques : 8, 16, 32, 64, 128 mots. Chaque point représente la moyenne de 450 réponses. Figure tirée de (Sumby & Pollack, 1954)

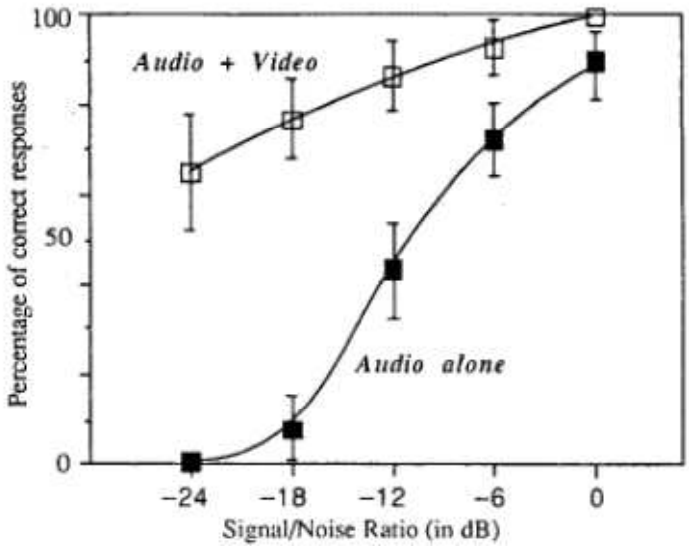


Figure 2 - Taux d'intelligibilité en fonction du rapport signal à bruit dans les conditions « auditif » et « audiovisuelle » pour 18 non-mots et 18 sujets. Figure tirée de (Benoit et al, 1994)

La lecture labiale améliore également la compréhension dans le cas d'un texte compliqué (Reisberg et al, 1987), (Thompson & Ogden, 1995) ou du traitement d'une langue étrangère (Davis & Kim, 2001). Notamment l'information visuelle sur la langue, les dents, les lèvres, la mâchoire, les joues améliore la précision de répétition de phrases courtes pendant l'apprentissage d'une langue étrangère (coréen pour les sujets anglophones), et facilite les performances de reconnaissance et de rapidité de phrases (Davis & Kim, 2001). Dans l'expérience de Davis et al. (Davis & Kim, 2001) les sujets, n'ayant jamais entendu le coréen, était enjoints à répéter des phrases courtes de durées 1-1.5s prononcées en coréen. Ces phrases étaient accompagnées soit par la partie haute du visage d'une locutrice native, soit par la partie basse. Les sujets présentent un score significatif de meilleure répétition s'ils ont vu la partie basse du visage (Figure 3, à gauche). Dans une deuxième session où les sujets ont déjà entendu la phrase présentée dans la première session et où on teste leur capacité de reconnaissance, on obtient le même effet (Figure 3, à droite).

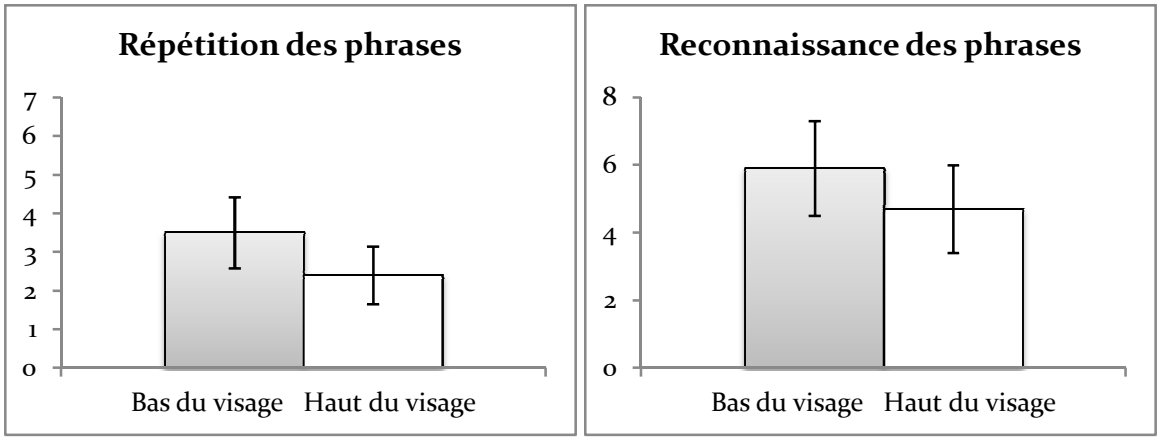


Figure 3 – Taux moyen de réussite à l'épreuve de répétition de phrases coréennes par des sujets ne connaissant pas le coréen dans l'expérience de Davis et al. (première session expérimentale, à gauche), et de reconnaissance si la phrase avait déjà été précédemment entendue (deuxième session, à droite) (d'après (Davis & Kim, 2001)).

La présentation des lèvres du locuteur améliore les résultats d'une tâche de répétition de phrases même dans la langue natale des sujets (anglais) (Figure 4) (Thompson & Ogden, 1995).

Reisberg (Reisberg, 1978) a fait une étude inverse, en étudiant l'influence de l'information auditive sur la compréhension de phrases par la lecture labiale. 15 phrases était préparées (présentées 3 fois par condition), que les sujets devaient identifier en choisissant une réponse parmi 15 dans la liste. Pour étudier l'influence de la présence d'un signal auditif deux conditions ont été proposées : (1)ajout d'impulsions auditives de fréquence constante mais d'intensité cohérente avec l'intensité du signal source (donc fournissant des informations sur l'intensité mais pas sur la fréquence fondamentale) et (2)impulsions de fréquence égale à la fréquence fondamentale du signal source d'origine, et les mêmes variations de l'intensité (donc fournissant des informations à la fois sur l'intensité et la fréquence fondamentale). Les résultats montrent une forte amélioration de la compréhension en rajoutant ces deux types d'information auditive, avec une augmentation d'intelligibilité plus forte si la quantité d'information auditive fournie est plus importante (Figure 5).

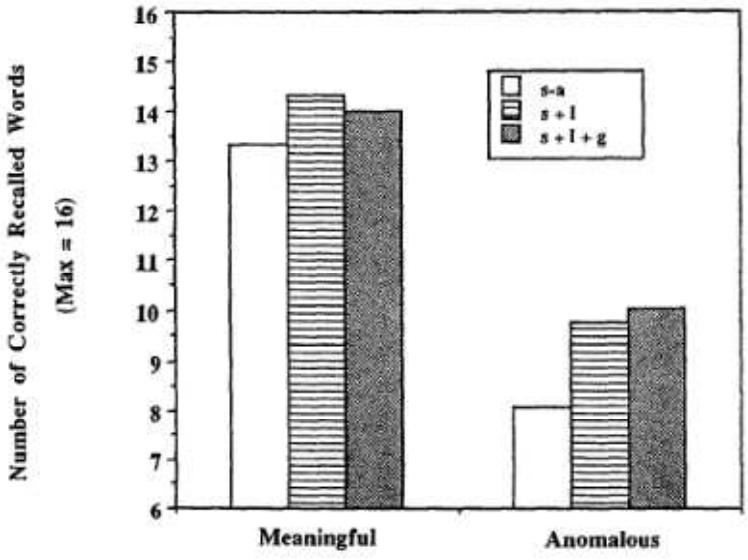


Figure 4 – Moyenne des mots correctement rappelés pour des phrases sémantiquement cohérentes de type « The bicycle tire exploded when he filled it with too much air at the gas station » ou incohérentes, telles que « The sandwiches soon let go when he swung it with too much zoo at the night» dans les conditions: s-a audio seul, s+l audio plus lèvres, s+l+g audio avec lèvres et geste iconique. Figure tirée de (Thompson & Ogden, 1995).

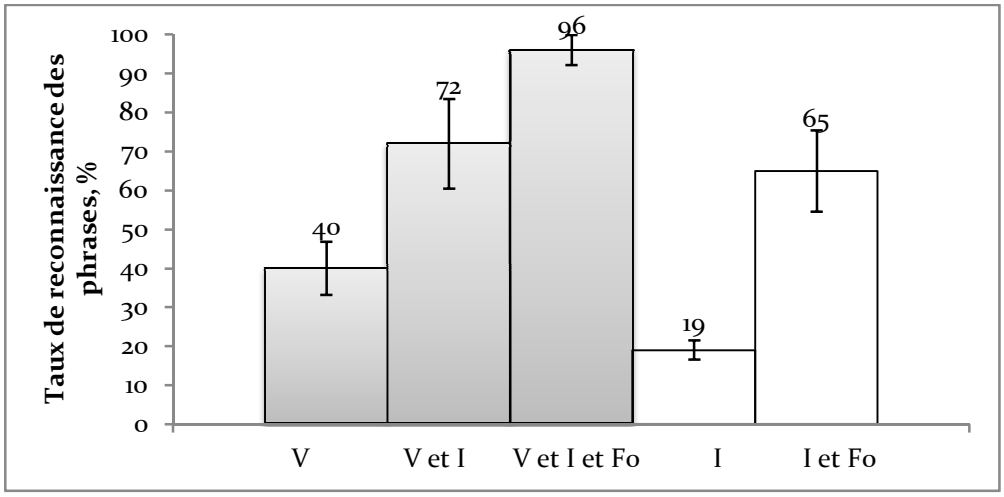


Figure 5– Taux de reconnaissance de phrases par la lecture labiale (V), par la lecture labiale avec une information sur l'intensité auditive (V et I), par la lecture labiale avec une information sur l'intensité et sur la fréquence fondamentale (V et I et Fo), comparées aux taux avec une information sur l'intensité auditive seule (I) et sur l'intensité et la fréquence fondamentale (I et Fo)

Grant et Greenberg (Grant & Greenberg, 2001) ont également étudié l'influence intermodale sur la compréhension de phrases. Le signal auditif était filtré par canaux 1/3 d'octave, en ne conservant finalement que le canal basse fréquence (298-375 Hz) et le canal haute fréquence (4762-6000 Hz). L'intelligibilité de parole auditive seule était autour de 19%, la lecture labiale à ~11% de compréhension, mais la combinaison des deux modalités améliore

considérablement la performance jusqu'à 63%. Une fois encore, et ici de manière très spectaculaire, la parole multimodale est beaucoup plus performante que chacune des modalités prise isolément.

La découverte de l'effet McGurk en 1976 (McGurk & MacDonald, 1976) nous donne une autre mise en évidence extrêmement frappante de la nature audiovisuelle de la parole. Dans l'exemple classique le montage d'un stimulus audio « ba » avec un stimulus vidéo « ga » est souvent perçu comme un stimulus « da » ou « tha ». A l'inverse, le montage d'un stimulus audio « ga » avec un stimulus vidéo « ba » produit un percept de combinaison phonétique de type « bga ». Cet effet montre l'influence d'un signal visuel incongruent sur un signal auditif dans le percept final. Cet effet est souvent utilisé dans la littérature comme une preuve majeure de la nature audiovisuelle de la parole. L'observation de l'effet McGurk à travers des langues et cultures différentes montre que cet effet est remarquablement stable, malgré d'importants facteurs de variation interindividuelle. Nous reviendrons en détail sur la phénoménologie de l'effet McGurk dans une section ultérieure de cette revue de questions.

Bien d'autres aspects des relation entre vision et auditions dans le traitement cognitif de la parole ont été étudiés dans la littérature au long de ces dernières années (Bernstein et al, 2002): attention et mémoire (Driver, 1996), plasticité de la parole (Beautemps et al, 1999), prosodie (Risberg & Lubker, 1978) , (Munhall et al, 2004), émotion (de Gelder & Vroomen, 2000), perception des tons (Burnham et al, 2001). Mais nous ne traiterons pas ces études en détail dans la suite de ce manuscrit.

Toutes ces données montrent que la parole n'est pas uniquement auditive et que dans la plupart des circonstances les deux modalités sont impliquées pour mieux percevoir le message. La section suivante vise à préciser le rôle d'un signal visuel dans la parole.

1.2 Rôle d'un signal visuel

1.2.1 Redondance et complémentarité

Le signal visuel dans le flux de parole possède par rapport au flux auditif des propriétés à la fois de redondance et de complémentarité. Une partie de l'information visuelle est étroitement corrélée à certaines propriétés spectro-temporelles du signal auditif, mais il y une autre partie qui rajoute une information complémentaire (Campbell, 2008) .

Le rôle de corrélation, qui duplique partiellement l'information auditive, permet au système de vérifier en permanence que les deux sources sont cohérentes l'une par rapport à l'autre. Si la corrélation produit cette cohérence, alors la partie complémentaire peut être prise en compte dans le traitement de la parole. Elle peut enrichir la quantité d'information transmise qui peut être prise en compte, préciser l'information peu claire, améliorer la détection et la prédiction de la parole, augmenter l'intelligibilité et la rapidité de traitement. Ainsi, dans le cadre d'une conversation téléphonique, nous avons souvent des difficultés à distinguer les sons « m » et « n ». La différence entre ces deux phonèmes, faible acoustiquement, est par contre forte visuellement, notamment par la différence du geste d'ouverture de la bouche, impliquant une fermeture labiale dans le premier cas et pas dans le second. Etudions maintenant à quoi peuvent servir ces propriétés de redondance et complémentarité du signal visuel.

1.2.2 Amélioration de la compréhension du message

Le signal visuel facilite le traitement et la reconnaissance de la parole, comme nous l'avons vu précédemment dans les travaux de (Sumby & Pollack, 1954), (Erber, 1969), (Benoit et al, 1994). Cette facilitation a lieu dans les conditions bruitées, mais aussi quand la parole acoustique est peu prédictible (Grant & Seitz, 1998). Mais cet effet de facilitation a lieu aussi dans les conditions normales. L'identification des stimuli bimodaux est meilleure (Reisberg et al, 1987), (Arnold & Hill, 2001), (Tanaka et al, 2009) et plus rapide (Besle et al, 2004), (Arnal et al, 2009), (van Wassenhove et al, 2005). Le gain de reconnaissance (intelligibilité) disparaît si nous remplaçons l'image des lèvres par des stimuli visuels de non-parole (par un cercle ou une barre simulant les mouvements labiaux) malgré le maintien de la corrélation temporelle entre les signaux acoustiques et visuels (Summerfield, 1979), (Schwartz et al, 2004). Par contre, le changement de dynamique temporelle, soit l'accélération soit le ralentissement, d'une des composantes résulte en une diminution des taux de reconnaissance (Munhall et al, 1996), (Tanaka et al, 2009).

1.2.3 Détection de la parole dans le bruit

La présence du signal visuel permet également de mieux détecter la présence de la parole. Grant et Seitz (Grant & Seitz, 2000) ont demandé à leurs sujets d'identifier si la phrase présentée dans la bruit est une phrase cible parmi trois cibles possibles : (1) « To make pure ice, you freeze water », (2) « Both brothers wear the same size », (3) « Watch the log float in the wide river ». Trois conditions étaient comparées: audio seul (A), audiovisuel cohérent (AVM) et audiovisuel incohérent (AVUM). La présence d'une entrée vidéo cohérente (AVM) améliore le seuil de détection jusqu'à 1.6 dB par rapport à une condition auditive seule tandis que la présence de vidéo incohérente (AVUM) ne change pas la performance auditive (Figure 6).

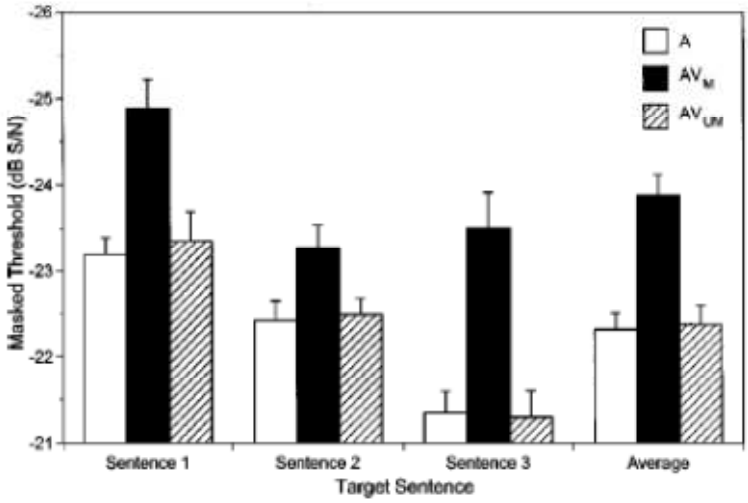


Figure 6 - Seuil de détection d'un signal de parole masqué par du bruit blanc en fonction des conditions expérimentales (A – audio seul, AVM – audio et vidéo cohérents, AVUM – audio et vidéo incohérents) et de la phrase cible. Figure tirée de (Grant & Seitz, 2000)

Pour expliquer la différence des résultats entre les phrases, notamment entre la phrase 3 et les phrases 1 et 2, Grant et Seitz (Grant & Seitz, 2000) ont fait l'hypothèse que l'effet visuel est dépendant de la corrélation entre les entrées auditive et visuelle. Il existe des relations

entre les mesures d'ouverture des lèvres et la modulation d'amplitude acoustique avec des variations des degrés de corrélation en fonction des phrases. Les auteurs ont ainsi mesuré la corrélation entre les mouvements des lèvres (largeur et hauteur d'ouverture de la bouche) et l'amplitude de l'enveloppe spectrale dans chacune des 3 régions formantiques F1 (100–800 Hz), F2 (800–2200 Hz), and F3 (2200–6500 Hz) ainsi qu'avec une large bande spectrale (100–6500 Hz). Les mesures, présentées sur la Figure 7 (phrase 2 à gauche et phrase 3 à droite), montrent une différence de corrélation audiovisuelle entre les phrases 2 et 3, et également que l'information présentée dans les bandes spectrales F2 et F3 fournit une bonne corrélation avec les variations de la cinématique labiale (Figure 8).

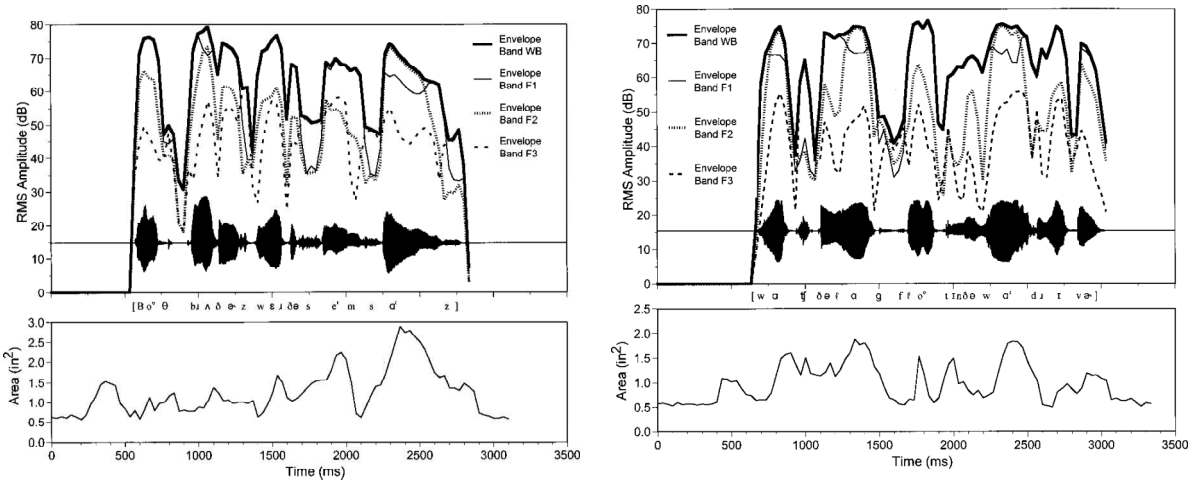


Figure 7 – Variations de la moyenne quadratique (RMS) de l'amplitude du signal acoustique et de l'aire de l'ouverture labiale pour la phrase 2 « Both brothers wear the same size » à gauche et la phrase 3 « Watch the log float in the wide river » à droite. WB- bande d'enveloppe ~100-6500 Hz, F1 ~100-800 Hz, F2 ~800-2200 Hz, F3 ~2200-6500 Hz

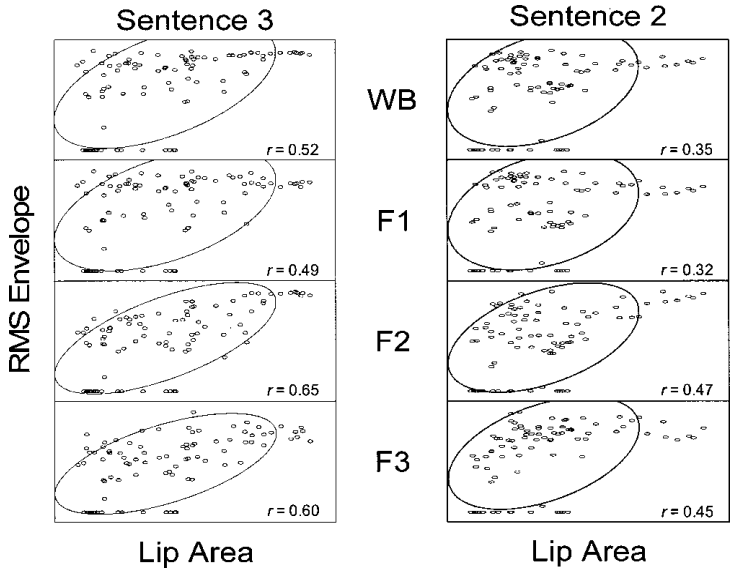


Figure 8 – Diagramme de dispersion montrant le taux de corrélation entre l'aire d'ouverture labiale et la moyenne quadratique (RMS) des fonctions d'enveloppe pour les phrases 3 et 2.

Kim et Davis (Kim & Davis, 2004)] ont aussi précisé que la facilitation de la détection n'a lieu que si la corrélation temporelle est forte entre les signaux auditifs et visuels. Dans leurs expériences l'effet d'amélioration de détection disparaît dans le cas d'un signal auditif inversé ou accéléré par rapport à l'image visuelle. Mais, par contre, un ralentissement du signal auditif provoque une légère amélioration de détection. Ces effets d'influence de la corrélation audiovisuelle résistent au passage à une langue étrangère (Kim & Davis, 2003), ils sont donc produits par les contenus acoustiques et visuels eux-mêmes, indépendamment d'une phase de compréhension du message.

1.2.4 Prédiction du son par l'image

Dans l'environnement écologique le signal visuel arrive en général plus tôt que le signal auditif. La raison est à chercher dans les processus de préparation du geste, qui sont visibles avant que le geste ne produise son effet audible. Ainsi, le mouvement d'un marteau se voit avant que le marteau ne frappe le clou et donc avant l'émission d'un son. De même, les processus de coarticulation produisent des phénomènes d'anticipation de la mâchoire et des lèvres qui souvent se voient avant de s'entendre (voir par exemple (Chandrasekaran et al, 2009)). Notre système de perception est calibré par rapport à cette différence, ainsi que la montre l'asymétrie de la fenêtre d'intégration temporelle dans la fusion de la parole audiovisuelle. Ainsi, dans l'expérience de Grant et Greenberg (Grant & Greenberg, 2001) un décalage temporel entre audition (deux bandes de fréquence 1/3 d'octave respectivement à 298-375 Hz et 4762-6000 Hz) et vision montre une décroissance forte et continue d'intelligibilité si le signal acoustique est en avance sur le signal vidéo de plus de 40 ms (Figure 9 en bleu) et une stabilité entre 40 et 200 ms pour une avance du signal vidéo par rapport à l'audio (Figure 9 en rouge). Cette stabilité est maintenue au niveau de la situation de synchronisation parfaite (intelligibilité audiovisuelle à 63%).

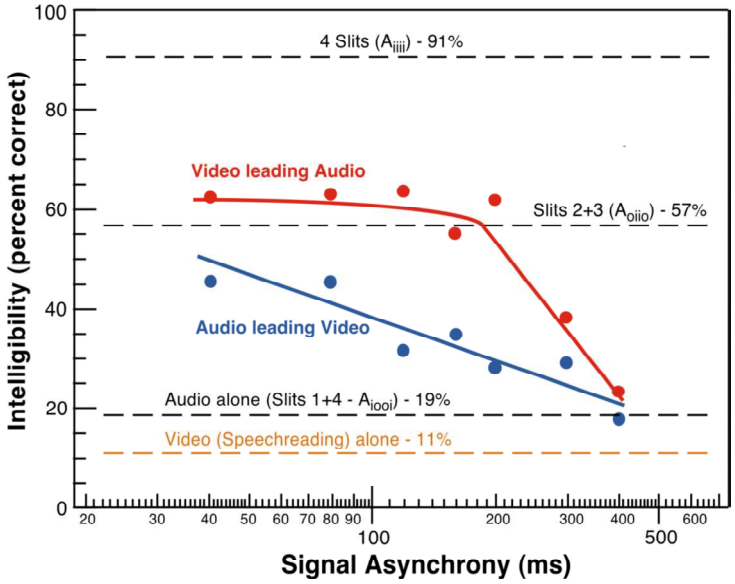


Figure 9 – Intelligibilité moyenne en fonction du décalage audiovisuel. Signal audio en avance sur le signal vidéo en bleu, signal vidéo en avance sur le signal audio en rouge ; la condition entrée audio seule, correspondant aux bandes de fréquences 298-375 Hz et 4762-6000 Hz (ligne pointillée noire du bas) est à 19% ; la condition entrée vidéo seule (ligne pointillée orange) est à 11%. Figure tirée de (Grant & Greenberg, 2001).

Le système peut également utiliser cette propriété écologique pour prédire le son émis à partir de l'entrée visuelle (Arnal et al, 2009), (Jiang et al, 2002).

En résumé, l'entrée visuelle fournit à la fois des possibilités d'amélioration de la détection, de la reconnaissance et de la prédiction du contenu acoustique de la parole.

1.3 L'effet McGurk

Nous avons déjà présenté l'expérience princeps de Harry McGurk et John MacDonald qui a donné lieu à la découverte de « l'effet McGurk » (McGurk & MacDonald, 1976). Dans leur publication d'origine impliquant un montage de stimuli « ba » avec « ga », et « pa » avec « ka », la majorité des participants ont répondu « da » et « ta » pour les cas A « ba » + V « ga » et A « pa » + V « ka », tandis que dans les combinaisons inverses A « ga » + V « ba » et A « ka » + V « pa » la majorité des réponses correspondent à l'entrée auditive, mais aussi à des combinaisons « gabga », « bagba », « baga », « gaba » et « kapka », « pakpa », « paka », « kapa ». Depuis cette publication initiale, l'effet McGurk est largement utilisé comme une mise en évidence de la fusion audiovisuelle.

1.3.1 Universalité vs. variations à travers les âges, les langues et les sujets

De nombreuses études ont porté sur ce paradigme expérimental. L'effet McGurk a été testé dans de nombreuses langues différentes. Il apparaît dans tous les langues où il a été testé: espagnol, allemand (Duran, 1995), italien (Bovo et al, 2009), néerlandais, chinois (Gelder et al, 1995), japonais (Sekiyama & Tohkura, 1993), hongrois (Grassegger, 1995), français (Cathiard et al, 2001) et autres. La diminution significative de perception de taux d'effet McGurk a été décrite pour les langues japonaise et chinoise en comparaison avec l'anglais, (Sekiyama & Tohkura, 1991), (Sekiyama, 1997), (Hisanaga et al, 2009). Pour expliquer ce phénomène, deux explications principales ont était proposées. La première repose sur l'influence supposée de la culture japonaise et chinoise, qui prescrit d'éviter de regarder son interlocuteur en face, ce qui rendrait les sujets moins sensibles à l'influence visuelle. La deuxième hypothèse est linguistique et non culturelle. Elle s'appuie sur la différence de la structure tonale et syllabique entre ces deux langues et la langue de référence de l'effet McGurk, l'anglais, ainsi que sur l'absence de clusters consonantiques dans les langues asiatiques. Les deux hypothèses restent considérées comme pertinentes à l'heure actuelle.

La sensibilité des enfants prélinguistiques à l'effet McGurk a été décrite dans le travail de Burnham et Dodd (Burnham & Dodd, 2004). Un groupe d'enfants de dix semaines, « habitué » à des stimuli McGurk « ba »A + « ga »V, ne manifestait pas de surprise à un stimulus audiovisuel cohérent « da » ou « tha », mettant ainsi en évidence une capacité d'intégration de type « McGurk » dès cet âge. Un groupe contrôle d'enfants « habitués » à des stimuli congruents « ba » ne présentait pas ce même effet, ce qui a validé la conclusion des auteurs de l'existence d'un mécanisme d'intégration de la parole audiovisuelle chez les enfants prélinguistiques. Des études ultérieures ont permis de mettre en évidence une augmentation de l'effet McGurk avec l'âge (Sekiyama & Burnham, 2008).

Enfin, il est important de considérer que l'effet McGurk dépend du sujet (Schwartz, 2010), avec de fortes différences interindividuelles, certains sujets manifestant un effet McGurk significatif et d'autre un effet faible ou nul. De nombreuses recherches actuelles visent à

chercher des corrélats neurocognitifs de ces différences, portant notamment sur les caractéristiques de la fenêtre d'intégration audiovisuelle qui prédirait la susceptibilité aux illusions audiovisuelles (Stevenson et al, 2012).

1.3.2 Variations dépendant des caractéristiques des stimuli

Des mécanismes de fusion audiovisuelle apparentés à l'effet McGurk ont été démontrés dans un contexte syllabique CV ou VCV, mais également, avec des effets plus ou moins forts, dans le cas de voyelles (Summerfield & McGrath, 1984), (Lisker & Rossi, 1992), dans des mots (Dekle et al, 1992), des phrases (McGurk, 1981) et même sur des stimuli non directement phonologiques comme des clicks qui sont reconnus comme des consonnes dans certaines langues africaines, mais qui sont considérés comme des événements non phonétiques pour des sujets anglais (Brancazio et al, 2006).

L'effet McGurk résiste à des incohérences audiovisuelles variées, telles que des discordances sur le sexe du locuteur entre visage et voix (Green et al, 1991) ou des différences de localisation spatiale entre le visage et la voix (Jones & Munhall, 1997), (Bertelson et al, 1994), (Colin et al, 2001).

Par contre, l'effet McGurk dépend du décalage temporel entre les signaux auditif et visuel selon une fenêtre d'intégration audiovisuelle qui présente la même asymétrie et les mêmes caractéristiques que celle présentées précédemment (Munhall et al, 1996), (van Wassenhove et al, 2007) (voir Figure 10, montrant que l'effet McGurk est obtenu sur une gamme de délais allant de faibles avances de l'audio à des fortes avances du vidéo, comme pour les scores de compréhension dans le bruit présentés précédemment).

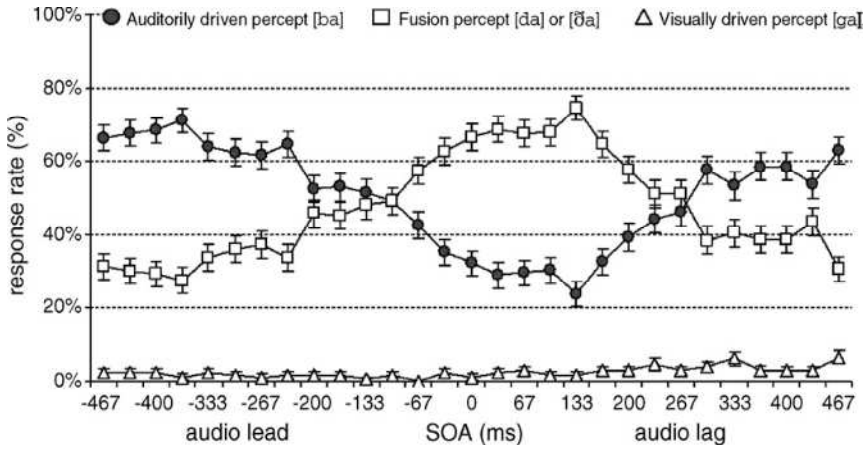


Figure 10 – Taux de réponses en fonction du délai entre les stimuli auditifs et visuels (SOA, stimulus onset asynchronies) en ms pour les stimuli McGurk A_bV_g. Figure tirée de (van Wassenhove et al, 2007)

L'effet McGurk dépend aussi du débit d'articulation (Colin & Radeau, 2003), (Munhall et al, 1996), ainsi l'effet McGurk est plus élevé dans le cas d'augmentation du débit auditif ou de ralentissement du débit visuel. Ceci peut-être expliqué par le fait que le ralentissement du débit vidéo laisse plus de temps pour lire sur les lèvres, et l'accélération du débit audio réduit l'intelligibilité du signal, donnant ainsi plus de poids au signal visuel.

Une augmentation de perception d'effet McGurk est également observée dans le bruit auditif (Sekiyama & Tohkura, 1991), (Colin et al, 2004) et une diminution de l'effet dans le cas d'une dégradation visuelle sous forme de réduction de la résolution d'image par un moyennage local des pixels (MacDonald et al, 1999).

1.3.3 Automaticité de l'effet McGurk

La vision classique considère une automaticité de la fusion dans le cas de l'effet McGurk en affirmant la nature préattentive du processus de fusion sans possibilité d'un contrôle volontaire (Massaro, 1987). Néanmoins, des études récentes ont mis en cause cette hypothèse d'automaticité et démontré des effets attentionnels. Ainsi, (Tiippana et al, 2004) ont montré une diminution significative de l'effet McGurk si l'attention du sujet est orientée vers un distracteur visuel tel qu'une feuille d'arbre qui glisse sur le visage du locuteur sans masquer les lèvres. Les travaux d'Alsius et al. (Alsius et al, 2005), (Alsius et al, 2007) montrent par ailleurs que l'effet McGurk décroît si l'on « charge » le système attentionnel par un paradigme de double tâche, que la tâche secondaire soit auditive, visuelle ou tactile.

Une autre mise en évidence de l'existence de mécanismes attentionnels concerne les résultats montrant la sensibilité des sujets aux composantes individuels de l'effet McGurk, démontrée par Soto-Faraco (Soto-Faraco & Alsius, 2009). Cette expérience montre en effet que les sujets sont capables de déterminer une éventuelle désynchronisation des deux modalités, et donc de « déconstruire » l'effet McGurk dans un paradigme adéquat.

1.4 Conclusion

La multimodalité de la parole apparaît donc comme un élément constitutif de la perception de la parole, imprégnant de nombreux aspects des mécanismes perceptifs, et affectant largement les signaux dans de nombreux types de paradigmes, et ce pour des sujets de tous types de capacités perceptives (normo-entendants ou malentendants), de toutes langues et de tous âges. Nous allons maintenant nous pencher sur les mécanismes cognitifs sous-tendant ces effets d'interaction multiensorielle en perception de parole.

Chapitre 2. Processus de fusion audiovisuelle en perception de parole

2.1 Les architectures cognitives

Dans le chapitre précédent nous avons montré que la parole n'est pas purement auditive et qu'il y a un effet de la modalité visuelle sur la perception de la parole. La première question qui se pose est celle de l'architecture cognitive permettant la combinaison/intégration/fusion de ces deux sources d'information. De nombreux travaux ont abordé cette question dans la littérature. Schwartz et collègues (Schwartz et al, 1998) ont résumé les approches existantes dans 4 architectures possibles (Figure 11).

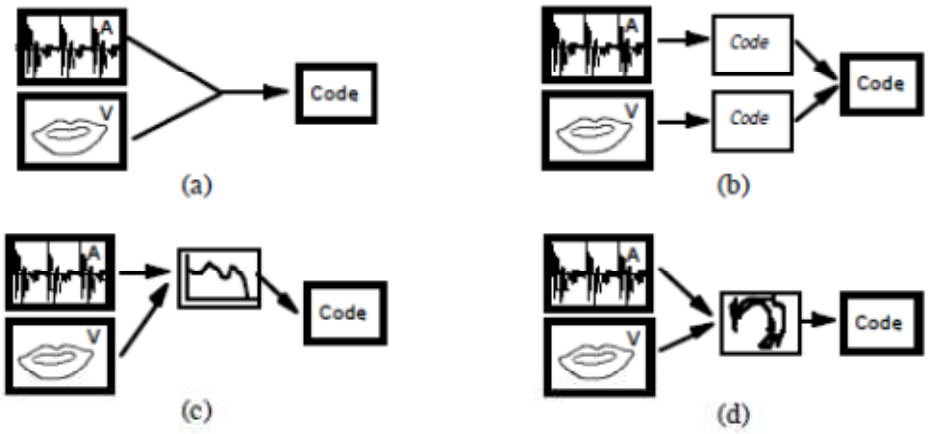


Figure 11 - Taxonomie des architectures de fusion audiovisuelle, d'après (Schwartz et al, 1998) (a) Identification directe, (b) Identification séparée, (c) Recodage dans la modalité dominante, (d) Recodage moteur

2.1.1 Modèle d'identification directe

Le premier type d'architecture suppose que les signaux auditifs et visuels peuvent se fusionner directement (Figure 11 (a)), c'est le « Modèle d'identification directe ». Ce modèle est issu du modèle "Lexical Access from Spectra" de (Klatt, 1979), et il suppose qu'un classifieur bimodal peut traiter directement la combinaison des signaux auditifs et visuels. Dans les discussions actuelles ce modèle est considéré comme peu probable en l'état. Son défaut principal est l'absence de représentation cognitive commune du son et de l'image, tandis que certaines données expérimentales montrent sa nécessité. Par exemple, l'argument fort en faveur d'une représentation commune est que les bébés pré langagiers savent détecter l'incohérence de la parole audiovisuelle dès 4 mois (Kuhl & Meltzoff, 1982), (Kuhl & Meltzoff, 1984). Ils savent associer un son et un visage prononçant un signal cohérent (son de « a » apparié au visage prononçant « a » de préférence au visage prononçant « i »). Ceci indique qu'il existe un processus de détection de congruence entre les deux modalités. Toutes les autres architectures sont basées sur l'existence d'une représentation cognitive commune du son et de l'image du locuteur.

2.1.2 Modèle d'identification séparée et modèles de fusion bayésienne

Le deuxième type est un « modèle d'identification séparée », qui suppose un recodage préalable de chaque modalité avant l'étape de fusion (Figure 11 (b)). Le recodage peut se faire sous forme de valeurs logiques, comme c'est le cas dans le modèle VPAM (Vision:Place Audition:Manner) (McGurk & MacDonald, 1976), (Summerfield, 1987), ou sous forme probabiliste ou équivalente, comme c'est le cas dans le modèle FLMP (Fuzzy-Logical Model of Perception) (Massaro, 1987), (Massaro, 1989). Dans ce type de modèles, l'intégration des inputs auditif et visuel se passe donc à un niveau postérieur à la catégorisation phonétique, les modèles correspondants sont donc baptisés de modèle à intégration tardive.

Dans le modèle VPAM chaque modalité prend en charge son propre ensemble de caractéristiques phonétiques. A partir du signal vidéo on extrait le lieu d'articulation tel que vélaire, bilabial etc., et à partir du signal auditif on extrait le mode tel que consonantique, nasal, etc. Ces informations sont fournies au processus de fusion pour la catégorisation. La critique majeure du modèle VPAM est la répartition stricte des rôles du signal auditif et visuel, tandis que certaines caractéristiques devraient être estimées à la fois visuellement et auditivement, à la fois pour le lieu et le mode. Ainsi, le modèle VPAM ne peut rendre compte de l'effet McGurk (« ba-audio » + « ga-vidéo » donne une fusion « da » ou « tha »), pour lequel ni le lieu (dans le cas du percept « da ») ni même le mode (dans le cas du percept « tha ») ne correspondent aux données susceptibles d'être fournies par l'audition (mode plosif) et la vision (lieu vélaire).

Cette faille est prise en compte dans le modèle FLMP, où chaque entrée est comparée analogiquement à un prototype unimodal. Les résultats de cette évaluation monosensorielle sont fusionnés par un processus multiplicatif normalisé :

$$P_{av}(C_i) = \frac{a_i v_i}{\sum_j a_j v_j}$$

où *a* est le taux de réponses en faveur de la catégorie C_i fourni par l'entrée auditive, *v* est le taux de réponses fourni par l'entrée visuelle, et P_{av} le taux de réponses en faveur de la catégorie C_i estimé en sortie du processus de fusion audiovisuelle (Figure 12).

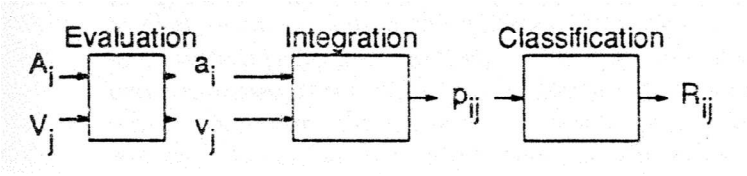


Figure 12 – Modèle FLMP de fusion audiovisuelle en perception de la parole. L'évaluation d'une source auditive Aj et visuelle Vj produit des valeurs aj et vj indiquant le degré de support de chaque source pour une catégorie donnée. Le résultat final repose sur le produit de ces degrés de support. Figure tirée de (Massaro, 1989).

Le modèle FLMP s'est avéré très populaire par sa simplicité et la possibilité de l'appliquer dans de nombreuses situations expérimentales avec des résultats de prédictions le plus souvent proches des observations empiriques. C'est en réalité un modèle général qui

permet d'effectuer la fusion de modalités différentes, pas nécessairement audiovisuelles. Ce modèle peut s'intégrer dans la catégorie plus générale encore des modèles de fusion bayésienne, s'appuyant sur une estimation de l'efficacité de chaque modalité par la théorie de l'intégration basée sur le maximum de vraisemblance (Maximum likelyhood integration theory) pour déterminer le poids de chaque modalité dans le processus de fusion intersensorielle, par exemple dans les expériences de perception visuo-haptique (Ernst & Banks, 2002), d'intégration visuo–vestibulaire (Angelaki et al, 2011), de détermination de la localisation des stimuli audiovisuels (Alais & Burr, 2004).

Néanmoins, l'hypothèse d'un processus de fusion audiovisuelle basée uniquement sur la sortie des processus de fusion auditive et visuelle sans prise en compte de facteurs de contrôle du processus de fusion est contestable. Ainsi, les données décrites précédemment sur la phénoménologie de l'effet McGurk ont conduit – ou pourraient conduire - à des variantes du FLMP intégrant une pondération des entrées du processus multiplicatif pour tenir compte de ces différents facteurs tels que la variabilité des sujets (Schwartz, 2010), le niveau de bruit (Berthommier, 2001), la langue (Sekiyama & Tohkura, 1991), (Sekiyama & Tohkura, 1993), ou l'attention (Schwartz et al, 2010).

2.1.3 Modèle de recodage dans la modalité dominante

La troisième catégorie, du « modèle à recodage dans la modalité dominante » (Figure 11 (c)), considère la modalité auditive comme une modalité dominante et plus adaptée à la parole. L'hypothèse est alors que la modalité visuelle doit être recodée sous un format imposé parla modalité auditive avant fusion. Le recodage pourrait se faire dans un espace spectro-temporel caractérisé par la forme de la fonction du filtre (ou fonction de transfert) du tractus vocal (Summerfield, 1987), (Summerfield & McGrath, 1984). Le cadre proposé est celui de la production de la parole qui détermine la parole comme une source d'énergie acoustique (produite par les poumons, et passant par le larynx) filtrée par les mouvements articulatoires, d'où le nom de fonction de filtre. Dans ce modèle, comme dans le suivant, l'intégration des inputs auditif et visuel se passe à un niveau préalable à la catégorisation phonétique et lexicale (Figure 13). Ces modèles sont donc dénommés, à ce titre, modèles d'intégration précoce.

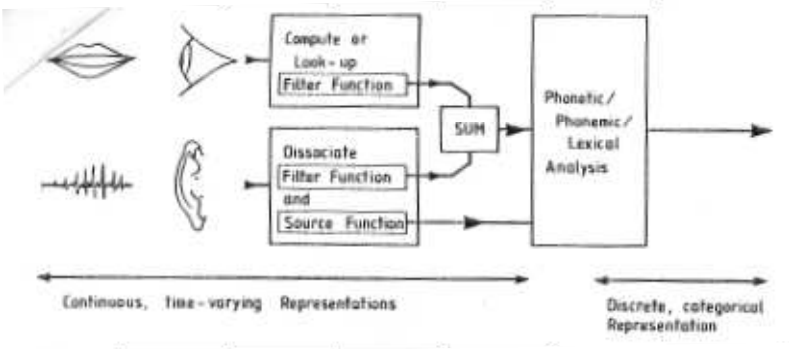


Figure 13 – intégration audiovisuelle basée sur la fonction de filtre du conduit vocal, dans le cadre des architectures de « Recodage dans la modalité dominante ». Figure tirée de (Summerfield, 1987).

2.1.4 Modèle de recodage dans la modalité motrice, théories motrices et perceptuo-motrices

La dernière architecture, le « modèle à recodage moteur », s'inspire de la théorie motrice (Liberman & Mattingly, 1985) et suppose le recodage des deux modalités dans un format ni auditif ni visuel, mais amodal. Elle est également une architecture à intégration précoce.

Dans la version classique de la théorie motrice, Liberman et Mattingly (Liberman & Mattingly, 1985) défendent l'idée d'un traitement de la parole fondé sur un processus d'analyse-par-la-synthèse. Les auteurs considèrent que la perception et la production de la parole sont deux faces d'un même processus. Dans la théorie motrice, la représentation qui est impliquée à la fois dans la perception et dans la production est basée sur la configuration du conduit vocal et les processus moteurs sous-jacents, ce qui peut être résumé sous le terme de « format moteur ». Or, les gestes de la parole sont également visibles, donc la modalité visuelle peut également être recodée dans un format moteur. Dans une version ultérieure de leur théorie, Liberman et Mattingly proposent que le recodage prenne en réalité la forme d'une représentation des intentions motrices, plutôt que du geste articulatoire lui-même. Ces intentions motrices seraient formées spécifiquement dans le cerveau du locuteur et un module spécialisé permettrait à l'auditeur de reconstruire cette représentation motrice avec un minimum d'effort. L'information recodée serait la base de la catégorisation. Liberman et Mattingly défendent l'idée que ce processus de traitement est spécifique à la parole, tandis que dans le cas de signaux acoustiques non langagiers la perception serait directe et non médiatisée par des processus de recodage moteur.

Cependant, une collègue de Liberman et Mattingly au sein des Laboratoires Haskins, Carol Fowler, a proposé quant à elle une théorie réaliste directe (Fowler, 1986). Elle pense comme Liberman et Mattingly qu'il existe un processus de recodage moteur, mais elle conteste l'idée que ce processus serait spécifique à la parole. Sa conception est qu'un auditeur récupère la cause physique de façon générale, quel que soit le processus (langagier ou non) et la modalité perceptive impliquée. Dans le cas de la parole la cause du signal peut-être la configuration articulatoire, dans le cas des autres signaux on récupère la forme de l'objet.

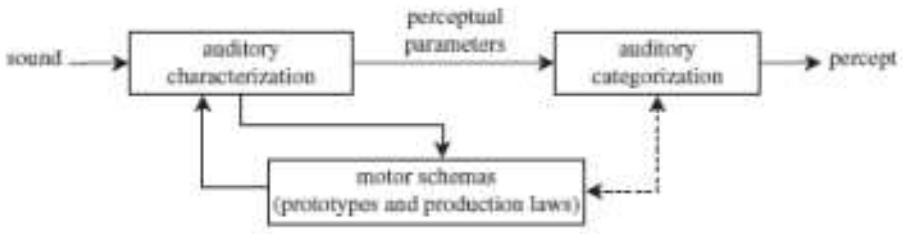


Figure 14–L'architecture générale de la PACT

Plus récemment, Schwartz et collègues (Schwartz et al, 2010) ont proposé une théorie perceptuo-motrice, PACT (« Perception for Action Control Theory »). Ils intègrent la nécessité de considérer l'existence d'interactions entre perception et action. Mais ils mettent également en avant l'existence de processus perceptifs pour caractériser les gestes, ainsi que le

démontrent notamment les mécanismes sous-jacents à l'organisation des systèmes sonores des langues du monde, avec des processus de détermination des frontières entre catégories phonétiques acoustiquement réglés sur la base de la dispersion perceptive ou des invariances sensori-motrices naturelles. Pour Schwartz et coll. la perception forme l'action et l'action met des contraintes sur la perception. Par rapport à la théorie motrice classique, la PACT insiste sur le fait que l'étape de catégorisation/décision doit en premier lieu prendre en compte les caractéristiques auditives et pas simplement les configurations articulatoires. Les percepts acoustiques sont façonnés par les connaissances articulatoires et l'unité de communication est une unité perceptuo-motrice. A la base cette théorie est issue d'une réflexion sur la modalité auditive, mais elle a également pris en compte des recherches du groupe sur le rôle de la modalité visuelle (Sato et al, 2007), (Sato et al, 2007), (Basirat et al, 2012).Le schéma général de la PACT est représenté sur la Figure 14.

2.2 Les processus de contrôle

Indépendamment des réflexions sur les architectures cognitives telles que présentées précédemment, se pose la question du mécanisme de fusion proprement dit. Bloch (Bloch, 1996) a proposé une taxonomie de ces mécanismes, organisée autour de la question du contexte général du processus de fusion.

- Processus Indépendant du Contexte et à Comportement Constant. La fusion se passe de manière fixe, c'est-à-dire que le résultat dépend d'un calcul automatique et de type fixé, opérant sur les sorties des processus à fusionner, par exemple par une loi additive ou multiplicative. C'est typiquement le mode opératoire du modèle FLMP.
- Processus Indépendant du Contexte et à Comportement Variable. La fusion se passe de manière variable, par exemple la fusion s'effectue par loi additive mais avec des coefficients de pondération variables, selon les valeurs d'entrée.
- Processus dépendant du contexte. La fusion prend en compte le contexte d'environnement extérieur, par exemple le niveau du bruit, pour pondérer l'entrée. C'est typiquement dans ce cadre qu'ont été réalisées un certain nombre d'adaptations du modèle FLMP, prenant en compte par exemple le niveau de bruit (Berthommier, 2001), le sujet (Schwartz, 2010) ou l'état attentionnel (Schwartz et al, 2010).

2.3 Les architectures neuroanatomiques sous-jacentes

2.3.1 Le modèle classique de Wernicke-Lichtheim-Geshwind

Les premières descriptions neuro-anatomiques de la parole viennent de la parole auditive. En 1861, Pierre Paul Broca décrit une perte de la capacité à produire de la parole chez son célèbre patient aphasique Leborgne, après un traumatisme dans la zone du gyrus frontal inférieur. La localisation précise post-mortem de la région détruite par le traumatisme a conduit à repérer cette zone qui porte dorénavant le nom « d'aire de Broca » et se situe dans les aires 44 et 45 de l'atlas de Brodmann (Figure 15). Depuis, l'aire de Broca est considéré comme jouant un rôle majeur dans les processus de production de la parole, notamment dans les aspects moteurs du langage.

Une dizaine d'années plus tard, Carl Wernicke a décrit une zone dans le gyrus temporal supérieur, qu'il a considérée être le lien focal potentiel d'une aphasie perceptive dénommée depuis aphasie de Wernicke. Les difficultés principales associées à cette zone sont des troubles de la compréhension du langage sous sa forme orale et écrite. Les personnes souffrant d'une aphasie de Wernicke perdent l'essentiel de leurs capacités de compréhension de la parole. Ils conservent une capacité à parler de façon naturelle avec un rythme et une syntaxe normale, mais le flux de parole reste incompréhensible. Cette zone, dénommée classiquement « aire de Wernicke », se situe dans la partie postérieure du lobe temporal à proximité du cortex auditif primaire, et correspond à la région 22 de l'atlas de Brodmann (Figure 15). L'aire de Wernicke est connectée avec l'aire de Broca par le faisceau arqué. La vision moderne considère que les fonctions de compréhension du langage précédemment attribuées à la seule aire de Wernicke sont plus largement distribuées dans le lobe temporal.

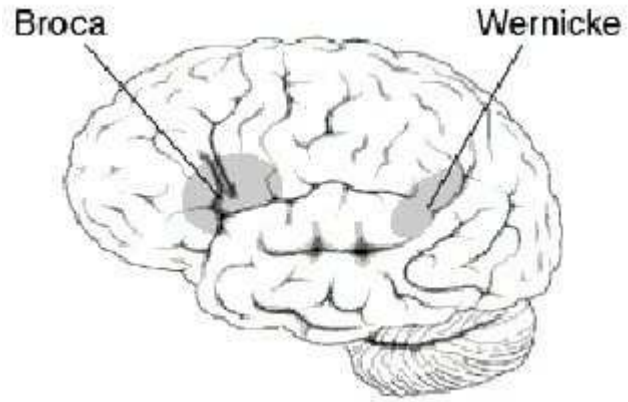


Figure 15 - L'aire de Broca et l'aire de Wernicke

Carl Wernicke a proposé l'un des premiers modèles neurologiques du langage, qui a été ensuite complété par Geshwind et Lichtheim. Selon ce modèle le chemin perceptif passe à partir des aires primaires auditives vers l'aire de Wernicke, qui est responsable de la compréhension de la parole. Pour produire de la parole l'information est envoyée par le faisceau arqué vers l'aire de Broca, où sont stockées les représentations articulatoires (Figure 16). Puis les instructions motrices sont envoyées de l'aire de Broca vers l'aire de la face dans le cortex moteur et vers les neurones moteurs de la face dans le tronc cérébral pour produire l'articulation orofaciale.

Le schéma expliqué concerne la perception auditive de la parole. Pour la tâche de lecture des mots ces deux aires ne sont pas suffisantes. Donc Lichtheim (Lichtheim, 1885) propose qu'une troisième aire, qu'il a appelée centre des concepts, réalise une association des représentations mentales des objets avec les mots. Geschwind (Geschwind, 1972) a proposé que cette zone se trouve dans le gyrus angulaire. Dans ce modèle l'information visuelle des mots passe des aires visuelles vers le gyrus angulaire (centre des concepts) pour la compréhension, et puis vers l'aire de Broca pour la production.

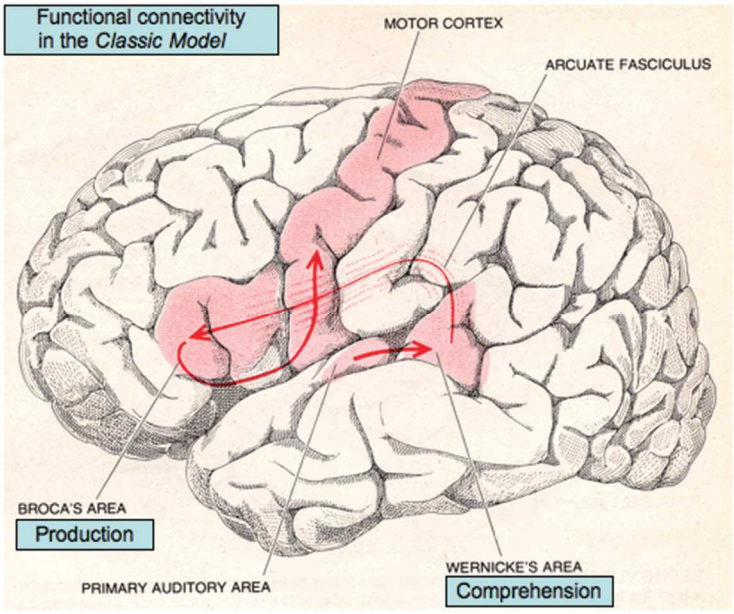


Figure 16 - modèle de Wernicke-Lichtheim-Geschwind. Figure tirée de (Shalom & Poeppel, 2008).

Ces premiers modèles, historiquement cruciaux dans l'histoire de la neurologie, proposent ainsi une vision très localisationniste des architectures corticales. Les évolutions des 20 dernières années, rendues possibles par le développement massif des techniques de neuroimagerie, ont conduit à des architectures beaucoup plus complexes et imbriquées, au sein desquelles les relations perceptuo-motrices jouent un rôle central, comme nous allons le voir maintenant.

2.3.2 Du système miroir au modèle à deux voies

Longtemps après l'introduction par Liberman et coll. de la théorie motrice, la découverte des neurones miroirs (Rizzolatti et al, 2001), (Kohler et al, 2002) lui a donné un renfort expérimental inattendu et inespéré. Le terme de neurone miroir désigne des neurones des aires frontales inférieures et des aires pariétales (classiquement impliquées dans la production d'actions), qui s'activent non seulement pendant la réalisation d'actions mais aussi pendant l'observation d'actions du même type. Ces neurones ont été observés par l'équipe de Parme chez les singes (Rizzolatti et al, 1996), (Rizzolatti & Craighero, 2004).

Des résultats de même type ont été obtenus chez l'humain par des techniques évidemment différentes, la neuroimagerie remplaçant les enregistrements cellulaires. Ainsi, de nombreuses études de neuroimagerie fonctionnelle ont montré l'activation de régions frontales et pariétales, à la fois dans des tâches motrices et perceptives, donnant lieu à l'introduction du concept de « système miroir » (Rizzolatti et al, 1996), (Iacoboni et al, 1999). Dans le cas de la parole, de nombreuses études montrent l'implication des zones motrices du cortex dans des tâches de perception auditive (Wilson et al, 2004), (Wilson & Iacoboni, 2006), (Pulvermüller et al, 2006).

C'est dans ce contexte que Hickok et Poeppel (Hickok & Poeppel, 2004), (Hickok & Poeppel, 2007) proposent un modèle qui sépare le processus de traitement de la parole en deux flux, inspirés d'une littérature classique sur le traitement visuel : une voie ventrale et une voie

dorsale (Figure 17). Dans ce modèle il est proposé que le traitement acoustique de la parole commence par une analyse spectro-temporelle dans les zones auditives du gyrus temporal supérieur (STG) de façon bilatérale. Puis le traitement se séparerait dans les deux voies, la voie ventrale étant essentiellement chargée des processus de compréhension de la parole et la voie dorsale de la mise en correspondance entre le signal acoustique et les représentations articulatoires/motrices.

La voie ventrale part du gyrus temporal supérieur (STG) et se poursuit vers le cortex temporal inférieur par des connexions ventro-latérales. Elle assurerait le passage des représentations phonologiques vers des représentations lexiques conceptuelles. La voie dorsale implique les zones des connexions à la frontière des lobes temporal et pariétal, qui servent comme une interface sensori-motrice, puis se poursuit vers les zones frontales motrices. Son rôle est d'associer représentations perceptives, phonologiques et articulatoires. La voie ventrale serait essentiellement bilatérale, ce qui peut expliquer l'absence de difficultés de compréhension de la parole dans le cas d'endommagement unilatéral d'un lobe temporal. Par contre la voie dorsale aurait une forte dominance gauche. Ceci peut expliquer les déficits de production issus des lésions dorsales des zones temporales et frontales, des déficits de l'hémisphère gauche provoquant une considérable diminution de performance dans les tâches de discrimination sous-lexicale.

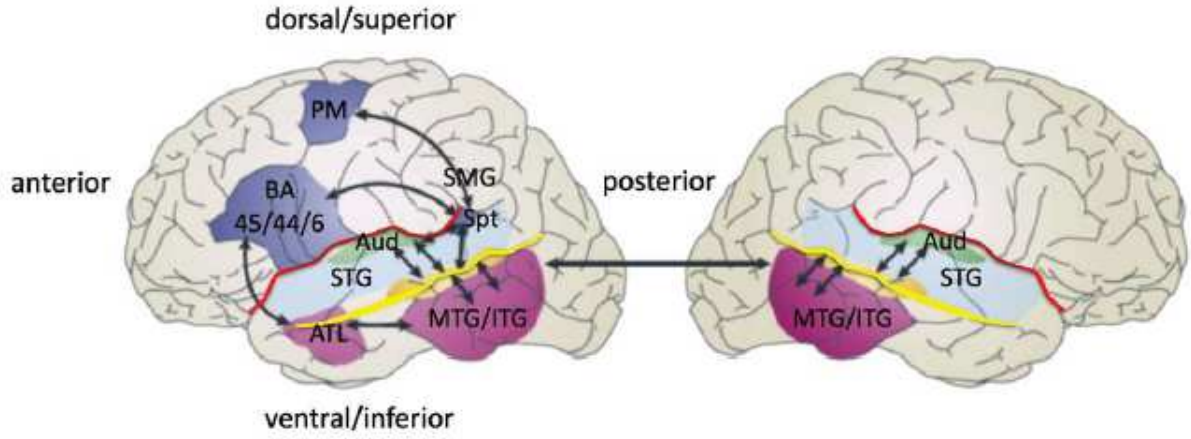


Figure 17 - Modèle proposé par Hickok et Poeppel pour le traitement cérébral de la parole. (Figure tirée de (Hickok, 2009)). En bleu – la voie dorsale, avec une latéralisation gauche, la zone Spt (région Sylvienne pariéto-temporale) est proposée être une interface sensori-motrice, tandis que les zones frontales correspondent à un réseau articulatoire. En rose – la voie ventrale, organisée de façon bilatérale, proposée être une interface lexicale qui lie l'information phonologique et sémantique. En vert – la surface dorsale du gyrus temporal supérieur (STG), impliquée dans l'analyse spectro-temporelle. La couleur jaune marque la partie postérieure de STS (sillon temporal supérieur), qui est impliquée dans le traitement phonologique. ATL –lobe temporal antérieur; Aud – cortex auditif primaire ; BA 45/44/6 sont les aires de Brodmann 45, 44 et 6 ; MTG/ITG est le gyrus temporal moyen et inférieur ; PM – cortex pré-moteur ; SMG – gyrus supramarginal ; la ligne rouge est la scissure Sylvienne, la ligne jaune est le sillon temporal supérieur.

2.3.3 Le réseau neuroanatomique de la perception audiovisuelle de la parole

Les architectures abordées précédemment considèrent essentiellement la modalité auditive. Le modèle de Hickok et Poeppel amène un nouveau regard sur l'organisation corticale du traitement de la parole, qui peut-être divisée selon deux flux relativement séparés. Nous allons maintenant aborder les études neuroanatomiques portant sur la parole audiovisuelle, ce qui nous permettra de passer à des modèles multimodaux.

Avec les techniques d'EEG (Electro-encéphalographie) et de MEG (Magnéto-encéphalographie) nous pouvons analyser les interactions audiovisuelles grâce à des mesures de potentiel évoqué, qui désignent la modification du potentiel électrique produit par le système nerveux en réponse à une stimulation externe (un son, une image, une toucher etc.) ou un événement cognitif interne, tel que l'attention, la préparation motrice, etc. Les potentiels évoqués sont décrits par des mesures d'amplitude et de latence, ou selon des analyses temps-fréquence permettant des analyses de synchronisation et de phase selon diverses bandes de fréquence.

En ajoutant l'information visuelle dynamique du locuteur à l'information auditive de la parole, on observe une modulation des potentiels évoqués auditifs avec une diminution d'amplitude du pic de la réponse, accompagnée d'une diminution de sa latence (Besle et al, 2004), (Colin et al, 2002), (van Wassenhove et al, 2005). Cette modulation, sur laquelle nous reviendrons au chapitre suivant, arrive assez tôt, entre 100 et 200 ms après le début du son, ce qui démontre l'implication du signal visuel dès le stade initial du traitement. L'anticipation du signal visuel, mentionnée au chapitre précédent, permet de supposer d'intervention de mécanismes de prédiction de caractéristiques acoustiques par le signal visuel. On peut se demander si cet effet de prédictibilité est spécifique à la parole ou si c'est une propriété de tous les événements multimodaux. L'expérience de Stekelenburg et Vroomen (Stekelenburg & Vroomen, 2007) montre que le même type de modulation peut être obtenu également avec des stimuli non phonétiques, tels qu'un claquement des mains. Et dans le cas contraire, quand le signal visuel est non prédictible, sa présence ne produit pas de modulation du cortex auditif au stimulus sonore. Ainsi, cette expérience met en évidence les effets de la prédictibilité du signal visuel même pour des signaux non phonétiques.

Si l'électrophysiologie fournit une grande résolution temporelle, ses limitations techniques ne permettent pas de localiser précisément les activations cérébrales dans le cerveau. La situation est inverse avec l'imagerie par résonance magnétique fonctionnelle (IRMf). Nous allons maintenant décrire ce que les études IRMf nous ont appris sur la neuroanatomie des mécanismes d'intégration audiovisuelle.

De nombreuses études insistent sur le rôle spécifique de la partie postérieure du sillon temporal supérieur (posterior superior temporal sulcus, pSTS) dans l'intégration audiovisuelle en perception de la parole (Calvert et al, 2000), (Wright et al, 2003). Ainsi, Calvert et al. (Calvert et al, 2000) dans leur étude IRMf ont observé un effet supra-additif d'activation dans le pSTS pour des stimuli congruents (AV>A+V) et un effet sub-additif pour des stimuli incongruents (AV<A+V). Les auteurs suggèrent que le pSTS pourrait jouer un rôle crucial dans l'intégration d'information audiovisuelle. Cependant cette vision est critiquée par Hocking et Price (Hocking & Price, 2008), qui proposent que pSTS ne serait qu'une partie d'un réseau plus large impliqué dans la mise en correspondance perceptive.

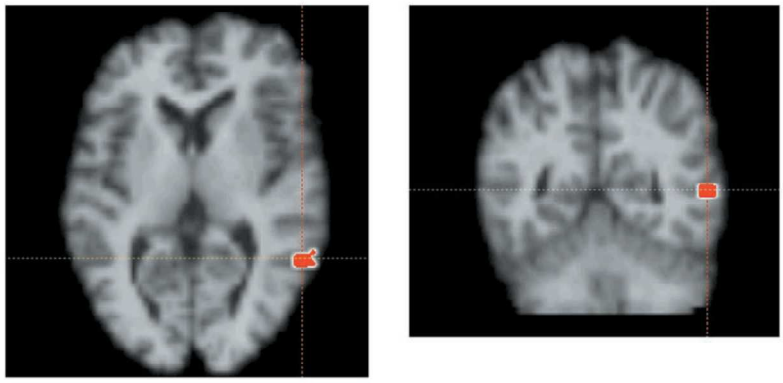


Figure 18 – Activation cérébrale dans la région du sillon temporal supérieur gauche qui est proposé comme le lieu d'intégration audiovisuelle de la parole. La partie gauche correspond à l'hémisphère droit selon la convention radiologique. Figure tirée de (Calvert et al, 2000).

Une autre observation importante de cette étude est le fait que l'activation dans le cortex auditif primaire et dans la zone occipitale du cortex, deux régions sensorielles primaires supposées jusqu'à cette époque être essentiellement cloisonnées et imperméables aux interactions multisensorielles, est augmentée en situation bimodale par rapport à la condition unimodale. Calvert et al. supposent que cette augmentation d'activité pourrait être produite par un retour de l'information concernant l'état d'intégration vers les zones du cortex primaire, en d'autres termes l'activité au sein de la région STS modulerait l'activité des cortex sensoriels correspondants. En étudiant l'intégration audiovisuelle chez les primates non humains, Ghazanfar et coll. (Ghazanfar et al, 2008) parviennent à la même conclusion, d'une rétropropagation d'information de STS vers le cortex auditif.

Cependant, Bernstein et al. (Bernstein et al, 2008) dans leur étude iRMf proposent le gyrus supramarginal (SMG) gauche, comme un lieu probable de fusion audiovisuelle. Leur paradigme aborde la question de l'intégration à partir du traitement de l'incongruence audiovisuelle. Trois types de stimuli avec différents degrés d'incongruence étaient présentés aux sujets : LI (Low Incongruity), MI (Medium Incongruity) et HI (High Incongruity). Les stimuli ayant une incongruité basse étaient cohérents : AbaVba et AlaVla. Les stimuli avec une incongruité moyenne étaient AbaVda, AlaVva, AbaVga et AlaVwa et les stimuli très incongruents étaient AbaVva, AlaVba, AbaVwa et AlaVda. Ces stimuli peuvent entraîner différents types de percepts : un percept cohérent avec la modalité auditive, cohérent avec la modalité visuelle, une combinaison des deux consonnes présentes respectivement dans la modalité auditive et visuelle (ex. /gi/ auditif + /bi/ visuel => /bgi/ audio-visuel) ou un percept fusion de type McGurk. La seule région qui était activée en proportion du degré d'incongruence de ces trois conditions était le gyrus supramarginal (SMG), dont les auteurs proposent qu'il serait impliqué dans l'analyse de la relation entre les entrées phonétiques auditives et visuelles, ainsi que dans la connaissance mémorisée d'une relation congruente entre patterns auditifs et visuels.

Une proposition alternative aux rôles de STS et SMG dans l'intégration audiovisuelle concerne l'intervention des interactions sensori-motrices. Ainsi, Skipper et al. (Skipper et al, 2007) supposent l'implication d'un système miroir dans l'intégration audiovisuelle. Par une étude IRMf ils ont observé une activation des régions motrices frontales impliquées dans la production de la parole, pendant la perception de syllabes McGurk (Vka avec Apa), et ils ont

pu montrer que cette activation est plus proche de l'activation induite par AVta, que de AVpa ou AVka. Par contre leur étude montre que l'activité pendant la perception d'effet McGurk dans les zones auditives et visuelles est similaire à l'activation pendant la perception des syllabes AVpa et AVka respectivement. L'interprétation est que l'activité des zones motrices pendant la perception correspond à une copie d'efférence du plan moteur inféré par le sujet percevant, et qui influence en retour l'interprétation phonétique.

D'autres études confirment l'implication du gyrus frontal inférieur, impliqué dans la production de la parole, lors de la perception audio-visuelle de la parole (Callan et al, 2003), (Ojanen et al, 2005). Ceci va nous conduire à la présentation d'un modèle cortical de fusion audiovisuelle, basée sur la théorie motrice.

2.3.4 Le modèle de Skipper

En partant de leurs propres résultats de neuroimagerie, Skipper et al. (Skipper et al, 2005), (Skipper et al, 2007) proposent un modèle neuroanatomique fonctionnel se basant sur l'hypothèse de l'existence de mécanismes d'analyse-par-synthèse au sein de la voie dorsale, mécanismes qui joueraient un rôle majeur dans le processus de perception audiovisuelle (Figure 19). Selon ce modèle les zones impliquées dans le traitement de la parole sont les aires visuelles, le cortex auditif primaire (A1), les aires temporales supérieures postérieures (STp), le gyrus supramarginal (SMG), le cortex somatosensoriel (SI/SII), le cortex ventral prémotor (PMv) et l'aire de Broca pars opercularis (POp). Le traitement commencerait dans les aires auditives et visuelles primaires, qui conduiraient vers les représentations multisensorielles sous forme d'hypothèses phonétiques dans les aires multisensorielles de STp. Ces hypothèses seraient formulées en termes d'intentions motrices sur le geste observé, dans l'aire de Broca(POp). Les intentions motrices conduiraient à des commandes motrices susceptibles de générer les gestes correspondants, selon un codage somatotopique approprié, dans les régions associées au contrôle orofacial dans les cortex prémoteur et moteur primaire PMv et M1.

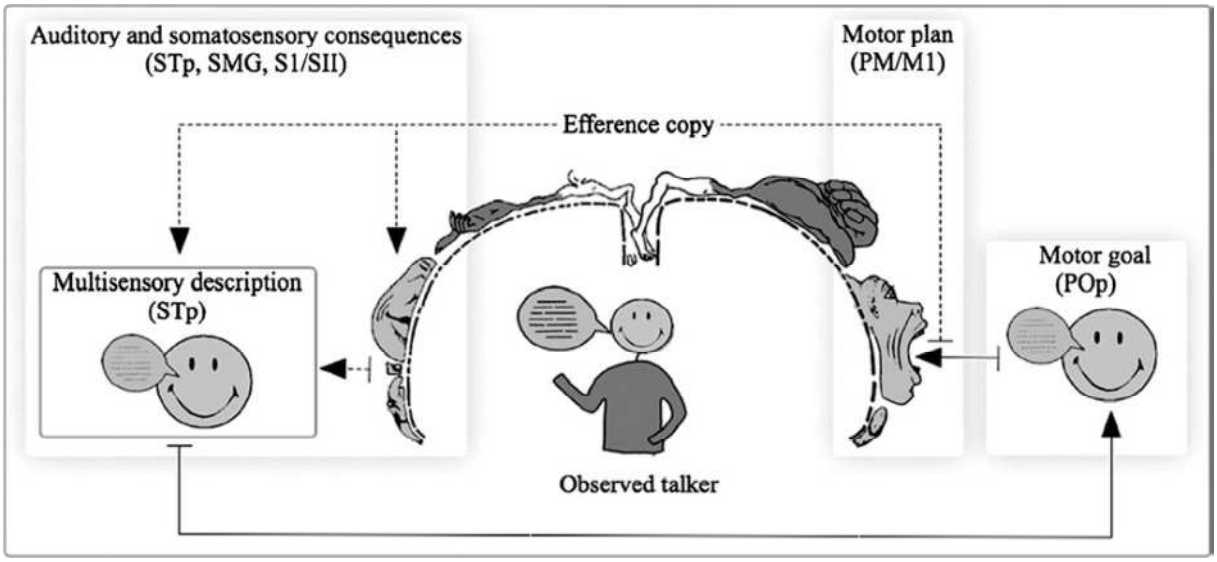


Figure 19 - Modèle de traitement cérébral de la parole audiovisuelle, proposé par Skipper (Figure tirée de (Skipper et al, 2007))

Ces commandes motrices fourniraient à leur tour une prédiction des conséquences auditives (STp) et somatosensorielles (SI/SII--> SMG--> STp) des gestes. Ces prédictions fourniraient enfin des contraintes sur le traitement de la parole en donnant la priorité aux interprétations particulières par rapport à des hypothèses primaires (STp).

2.3.5 Les mécanismes d'interaction multisensorielle de Senkowski

La description générale que proposent Senkowski et al. (Senkowski et al, 2008) des mécanismes d'interactions multisensorielles implique les mécanismes de synchronisation neuronale et de cohérence des ondes cérébrales, sur lesquels nous reviendrons. Dans ce cadre, Senkowski et al. proposent différents scénarios de traitement de la parole audiovisuelle (Figure 20). Le scénario le plus simple prédit que pendant les interactions multisensorielles, la synchronisation neuronale s'établit directement entre les aires sensorielles primaires, et que c'est alors dans ce mécanisme de connexion directe entre cortex auditif et visuel primaire que se réalisent les interactions audiovisuelles (Figure 20(a)). Une seconde possibilité est que la cohérence neuronale mette en jeu les aires associatives multisensorielles, telles que les régions pariétales ou temporales supérieures (Figure 20(b)). Les mécanismes de cohérence peuvent également impliquer à la fois des mécanismes d'interaction directe entre régions unimodales et augmentation d'activité oscillatoire dans les aires multisensorielles. Ceci peut refléter des interactions bottom-up et top-down entre les aires unimodales et multimodales (Figure 20(c)). Les changements dans les aires multisensorielles impliquent souvent les régions frontales et préfrontales, qui peuvent à leur tour moduler les patterns temporels dans les aires pariéto-temporales par couplage oscillatoire (Figure 20(d)). Enfin, l'hypothèse jugée la plus probable par les auteurs est celle qui regroupe tous ces mécanismes. Les interactions corticales sont combinées dans un modèle complexe, qui implique à la fois les régions frontales, temporo-pariétales, ainsi que les cortex unimodaux (Figure 20(e)). Il est probable que ce réseau de cohérences temporo-pariéto-frontales se coordonne avec des structures subcorticales, telles que les noyaux thalamiques (Figure 20(f)).

2.4 Conclusion

Ainsi, on le voit, les propositions sur les architectures cognitives et les organisations corticales sous-jacentes sont multiples, impliquant divers types de processus de fusion et de flux d'information entre représentations sensorielles et motrices. Néanmoins, toutes ces propositions reposent sur un principe implicite de prise d'information indépendante entre voies auditive et visuelle avant interaction. C'est ce principe que nous allons tenter de remettre en cause dans le chapitre suivant.

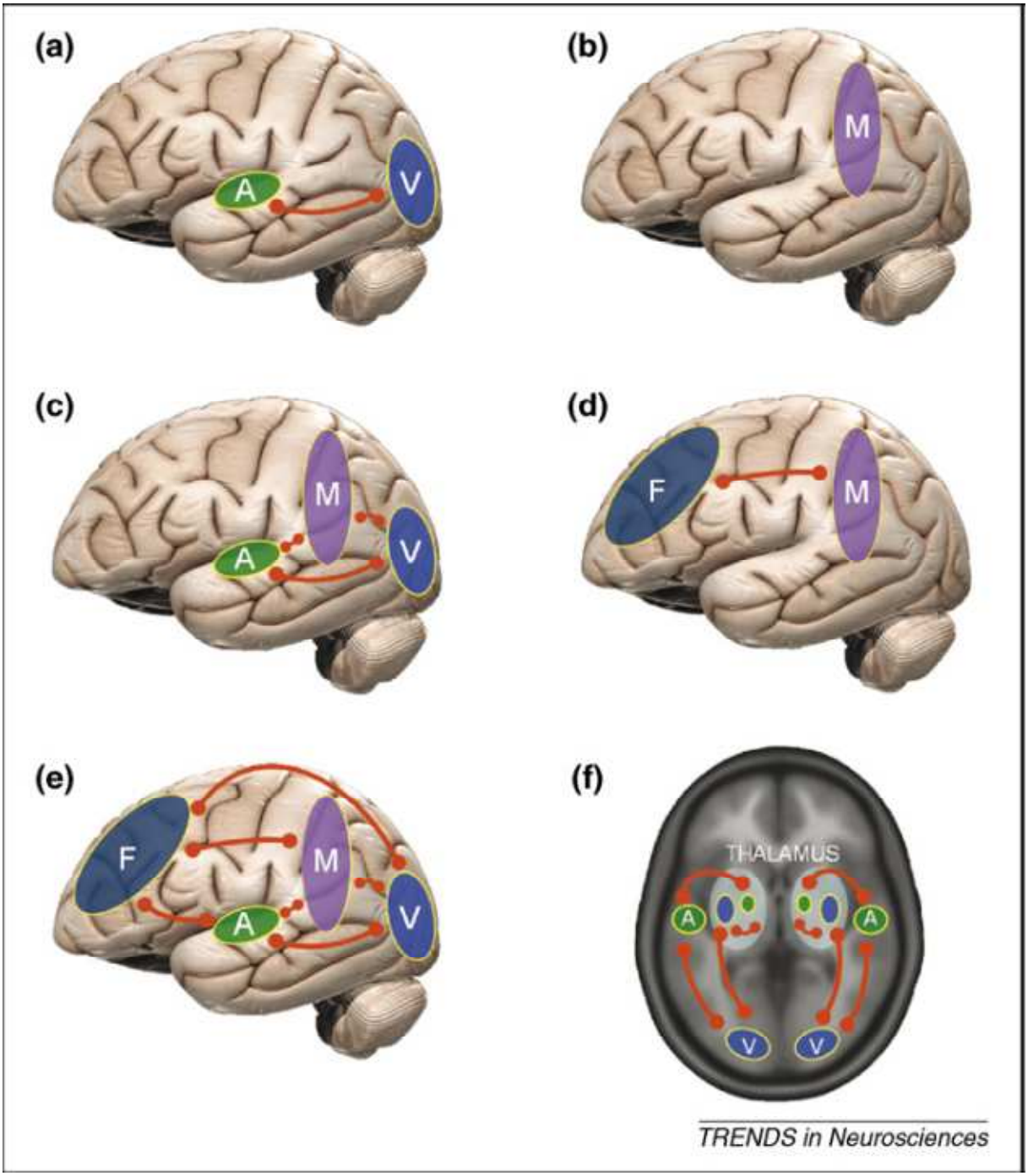


Figure 20 - Les scénarios hypothétiques pour les interactions multisensorielles à travers la cohérence neuronale. (Figure tirée de (Senkowski et al, 2008)). Les abréviations : A – cortex auditif, V – cortex visuel, M – régions multisensorielles, F – cortex préfrontal.

Chapitre 3. Un mécanisme de liage audiovisuel préalable à la fusion ?

3.1 Eléments de mise en évidence d'un niveau d'interaction précoce

Pour résumer, différents modèles d'intégration audiovisuelle existent dans la littérature, nous en avons présenté un certain nombre, tant d'un point de vue fonctionnel que neuroanatomique. Mais jusqu'à présent il n'y a pas de consensus clair sur la question de la convergence audiovisuelle. La vision classique considère que l'information des modalités différentes est élaborée indépendamment avant convergence. Néanmoins, des architectures telles que celle présentée par Senkowski et al. mettent en jeu des processus plus complexes alliant interactions directes entre représentations monosensorielles et processus de convergence à plus haut niveau, impliquant des boucles de retour. Plus concrètement, nous allons voir que plusieurs séries de données expérimentales suggèrent l'existence d'un processus d'interaction précoce entre flux auditif et visuel avant convergence et fusion.

3.1.1 Interactions audiovisuelles précoces en électrophysiologie

Nous avons mentionné précédemment les études d'électrophysiologie montrant des effets d'interaction audiovisuelle à des stades précoces. Ainsi, plusieurs études montrent des effets de la modalité visuelle sur l'onde N1. N1 (ou N100) correspond à un pic négatif de la réponse de potentiel évoqué mesuré par électroencéphalogramme (EEG) dans les régions fronto-centrales du cortex, autour de 100 à 150 ms par rapport au début de la stimulation. La source d'activité N1 est censée être principalement générée dans le cortex auditif primaire et les aires associatives du gyrus temporal supérieur (gyrus de Heschl et planum temporale). Elle est généralement considérée comme correspondant à une étape pré-représentationnelle et pré-attentive. Or, si l'on compare les potentiels évoqués fronto-centraux pour des stimulations de parole auditive vs. audiovisuelle, il apparaît que la composante visuelle produit une modulation de N1, impliquant à la fois une diminution d'amplitude et une anticipation temporelle légère (Besle et al, 2004), (van Wassenhove et al, 2005) (Figure 21).

Ainsi, l'influence du signal visuel sur l'activité N1 apparaît comme un effet précoce, peu compatible avec un mécanisme de fusion proprement dit mais plutôt comme une modulation de l'information visuelle dès les premières étapes du traitement auditif.

D'autres études portent sur le mécanisme de « négativité de discordance » (en anglais MMN ou mismatch negativity) qui apparaît sur les électrodes de la région temporale autour de 100 ms après l'arrivée d'un signal acoustique discordant au sein d'une séquence de stimuli acoustiques récurrents. Une stimulation visuelle appariée à une séquence de stimuli auditifs constants, et induisant occasionnellement un effet McGurk, conduit à l'apparition d'un effet MMN, environ 85 ms relativement à l'onset acoustique, ce qui confirme l'implication de l'entrée visuelle dans les processus précoces de traitement de la parole audiovisuelle (Colin et al, 2002), (Colin et al, 2004), (Ponton et al, 2009).

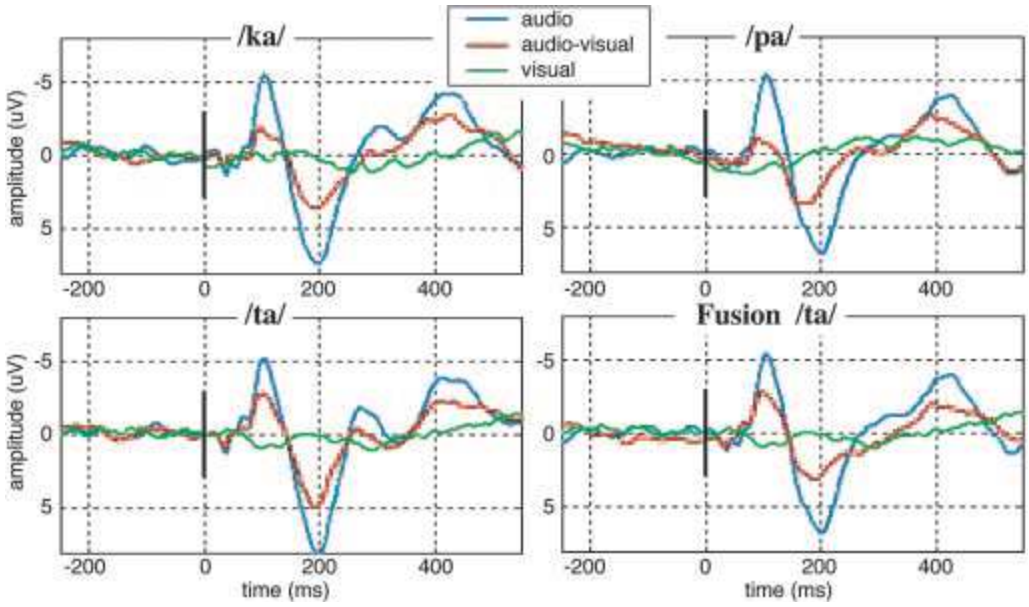


Figure 21 – Potentiel évoqué (ERP) moyen dans la zone centropariétale (CPz). La ligne verticale (oms) marque le début du signal auditif. Figure tirée de (van Wassenhove et al, 2005).

3.1.2 Facilitation audiovisuelle de la détection de traits phonétiques induisant un gain de reconnaissance

Nous avons montré dans le chapitre 1 que la parole visible fournit à la fois un gain d'intelligibilité (Sumby & Pollack, 1954), (Erber, 1969), (Benoit et al, 1994) et une amélioration de la détection de la parole dans le bruit (Kim & Davis, 2004), (Grant & Seitz, 2000). L'expérience de Schwartz et al. (Schwartz et al, 2004) a permis de montrer que le gain de détection pouvait se traduire par un gain d'intelligibilité d'une nature différente de celle de la lecture labiale. Pour cela, les auteurs ont associé un même mouvement des lèvres (pour la production d'une voyelle arrondie telle /u/ ou /y/) à différentes syllabes auditives dans le bruit. Toutes ces syllabes correspondent au même geste visuel : /tu/, /du/, /ku/, /gu/, /ty/, /dy/, /ky/, /gy/). Les résultats montrent un gain d'intelligibilité, qui ne peut être dû à la lecture labiale en soi, puisque le même geste visuel est toujours présenté. La simple présence du signal visuel, qui démarre 100 ms avant le signal auditif, améliore l'intelligibilité des syllabes d'environ 5-10% (Figure 22). Les auteurs expliquent ce résultat par un effet attentionnel, dans lequel le mouvement des lèvres, qui précède le début de voisement d'environ 200 ms (Figure 23), annonce au locuteur la nécessité de se préparer à extraire les informations auditives pertinentes. L'analyse plus détaillée montre que la présence du signal visuel améliore essentiellement la reconnaissance sur le voisement - et pas sur le lieu d'articulation – ce qui suggère que c'est l'indice de prévoisement (barre horizontale sur la figure) qui bénéficie de la détection améliorée. L'interprétation conduit donc à l'hypothèse qu'il y a deux mécanismes combinés qui produisent un gain de compréhension, l'un lié à l'amélioration de la détection des indices acoustiques par le guide visuel et l'autre à la lecture labiale proprement dite.

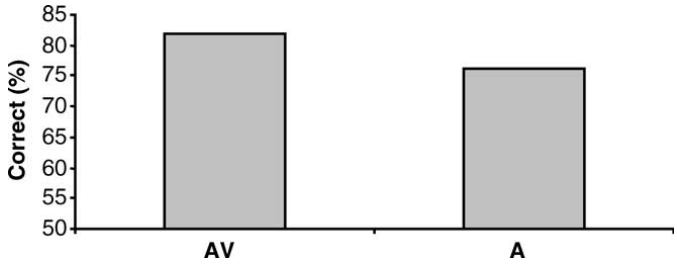


Figure 22 - L'effet d'amélioration d'intelligibilité des phonèmes par une simple présence du signal visuel. Figure tirée de (Schwartz et al, 2004).

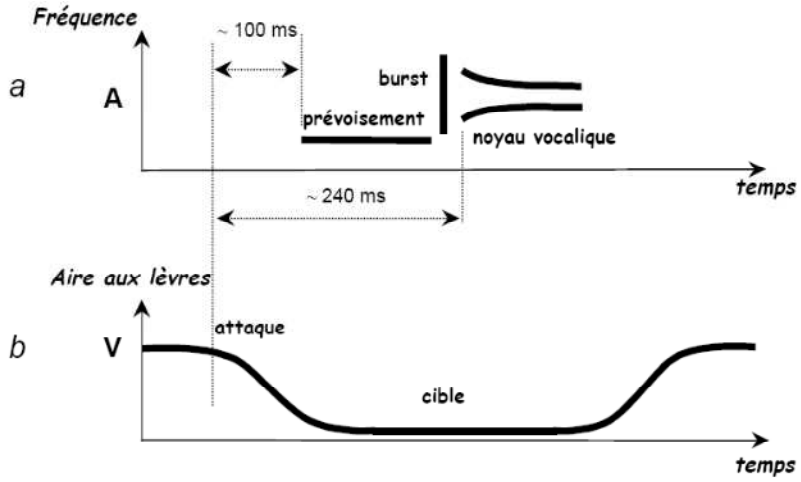


Figure 23 - Schéma de la structure acoustique d'une séquence plosive-voyelle en français (a) et geste labial correspondant (b). Figure tirée de (Schwartz et al, 2004)

Notons que le gain de reconnaissance disparaît si nous remplaçons une image des lèvres par des stimuli visuels non-parole : une barre verticale simulant les mouvements labiaux malgré le maintien de la corrélation temporelle parfaite entre les signaux acoustiques et visuels (Schwartz et al, 2004), ou un rectangle dynamique ou statique, corrélé avec l'enveloppe acoustique (Bernstein et al, 2004), voire la phrase écrite présentée simultanément à la parole acoustique.

3.1.3 Influence réciproque de la modalité auditive sur la perception visuelle

Nous n'avons mentionné jusqu'à présent que la modulation d'intelligibilité de la parole auditive par un signal visuel. On peut également trouver dans la littérature des mises en évidence d'influence de la lecture labiale par le signal auditif. Ainsi, les études de Brungart et al. (Brungart & Simpson, 2005) présentent des effets de distracteur audio sur l'intelligibilité d'un signal visuel en lecture labiale de mots (noms et digits). Ces effets se traduisent en l'occurrence par une diminution d'intelligibilité du signal visuel en présence du signal audio incohérent. Mais les auteurs montrent que le signal auditif doit être relié au signal visuel par des propriétés de cohérence minimale pour que l'interférence puisse avoir lieu. A minima deux conditions doivent être remplies : le distracteur auditif doit être un signal de parole et il doit être synchronisé temporellement avec le stimulus visuel.

3.2 Analyse des scènes perceptives

Ainsi, un certain nombre de données expérimentales, provenant à la fois de la neurophysiologie et de la psychologie comportementale, suggèrent qu'il pourrait exister un processus d'interaction audiovisuelle précoce avant fusion et décision.

Dans la littérature sur la perception auditive ou visuelle la question du traitement précoce est associée en général au problème d'analyse des scènes visuelles ou auditives. Les questions principales abordées concernent comment on extrait des objets différents, comment on relie l'information sur un même objet et également comment on relie ces différentes informations au sein de différentes cartes corticales. D'un point de vue pratique, les enjeux sous-jacents peuvent concerner la question de savoir comment un bébé reconnaît et suit la voix de sa maman ? Comment on peut percevoir la qualité individuelle des instruments de musique dans un orchestre (Helmholtz, 1877) et pourquoi parfois on perçoit l'orchestre et parfois les instruments ? Comment on peut suivre une voix particulière en présence des sons ou des bruits (de fond) dans l'environnement, comme c'est le cas dans l'effet « Cocktail party » (Cherry, 1953) ?

Nous allons passer en revue trois cadres théoriques incontournables dans la réflexion sur les mécanismes d'analyse de scènes dans notre contexte audiovisuel.

3.2.1 La psychologie de la forme (Gestalt)

C'est au plein cœur de ces questions que s'est développée la psychologie de la forme au début de XXème siècle en Allemagne, autour de ses auteurs principaux Wertheimer, Kohler et Koffka. La vision générale se base sur le principe que la perception sous sa forme de représentation mentale traite la scène spontanément comme des ensembles structurés, c'est-à-dire des formes et non comme une simple addition ou juxtaposition d'éléments. L'hypothèse est donc que le système perceptif perçoit les objets dans leur entièreté avant la perception des parties individuelles. Le but principal des gestaltistes est de comprendre l'organisation des éléments dans une unité, à partir de principes tels que l'émergence, la réification, la multistabilité et l'invariance.

L'illustration classique d'un ***principe d'émergence*** est celle du chien, qui apparaît spontanément dans l'ensemble des traces (Figure 24). Dans cet exemple nous reconnaissons un chien comme un objet entier, sans parcourir préalablement les parties que constituent les pieds, le nez, les oreilles.

Figure 24 -illustration du principe d'émergence, tirée de (Lehar, 2003)

La ***réification*** est un principe de construction spatiale perceptive d'un percept qui contient une information plus explicite par rapport aux stimuli sensoriels présentés. Sur la Figure 25 (A) on perçoit le triangle, tandis que ce triangle n'est pas dessiné. Sur les figures (B) et (D) nous pouvons percevoir que les formes dessinées font la partie d'une seule forme, et sur (C) on

reconstitue une figure 3D.

Le troisième principe est un principe de perception multistable ou ***principe de multistabilité***. Dans le cas de stimuli ambigus, le percept bascule entre les interprétations possibles. Des exemples classiques sont ceux du cube de Necker et du vase de Rubin (Figure 26). La perception d'un cube bascule d'une face avant à l'autre, et le vase alterne avec les deux visages.

Le ***principe d'invariance*** décrit une propriété de la perception des formes, qui est que nous n'avons pas de difficulté à reconnaître une forme malgré ses transformations spatiales, telles que rotation, translation, échelle, déformation élastiques, variation des lumières. Sur la Figure 27 (A) les objets sont facilement reconnus avoir la même forme. Cette forme est bien différente des formes sur l'image (B). Elle peut être facilement reconnue malgré les déformations élastiques (C) ou malgré l'utilisation des éléments graphiques différent s (D).

Les quatre principes sont considérés par les gestaltistes comme des aspects différents d'un mécanisme dynamique unifié qui devraient être modélisés conjointement.

En plus de principes de perception des formes les gestaltistes ont formulé des lois de groupement selon lesquelles les parties séparées sont perçues comme appartenant au même objet ou au même groupe d'objets. Notre système perceptif accéderait aux formes perceptives en utilisant les lois de groupement.

L'idée générale des gestaltistes est exprimée par la ***loi de la bonne forme***, qui postule qu'un système perceptif tente d'organiser l'expérience perçue de manière régulière, ordonnée, symétrique et simple. Les individus auraient tendance à percevoir le monde dans ses formes de base simplifiées. Ce principe général est explicité sous la forme des lois de proximité, de similarité, de fermeture, de symétrie, de destin commun, de continuité et de l'expérience passée, que nous expliciterons en partie dans la section suivante dans le contexte de l'audition.

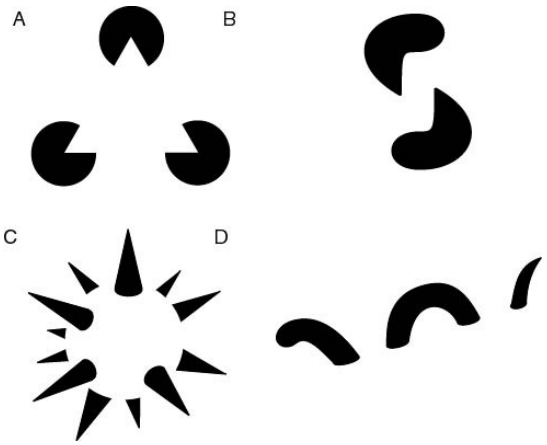


Figure 25 – illustration du principe de réification, tirée de (Lehar, 2003). A : le triangle de Kanizsa, B : un ver volumétrique de Peter Tse, C : une sphère à pointe de Idesawa, D : un monstre marin de Peter Tse

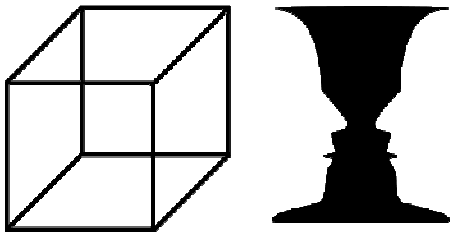


Figure 26 – illustration du principe de multistabilité: le cube de Necker et la vase du Rubin, tirés de (Lehar, 2003).

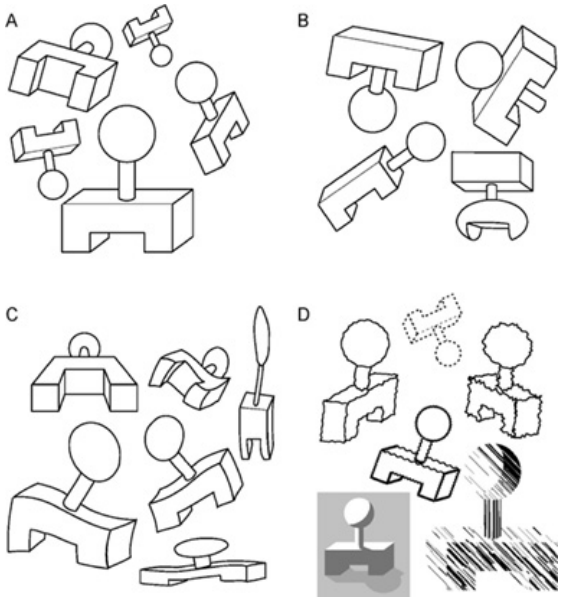


Figure 27 – Illustration du principe d'invariance, tiré de (Lehar, 2003)

Dans la littérature moderne la critique majeure de la théorie de la Gestalt est l'absence de modèle perceptif concret capable d'expliquer la mise en œuvre de ces principes, formulés de manière globale et qualitative, et de manière générale trop vague pour fournir des principes computationnels clairs.

Néanmoins, un effet important des théories gestaltistes est d'avoir favorisé le développement des recherches sur les mécanismes d'analyse de scène. En effet, bien que la plupart des recherches des gestaltistes se soient concentrées sur les supposées lois de la perception visuelle, les mêmes principes de groupement peuvent être introduits et explorés dans la perception auditive, tactile, voire olfactive (voir par exemple Schwartz et al. (Schwartz et al, 2012) pour une introduction à un numéro spécial de revue sur l'étude des mécanismes de multistabilité à travers les modalités perceptives).

Bregman a joué un rôle majeur dans le transport et l'adaptation des principes d'organisation perceptive à la modalité auditive autour du concept, qu'il a largement popularisé, d'analyse de scènes auditives. C'est ce que nous allons décrire maintenant.

3.2.2 Bregman et l'analyse des scènes auditives

La vision des gestaltistes est que les lois formulées sont les propriétés fondamentales d'un système perceptif, qui sont à la base d'une capacité humaine à donner un sens aux entrées sensorielles. Ces propriétés d'organisation des scènes perceptives seraient innées. La vision de Bregman (Bregman, 1990) est plus complexe, de nature heuristique et évolutive. Il suppose que ces lois sont dérivées des caractéristiques générales du monde externe et basées sur l'expérience. L'environnement impose le problème tandis que le cerveau humain essaie de décrire l'état de son environnement.

Par rapport aux gestaltistes, Bregman contraste deux séries de mécanismes impliqués dans l'analyse de scène auditive : le groupement auditif bottom-up par des primitives et l'appel à des processus top-down par des schémas appris. L'organisation par les primitives est un processus pré-attentif, qui permet de grouper spontanément les différentes composantes. Au contraire, le groupement dirigé par les schémas est guidé par les processus attentionnels. L'apport principal de Bregman porte sur les mécanismes de groupement primitifs, s'appuyant sur des analyses variées du flux d'entrée, exploitant des paramètres et des sous-processus tels que hauteur, intensité, fluctuations d'enveloppe, cohérences de fréquence, localisation, etc., à partir desquels le système tente de résoudre le problème du groupement pour s'assurer que toutes ces propriétés concernent un même événement ou un même objet.

Au cœur du mécanisme de groupement par primitives, Bregman considère le mécanisme de détermination du « destin commun ». Ce mécanisme considère que si on observe des variations cohérentes dans les parties différentes d'une scène donnée, il y a une forte chance que ces parties appartiennent au même objet. Par exemple quand un son harmoniquement structuré change dans le temps, tous ses harmoniques sont modulés en fréquence et en amplitude de manière à maintenir la relation harmonique. Ces régularités peuvent être utilisées dans le sens inverse pour déduire la structure sous-jacente. Quand les relations entre composantes fréquentielles maintiennent une relation harmonique malgré des changements de fréquence, d'amplitude ou de localisation de chaque composante individuelle, il est probable que toutes ces composantes soient associées à un événement physique cohérent. Pour Bregman, les systèmes perceptifs animaux ont évolué pour répondre à certains

facteurs constants dans leur environnement, notamment ceux associés à ce principe de « destin commun ».

Contrairement aux travaux des gestaltistes qui sont principalement centrés sur la perception visuelle, Bregman s'est attaché à montrer que certains principes gestaltistes peuvent être appliqués aussi à l'audition (Bregman, 1990).

Ainsi, pour illustrer la ***loi de continuité***, qui prédit que nous avons une tendance à associer les éléments qui sont en continuité dans l'espace, si on fait alterner un son doux et un son fort puis à nouveau un son doux, plutôt que de percevoir une modulation d'intensité, on perçoit un son doux stable, sur lequel se superpose temporairement un second son doux : le son fort est ainsi décomposé en une base, qui assure la continuité du son doux, et un second son superposé.

La ***loi de similarité*** permet d'assurer l'appartenance à une même source. En vision le groupement peut passer par une association perceptive de stimuli de même couleur, et l'analogue auditif est fourni par le groupement par le timbre.

La ***loi de proximité*** peut-être réalisée par la proximité temporelle ou la proximité fréquentielle, comme le montre la fameuse expérience de Van Noorden (Van Noorden, 1975), alternant sons graves (A) et aigus (B) dans une séquence de type (ABA_ABA_ABA…) où « _ » est un silence de même durée que les sons A et B. Cette séquence conduit soit à la perception d'un « galop » de séquences « ABA », tous les sons étant alors groupés dans un seul flux, soit à la perception de deux flux indépendants de « A » et de « B ». Le groupement en un ou deux flux est géré par la proximité temporelle et spectrale : des sons qui sont très proches sur l'axe temporel ou fréquentiel ont plus de chance d'appartenir à la même source.

Un autre exemple est le principe ***d'allocation exclusive*** (conduisant au phénomène de ***multistabilité***), qui est illustré dans la vision du vase de Rubin, que nous avons déjà présenté (Figure 26). La multistabilité provient de ce que notre système perceptif attribue la ligne de contour soit au vase soit aux deux visages. Cette propriété d'allocation exclusive peut se retrouver dans l'audition avec par exemple une séquence de tons telle que celle de la Figure 28 (Bregman & Rudnicky, 1975). Dans cette expérience, la tâche des sujets est de décider l'ordre des tons cibles (A et B) intégrés dans la séquence. Quand ils sont présentés isolés la décision est facile. Mais quand ils sont environnés par des tons F (FABF), il devient difficile d'entendre l'ordre au sein de cet objet FABF complexe. La question que se sont alors posée les auteurs est de savoir comment séparer les tons cibles A et B et les tons perturbateurs F dans des flux différents (FF et AB) pour que l'ordre des tons A et B redevienne clairement audible. Pour cela ils ont introduit des séquences de tons C avec une fréquence spécifique. Quand la fréquence des tons C était beaucoup plus basse que celle des tons F, les tons F étaient groupés avec les tons A et B (FABF), donc l'ordre des tons A et B n'était pas clair pour les auditeurs. Mais quand la fréquence des tons C était plus proche de celle des tons F, ils étaient groupés ensemble dans un flux CCCFFCC qui « éliminait » les tons perturbateurs F et rendait l'ordre des tons AB facile à déterminer, car ils étaient extraits dans un flux séparé. Dans cet exemple le principe d'allocation exclusive a permis de rendre inopérant un ensemble de composantes en les allouant à un flux « parallèle » au flux cible A-B.

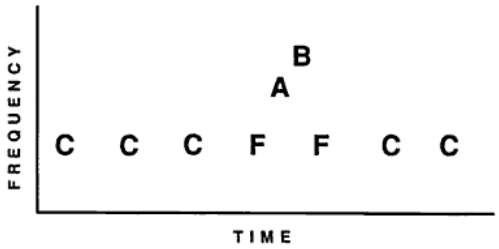


Figure 1.7
A tone sequence of the type used by Bregman and Rudnicky (1975).

Figure 28– Illustration du principe d'allocation exclusive en audition. Figure tirée de (Bregman & Rudnicky, 1975).

Cependant, il n'existe pas de principe général qui garantisse en tout circonstance le succès du groupement par des primitives, car les conditions de décomposition des scènes peuvent être extrêmement variables, et les primitives peuvent conduire à des décisions parfois contradictoires. Dans ce contexte, Bregman propose que le groupement fasse intervenir des mécanismes de type votes, avec des effets de compétition ou au contraire de renforcement. Cette approche peut expliquer l'instabilité du résultat dans des situations ambigües. Dans des cas non ambigus, le système perceptif pourra attribuer une composante donnée à un flux plutôt qu'un autre. En cas de concurrence forte entre deux organisations, elles peuvent être « viables » l'une et l'autre. Ainsi, dans l'expérience de Bregman et Pinker (Bregman & Pinker, 1978) groupement temporel et spectral sont mis en concurrence (Figure 29) : les tons A et B peuvent être regroupés en un flux A-B par un mécanisme de proximité fréquentielle (primitive 1), dans ce cas le sujet va percevoir les deux flux AB et C. Au contraire les tons B et C peuvent être regroupés en un objet BC par un mécanisme de cohérence temporelle (primitive 2), donc la séquence sera perçue comme deux flux A et BC. La concurrence entre les décisions prises par chaque primitive implique un système de « gestion des conflits » qui pour Bregman peut être de type « vote » (et que l'on dénommerait actuellement « fusion de décision ».

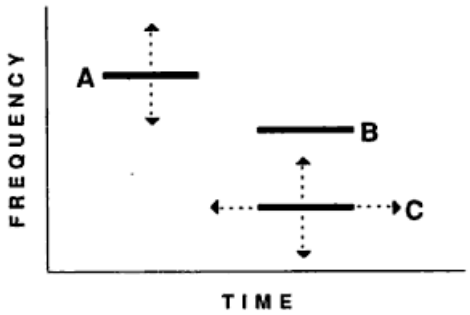


Figure 29– Concurrence entre deux lois de groupement. Figure tirée de (Bregman & Pinker, 1978).

3.2.3 Le modèle de Treisman

Dans les années 80s Anne Treisman a introduit le modèle FIT (Feature Integration theory), qui essaie d'expliquer le problème du liage visuel en soulignant le rôle de l'attention spatiale dans la combinaison de différentes caractéristiques perceptives d'un objet (Treisman & Gelade, 1980). Par rapport aux gestaltistes et leur centrage sur la notion de formes, et à Bregman, qui sépare les notions de primitives et de schémas dans l'analyse des scènes auditives, Treisman, dans l'analyse des scènes visuelles, introduit la question du liage entre des cartes de primitives comme la couleur, l'orientation, la fréquence spatiale, la luminosité, la direction de mouvement, etc. Chaque carte est associée à un ensemble de valeurs primitives, et les cartes différentes sont traitées par des canaux indépendants et codées séparément. Pour caractériser un objet unique il faut alors associer sous une forme ou une autre ces cartes, ce qui se fait, selon Treisman, à l'aide de l'attention spatialement focalisée, qui joue un rôle de « colle » qui intègre les différents attributs de l'objet (Figure 30). Quand l'attention est chargée ou distraite, il devient plus difficile d'effectuer cette intégration ce qui peut conduire à des conjonctions illusoires, en associant par exemple les caractéristiques (telles que la couleur et la forme) de deux objets différents en un seul objet « chimérique » (Treisman & Schmidt, 1982).

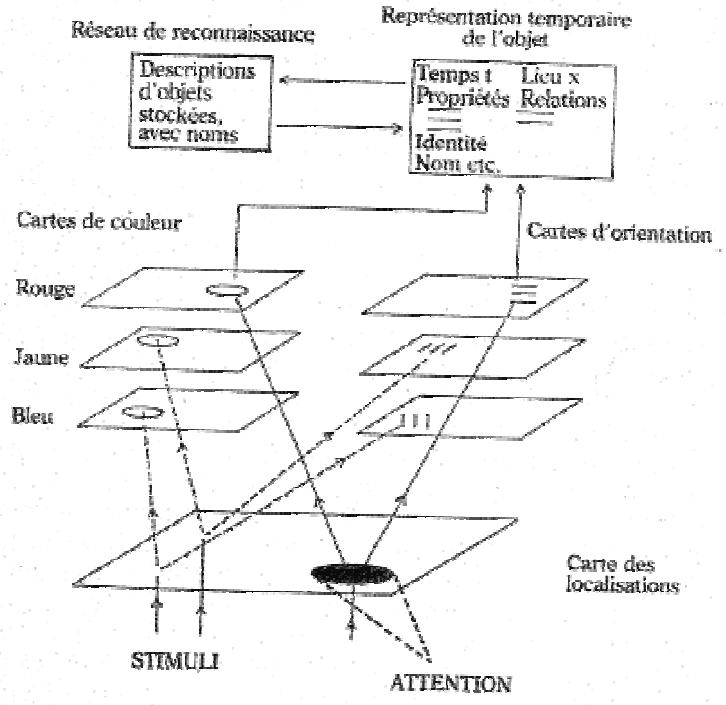


Figure 30 – Schéma du modèle pour la perception des traits et des objets. Figure tirée de (Treisman, 1992).

Pour faire une illustration, nous avons choisi une expérience parmi de nombreuses réalisées par Treisman (Treisman & Gelade, 1980). Dans cette expérience la tâche des sujets était de trouver une cible le plus vite possible dans deux conditions. Dans la condition « conjointe », la cible était décrite par des caractéristiques de couleur et de forme (par exemple chercher un T vert). Dans la condition « disjointe » le sujet devait identifier deux cibles en parallèle, chacune avec sa caractéristique : une couleur (bleu) et une forme (S). Dans chaque

condition le sujet devait analyser les deux cartes sensorielles requises, carte de couleur et carte de forme, dont on suppose qu'elles sont codées indépendamment dans le cerveau. Mais dans la condition « conjointe » il devait tester également leur combinaison.

Les cibles étaient présentées dans 4 tailles d'écran différentes, avec 1, 5, 15 ou 30 objets (cibles et distracteurs). Dans chaque condition les distracteurs étaient les mêmes : X verts et T marrons. Les résultats (Figure 31) font apparaître des performances très différentes selon la tâche et selon qu'elle donne un résultat positif (une cible est présente) ou non.

Dans l'expérience de disjonction, les cartes peuvent être exploitées indépendamment. Si une cible existe dans l'une ou l'autre carte, elle est perçue rapidement et indépendamment du nombre de distracteurs, par un effet de « pop out », ce qui explique le temps de réaction faible et constant dans cette situation (DISJUNCTION-POS). Si aucune cible n'est présente, il faut explorer complètement une des deux cartes (ce qui suffit à éliminer l'hypothèse), et le temps varie linéairement avec le nombre d'objets (DISJUNCTION-NEG).

Dans la condition de conjonction, il est nécessaire d'explorer chaque objet et de lier dans les deux cartes les attributs « forme » et « couleur », ce qui prend un temps proportionnel au nombre d'objets et double du temps d'une seule carte. Ceci explique la valeur double des temps dans le cas (CONJUNCTION-NEG) par rapport à (DISJUNCTION-NEG). Enfin, dans cette même condition de conjonction, une cible positive, qui ne peut pas surgir par « pop out », demande statistiquement d'explorer la moitié des objets, selon que le hasard fait rencontrer la cible au début ou à la fin de l'exploration (CONJUNCTION-POS).

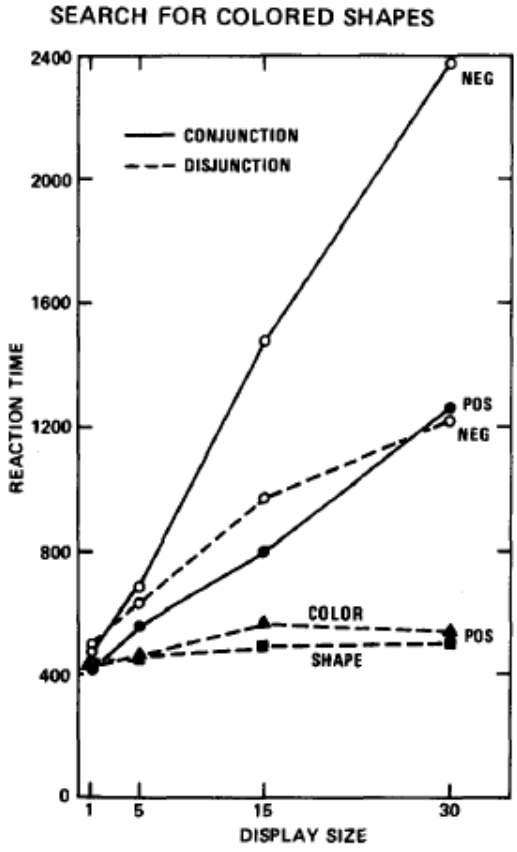


Figure 31 - Temps de recherche de la cible, selon les conditions et les réponses, Figure tirée de (Treisman & Gelade, 1980)

Ainsi, l'interprétation de Treisman implique que les sujets scannent successivement les objets avec l'attention spatialement focalisée, ce qui assure l'intégration correcte des traits dans l'espace multidimensionnel. Cette expérience est intéressante car elle démontre un liage entre deux cartes différentes et le rôle des mécanismes attentionnels (ici guidés par la composante spatiale) pour relier ces deux cartes.

Un intérêt majeur pour nous du modèle de Treisman est de nous rapprocher de la question, centrale pour nous, du liage audiovisuel. Même si le travail de Treisman se fait purement dans le domaine de l'analyse des scènes visuelles, il permet d'introduire un élément nouveau qui est celui de la fusion de traits, peu abordé par Bregman. Ceci était rendu possible grâce à tout le travail de recherche préalable dans le domaine visuel sur les cartes de traits, travail qui n'est pas réellement abouti jusqu'à présent dans le domaine auditif. Avec ce modèle, nous avons une indication sur le fait que la jonction ou le liage à un niveau plus élevé que les traits primitifs n'est pas un processus purement automatique et nécessite l'intervention de processus attentionnels.

3.2.4 Mécanismes neurophysiologiques sous-jacents

La question des mécanismes généraux de liage, qui désigne l'ensemble des mécanismes permettant au cerveau de produire une représentation cohérente du monde extérieur, à travers la multiplicité des canaux sensoriels véhiculant l'information, et la multiplicité des représentations mentales portées par des aires cérébrales différentes (Revonsuo & Newman, 1999), est très vaste. Elle a alimenté un nombre considérable de recherches sur les mécanismes neurophysiologiques sous-jacents. Sans entrer dans une présentation détaillée de ces mécanismes, ce qui serait hors du propos de cette thèse, mentionnons pour mémoire les propositions sur les mécanismes de synchronisation neuronale, qui jouent un rôle majeur dans les recherches actuelles. Les oscillations neuronales pourraient fournir la base de mécanismes de mise en cohérence par processus de synchronisation entre assemblées neuronales dynamiques. Ces synchronisations permettraient d'établir des processus de connectivité fonctionnelle entre des populations de neurones, spatialement distribués (Senkowski et al, 2008). Dans un modèle de « codage multiplexe » distribuant l'information neuronale dans des bandes de fréquence différentes, l'activité rythmique dans la bande dite « gamma » (30-50 Hz) est considérée jouer un rôle majeur dans ce processus de liage par synchronisation (Fries et al, 2007). L'activation gamma semble intervenir à la fois dans le liage précoce des primitives (Engel et al, 1999), (Singer, 1999), (Tallon-Baudry et al, 1996) et dans le liage top–down dirigé par des schémas (Tallon-Baudry et al, 1997).

3.3 Corrélations audiovisuelles

Notre travail de recherche est centré sur la question du liage entre les modalités auditives et visuelles, ce qui nous conduit un pas plus loin que les travaux précédents de la Gestalt, de Bregman ou Treisman. Dans notre cas, nous devons supposer l'existence de traits primitifs représentés dans des cartes de primitives monosensorielles, et il se pose alors la question du liage entre ces cartes dans les modalités auditive et visuelle.

On le voit, tous les modèles et toutes les propositions théoriques attribuent un rôle central aux propriétés de cohérence – essentiellement spatiale/positionnelle et temporelle –

entre les éléments à lier. Ceci nous ramène à la description initiale de l'information visuelle que nous avions introduite dans le chapitre 1, entre redondance et complémentarité. Ce sont bien les propriétés de redondance audiovisuelle qui sont susceptibles de servir de base aux processus de liage, s'ils existent. Nous allons donc présenter maintenant les études qui ont cherché à caractériser ces redondances, et à établir le contenu des corrélations audiovisuelles.

3.3.1 Yehia et collègues

Yehia et ses collaborateurs (Yehia et al, 1998) ont mesuré la corrélation entre les mouvements faciaux, les mouvements du tractus vocal et le signal acoustique. Pour recueillir leurs données ils ont utilisé un système OPTOTRAK (qui permet de mesurer une trajectoire 3D de capteurs infrarouges, placés sur la joue, le menton, les lèvres, etc.) avec 12 points pour mesurer la dynamique labiale, ainsi qu'un système EMMA (articulographe électromagnétique) avec 7 points pour capturer le mouvement du tractus vocal. Leurs résultats précurseurs montrent que 80-90% de la variance des mouvements faciaux peut être déterminée à partir des mesures sur le tractus vocal et vice versa. Ils observent également une corrélation de 70 à 85% entre la géométrie du tractus vocal et les paramètres caractérisant le signal acoustique. Ils montrent enfin que la forme de la langue peut être estimée assez correctement à partir des mouvements faciaux grâce à ces corrélations élevées (Figure 32).

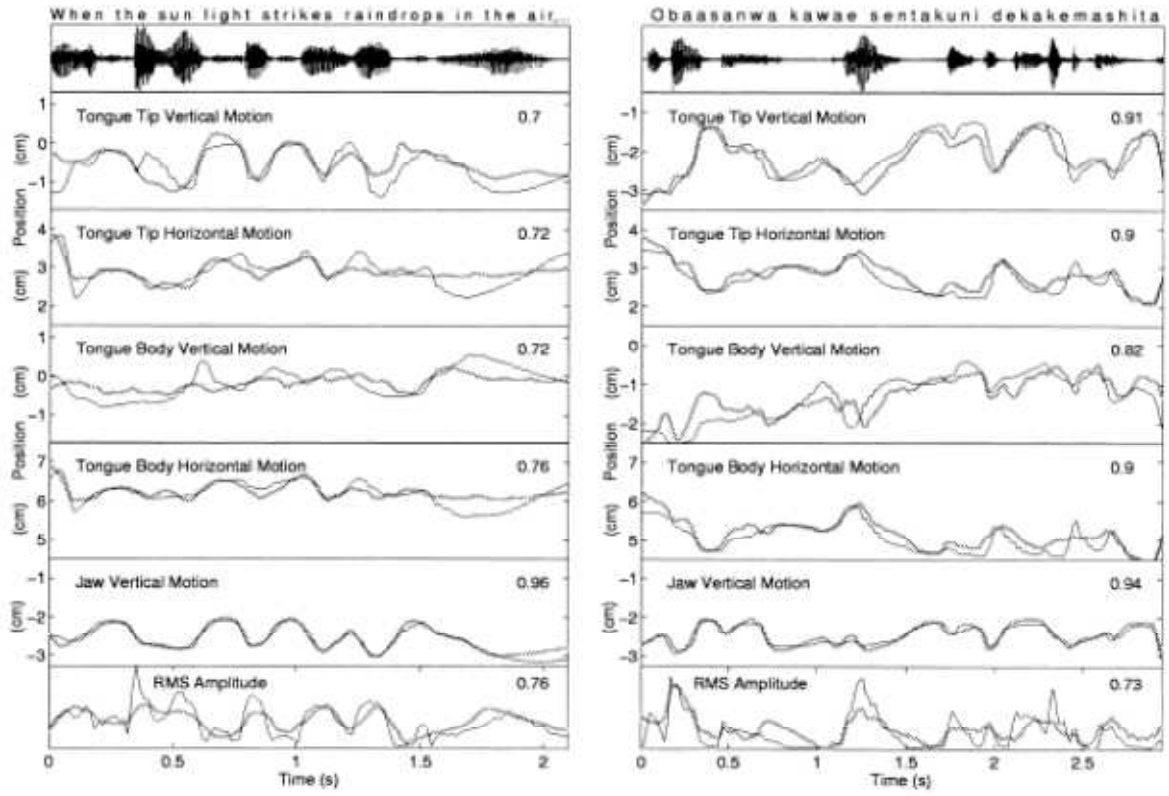


Figure 32 – Reconstruction du mouvement de la langue à partir des données faciales pour deux sujets masculins. La première ligne contient le signal acoustique, puis les lignes 2 à 6 présentent en noir les patterns articulatoires temporels et en gris les patterns estimés à partir de l'information faciale. Sur la dernière ligne on compare en noir l'amplitude RMS (moyenne quadratique) mesurée sur le signal et en gris l'amplitude RMS estimée à partir de l'information faciale. Le coefficient de corrélation pour chaque comparaison entre données et estimations est inscrit dans le coin supérieur droit. Figure tirée de (Yehia et al, 1998).

3.3.2 Barker et Berthommier

Barker & Berthommier (Barker & Berthommier, 1999) ont proposé une étude similaire, mais avec une technique de mesure d'articulation visuelle par« chroma key », qui permet une mesure précise de la dynamique labiale (technique que nous avons utilisée également, qui extrait un contour de lèvres maquillées en bleu à partir d'enregistrement visuel, et qui sera présentée plus en détail dans la suite de cette thèse).Ils obtiennent des résultats similaires (70-75% pour la reconstruction d'information acoustique à partir des données visuelles et 55-60% pour reconstruire le mouvement labial à partir des mesures acoustiques). Ils observent cependant que la reconstruction du signal acoustique à partir des seules données labiales est moins efficace.

3.3.3 Grant et collègues

Grant et al. (Grant & Seitz, 2000) ont mesuré la corrélation entre les mouvements labiaux et l'enveloppe de l'amplitude acoustique, qui était précédemment séparée en 3 sous-bandes. Ils ont obtenu une cohérence temporelle entre les variations d'ouverture des lèvres et l'enveloppe acoustique pour la bande intermédiaire. Ils ont pu ainsi mettre en correspondance ces taux de corrélation audiovisuelle avec les effets de la modalité visuelle sur la détection auditive, que nous avons déjà présentés dans la section1.2.3.

3.3.4 Chandrasekaran et collègues

Chandrasekaran et al. (Chandrasekaran et al, 2009) ont également observé une corrélation robuste et une forte correspondance temporelle entre l'ouverture de la bouche et l'enveloppe acoustique, ainsi qu'entre l'ouverture de la bouche et la première résonance du tractus vocal (~75%) (Figure 33). Ils montrent également que l'ouverture de la bouche et l'enveloppe auditive sont modulées temporellement dans une fenêtre de fréquences de l'ordre de 2-7 Hz.

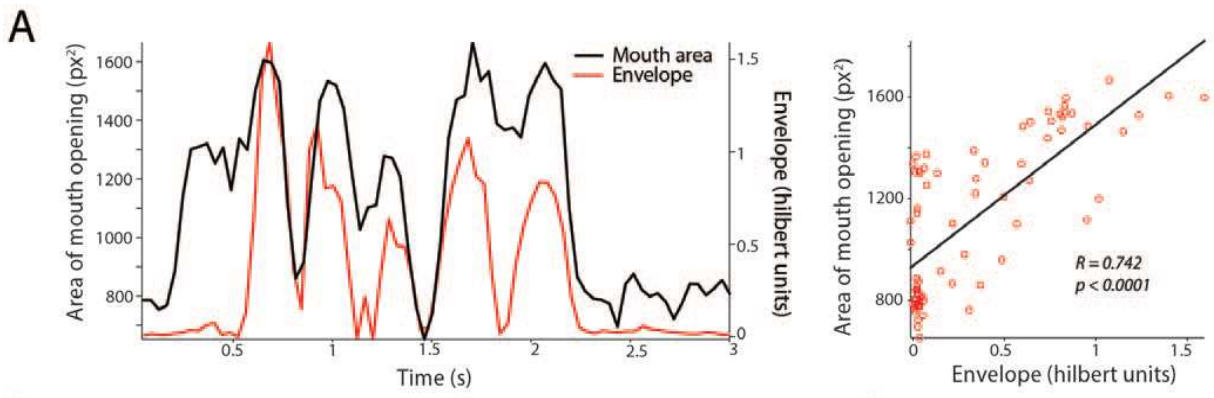


Figure 33– Corrélations moyennes entre l'aire de l'ouverture de la bouche et l'enveloppe acoustique. Figure tirée de (Chandrasekaran et al, 2009)

3.3.5 Jiang et collègues

Jiang et al. (Jiang et al, 2002) ont étudié la régression multilinéaire entre le mouvement du visage, de la langue et le signal acoustique. En termes de lieu d'articulation le lieu de la langue est un meilleur prédicteur que les lieux bilabial ou glottal. Il existe une certaine asymétrie dans la prédiction, en ce sens qu'il est plus simple de prédire les mouvements articulatoires que l'inverse. Ceci peut être dû au fait que l'acoustique de la parole est plus informative que les mouvements visuels. Ce travail montre aussi que les prédictions sont meilleures pour les syllabes que pour des phrases.

3.3.6 Berthommier

Berthommier (Berthommier, 2004) propose d'appliquer les propriétés de cohérence audiovisuelle pour des applications réalistes qui pourraient permettre de synthétiser l'information vidéo à partir du signal auditif et ainsi d'augmenter l'intelligibilité d'un signal auditif dans des conditions très bruités grâce à l'information visuelle. Pour ce faire il applique sur un signal audio bruité un filtre estimé à partir de l'information visuelle. Son étude montre un gain d'intelligibilité d'environ 4 dB, correspondant à environ 20% d'augmentation du score de compréhension de mots, consistant en des nombres ou des chiffres (Figure 34).

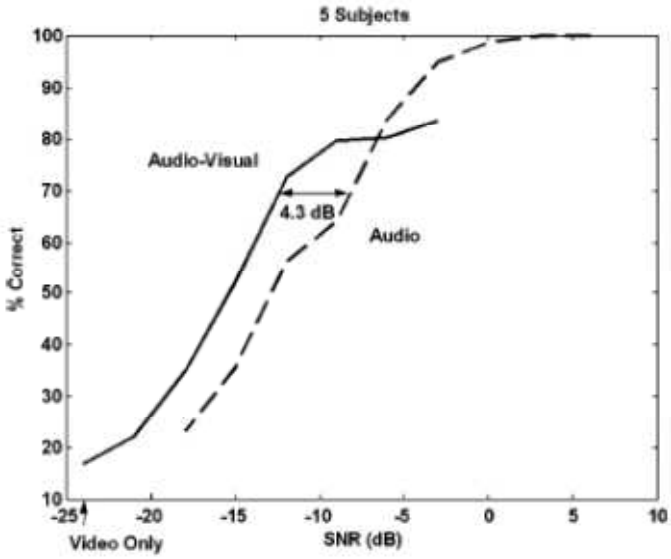


Figure 34 – Effet d'amélioration d'un signal auditif bruité a l'aide de données visuelles. Figure tirée de (Berthommier, 2004)

Pour expliquer les résultats obtenus, Berthommier (Berthommier, 2004) propose un modèle, où l'étape de mesures de corrélations audiovisuelles est une étape de bas niveau, préalable au traitement et à la fusion ultérieure des signaux auditifs et visuels avant décision. Nous discuterons ce modèle en détail dans le chapitre suivant.

3.4 Conclusion

Nous avons ainsi passé en revue les éléments factuels qui nous conduisent à supposer l'existence d'un mécanisme de liage audiovisuel préalable à la fusion et à la décision, présenté quelques architectures cognitives disponibles dans la littérature pour traiter de ce type de mécanismes, et décrit les principales études permettant de mettre en évidence des effets de corrélation audiovisuelle susceptibles de servir de base aux mécanismes de liage. Nous allons maintenant aborder notre thème central, qui est celui de la mise en évidence explicite d'un processus de liage audiovisuel en perception de la parole.

Chapitre 4. Stratégie expérimentale et plan du travail

4.1 Une hypothèse

Nous pouvons maintenant formuler explicitement l'hypothèse de cette thèse, en repartant de la formulation de Campbell (Campbell, 2008), qui organise le rôle de la modalité visuelle dans la perception de la parole entre son rôle complémentaire et son rôle de corrélation par rapport l'information auditive. Pour Campbell, ces deux propriétés font appel à des systèmes séparés, correspondant possiblement à des voies différentes dans l'architecture corticale sous-jacente.

Notre vision est différente, car notre proposition est que ces deux processus pourraient renvoyer à une architecture séquentielle-hiérarchique plutôt que parallèle-autonome. Dans un modèle proposé par Berthommier (Berthommier, 2004), une première étape d'évaluation de la corrélation entre les deux modalités conditionnerait la fusion, qui ne prendrait effet que si le taux de corrélation est suffisant (Figure 35). C'est cette architecture hiérarchique qui fournit le cadre de notre thèse.

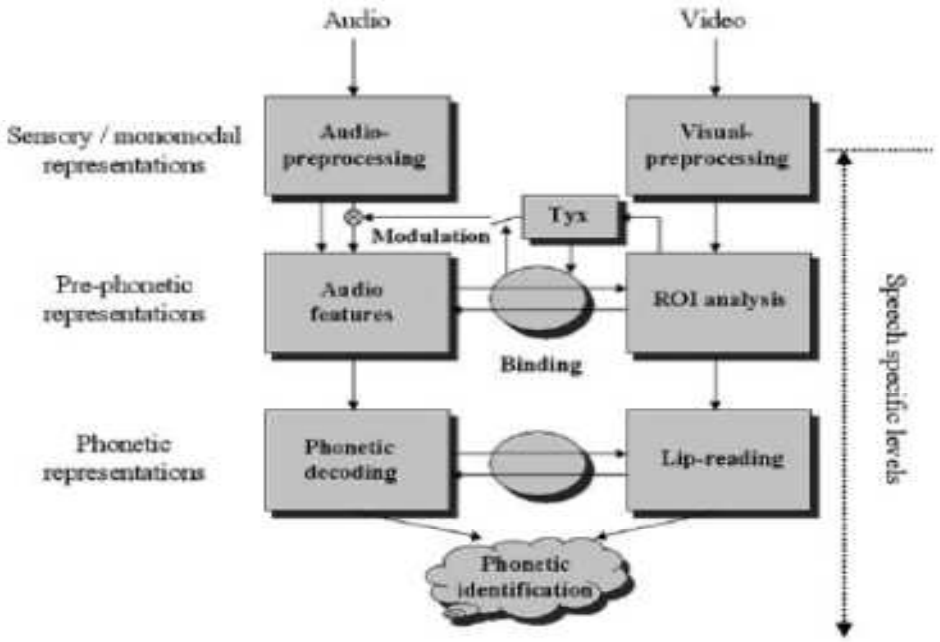


Figure 35 - Modèle avec plusieurs niveaux d'interaction audiovisuelle, qui contient la notion de liage. (Figure tirée de (Berthommier, 2004))

En étudiant la bibliographie sur la fusion audiovisuelle nous avons vu que jusqu'à présent la fusion audiovisuelle était toujours considérée comme un processus automatique relevant d'un seul niveau de fusion. Nous suggérons qu'il existe en réalité au minimum deux niveaux d'interaction : un niveau précoce (détection/comparaison /modulation) et un niveau tardif (fusion/reconnaissance/décision). Dans ce mémoire nous appellerons le niveau précoce « processus de liage » et le niveau tardif « fusion » (Figure 36). Les rôle et mécanismes de chacun de ces niveaux sont encore peu compris, mais nous suggérons que le système de liage peut moduler, sous une forme à déterminer, le processus de fusion audiovisuelle. C'est cette hypothèse qui fournit le cœur de notre travail.

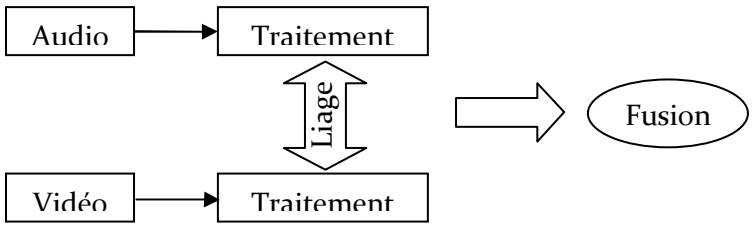


Figure 36 - Modèle qui représente l'idée centrale de ce manuscrit, avec l'existence supposée d'un processus de liage qui précède la fusion.

4.2 Un paradigme

La question qui se pose pour nous est donc de savoir si le mécanisme de détection précoce que nous avons décrit dans la Section 3.1 fait partie d'un système plus large assurant un rôle de liage conditionnel. Ce système permettrait, au cas par cas, de lier les entrées auditives et visuelles en un même flux, ou au contraire de les séparer en deux flux différents. Si c'est le cas, on doit pouvoir construire des situations expérimentales où on peut « débrancher » le second niveau de fusion, comme c'est probablement le cas dans les films doublés, où il ne faut pas intégrer les entrées auditive et visuelle dans la reconnaissance, puisqu'elles ne portent pas d'information cohérente.

Pour mettre en évidence l'existence d'un tel processus du liage, nous avons élaboré un paradigme expérimental original. Il nous fallait d'abord trouver une manière de mesurer le degré de fusion, et nous avons considéré que l'effet McGurk fournissait un bon outil. Il a été depuis sa découverte considéré comme une mise en évidence claire et sans équivoque de la fusion audiovisuelle. Nous avons vu que la perception d'effet McGurk dépend de nombreux facteurs : du locuteur, de l'intensité du signal, du bruit environnant, du sujet, de l'âge, de la langue, etc. A l'inverse, cet effet semble précisément assez robuste à des variations expérimentales telles que les incohérences de localisation entre flux auditif et visuel et jusqu'à un certain point à l'attention (Section 1.3).

Une hypothèse complémentaire à l'hypothèse principale est que, s'il existe effectivement un processus de liage audiovisuel préalable à la fusion, le liage est un processus dynamique, dans lequel l'état du système change au fur et à mesure qu'il enregistre des informations lui permettant d'estimer la cohérence audiovisuelle au cours du temps. C'est cette hypothèse qui va nous permettre de mettre en évidence l'existence de ce processus de liage. Nous proposons ici de mettre l'effet McGurk sous condition de variations contextuelles, et nous nous attendons à observer que les variations de contexte vont moduler la fusion audiovisuelle. La question centrale de ce mémoire est donc de savoir si l'effet McGurk résiste à des variations du contexte préalable, qui permettraient de lier/délier les flux auditif et visuel. Nous supposons que par manipulation du contexte, on peut produire un « décrochage » du lien audiovisuel, conduisant à une diminution de la fusion et donc à une réduction d'effet McGurk (Figure 37).

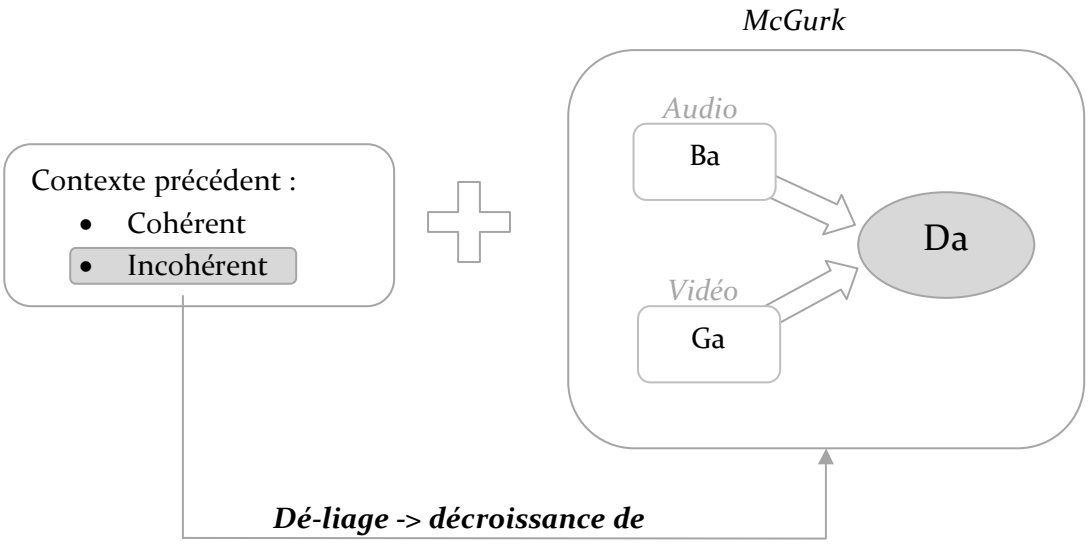


Figure 37 - Principe expérimental

4.3 Un programme expérimental

La suite de ce manuscrit est organisée en trois parties. Dans une première partie expérimentale, nous allons entreprendre de mettre en évidence ce processus de liage par modulation contextuelle de l'effet McGurk. Nous verrons qu'il nous faudra plusieurs expériences de mise en évidence et de contrôle avant de parvenir à une démonstration expérimentale claire.

La partie suivante s'attachera à tenter de mieux caractériser ce processus de liage, à en déterminer les contenus, les constantes de temps, les mécanismes possibles de réinitialisation.

Enfin, forts de ces avancées expérimentales, nous pourrons proposer une discussion générale sur nos résultats et sur le modèle cognitif sous-jacent, ainsi qu'un certain nombre de perspectives pour les travaux à venir.

Partie II
Mise en évidence comportementale de l'existence d'un processus de liage audiovisuel conditionnant la fusion

Chapitre 5. Mise en place de la méthodologie sur une expérience princeps

5.1 Introduction

Nous démarrons notre projet expérimental par la mise en place d'une expérience « princeps », qui doit à la fois nous servir de pilote, et de première mise en évidence de l'existence d'un processus du liage. Pour cela nous nous mettons dans les conditions les plus favorables, c'est-à-dire en utilisant des contextes fortement incohérents *vs*. des contextes fortement cohérents suivis par une cible. Cette séquence complète, constituée d'un contexte et d'une cible, est appelée dans la suite de ce document le « stimulus ».

Pour réaliser notre première expérience nous construisons donc deux types de contextes : « cohérent » et « incohérent ». Dans le cas cohérent le contexte consiste en une séquence de syllabes, présentées en modalité audiovisuelle : le sujet voit le visage du locuteur qui prononce des syllabes synchronisées avec les syllabes audio. Le contexte incohérent est constitué du même matériel audio, superposé avec la vidéo du même locuteur, qui prononce une série de phrases non contrôlées et non pas des syllabes. Ainsi, nous considérons ici une incohérence « maximale », puisque les contenus auditifs (syllabes) et visuels (phrase) sont totalement différents.

Cette expérience nous a permis d'introduire tout notre matériel expérimental, qui sera donc présenté ici avec un soin tout particulier. Elle a été aussi l'occasion de préciser notre paradigme et de mettre au point nos outils d'analyse. C'est l'ensemble de ces processus de mises au point méthodologiques que nous allons présenter dans ce chapitre, avant de décrire l'expérience proprement dite, et ses résultats, dans le chapitre suivant.

5.2 Paradigme expérimental

Le principe général de l'expérience consiste à présenter à des sujets des cibles congruentes « Ba » (audio « ba » + vidéo « ba ») ou incongruentes « McGurk » (audio « ba » + vidéo « ga » ,dont on attend qu'elles soient souvent perçues « da »), et ce dans un contexte cohérent ou incohérent, ne comportant aucun stimulus « ba », « ga » ou « da ». Cette approche est décrite par un paradigme « oddball » qui consiste à détecter des cibles qui sont singulières par rapport à un fond, c'est-à-dire par rapport au contenu du contexte.

La tâche est une tâche de « monitoring » avec un choix forcé, c'est-à-dire de détection en ligne de cibles perceptives « ba » ou « da », sans que les sujets ne sachent quand une cible est susceptible d'apparaître. Dans cette expérience nous remplaçons l'approche classique, où les stimuli sont présentés en séquence avec une réponse par stimulus, par un approche « runtime » ou « monitoring », où les sujets doivent détecter les cibles pour répondre, et pour cela doivent analyser les éléments un par un. Tous les stimuli sont mélangés aléatoirement dans un film.

Si nous faisons l'hypothèse qu'un mécanisme de liage est en jeu dans cette expérience, nous n'avons par contre pas d'idée précise a priori concernant ses constantes de temps, c'est-à-dire la durée nécessaire de contexte incohérent permettant de perturber l'effet McGurk, donc nous avons décidé de tester plusieurs durées de contexte (5, 10, 15, 20 syllabes), ce qui présente l'intérêt supplémentaire de rendre l'arrivée de la cible imprévisible. En effet, pour favoriser la mise en évidence d'un effet de liage et de mécanismes contextuels, le moment d'arrivée d'une cible ne doit pas être prédictible, pour que le sujet ne puisse pas l'anticiper. Une anticipation trop forte par le sujet risquerait de le conduire à négliger les séquences contextuelles pour se concentrer uniquement sur les instants d'arrivée de la cible, ce qui pourrait éliminer ou réduire l'effet du contexte, s'il existe. Donc la variabilité des durées favorise l'imprévisibilité des arrivées de cible.

Afin de pouvoir répondre, les sujets doivent catégoriser. Cela induit une certaine charge cognitive, et les sujets ont intérêt à utiliser toute l'information disponible pour effectuer cette tâche de détection/catégorisation. Lorsque le contexte est cohérent, ils effectuent en continu une tâche de catégorisation audio visuelle, qui est plus facile que la catégorisation audio seule. Lorsqu'il est incohérent, ils pourraient aussi s'appuyer sur le fait que l'information audiovisuelle est au moins partiellement cohérente pour détecter la cible (ils détecteraient alors la variation de cohérence audiovisuelle). Au total, les sujets sont conduits à être attentifs au mouvement des lèvres tout au long de la présentation pour effectuer la tâche.

Le schéma du principe expérimental est présenté sur la Figure 38.

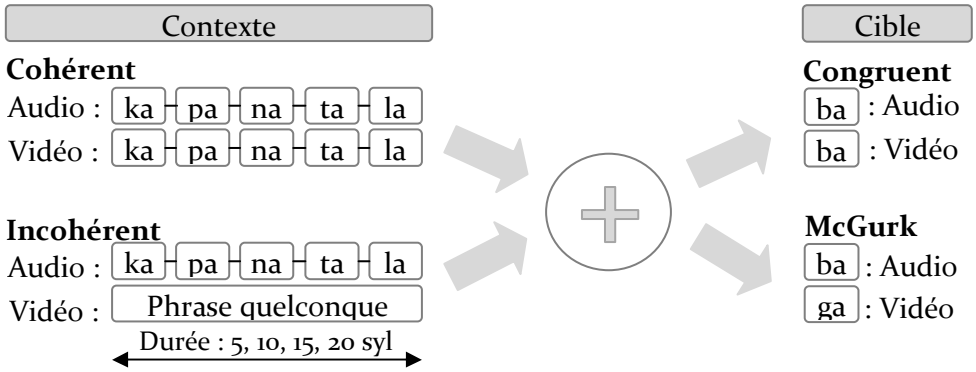


Figure 38 – Schéma de l'expérience 1

Les cibles « ba » ne présentent pas d'intérêt direct dans cette expérience, puisque nous prédisons qu'elles devraient être identifiées correctement « ba » quelque soit le contexte. Mais elles jouent un rôle de stimuli contrôles, qui nous permet d'une part d'établir un équilibre de la proportion entre les 2 types de réponses et d'autre part qui fournit une référence perceptive stable. Les cibles « Ba » nous permettent aussi de vérifier que les sujets sont attentifs pendant toute la durée de l'expérience et répondent réellement aux stimuli présentés.

Les stimuli « McGurk » sont au cœur de notre intérêt. Notre prédiction est qu'ils produisent moins de réponses de fusion « da » (et plus de réponses « auditives » « ba ») dans le cas de contexte incohérent. Les données empiriques montrent que, en français, l'effet McGurk apparaît en moyenne dans 35-50% des cas, tandis que les stimuli « ba » produisent des réponses « ba » dans presque 100% des cas (Cathiard et al, 2001). Pour équilibrer la fréquence attendue des réponses « ba » et « da », et pour optimiser le nombre de cibles «McGurk » qui concentrent notre intérêt, nous avons décidé de présenter les stimuli dans les proportions : ¼ des stimuli « Ba » et ¾ des stimuli « McGurk ».

Pour résumer, l'expérience est caractérisée par les éléments principaux suivants :

- Cible : ¾ de « McGurk » versus ¼ de « Ba »
- Contexte : cohérent versus incohérent
- Durée de contexte : 5, 10, 15, 20 syllabes.
- Choix forcé des réponses : soit « ba » soit « da ».

5.3 Préparation des matériaux expérimentaux

Le matériau de base pour la préparation des cibles et des contextes sera le même dans toutes nos expériences. Nous allons donc le décrire avec précision dans cette partie.

5.3.1 Enregistrement

Nous avons enregistré 80 séquences audiovisuelles de contextes de durée variée, se terminant toujours par la cible « ba » ou « ga ».

- Les 40 séquences destinées à produire le contexte audio pour toute l'expérience, et le contexte vidéo pour le cas de contexte cohérent, sont produites par des arrangements aléatoires de 13 syllabes françaises : « pa », « ta », « va », « fa », « za », « sa », « ka », « ra », « la », « ja », « cha », « ma », « na ». 20 séquences se terminent par une syllabe « ba » et 20 par une syllabe « ga ». La longueur des séquences (sans compter la syllabe finale « ba » ou « ga ») est 5, 10, 15, 20 syllabes, correspondant à des durées de l'ordre de 4, 7, 10 et 13 s. Les séquences ont été générées préalablement par script MATLAB, et étaient présentées au locuteur sur un écran de contrôle. Le locuteur devait répéter les séquences proposées, en laissant à chaque fois un silence court entre deux syllabes consécutives, de façon à fournir des points de montage acoustique simples. On évite ainsi la coarticulation entre deux syllabes consécutives, de façon à bien séparer chaque syllabe en terme acoustique et articulatoire. Des effets de coarticulation compliqueraient grandement le montage et l'interprétation. La cible ne doit pas être co-articulée de façon à bien contrôler le mouvement des lèvres et à la rendre séparable acoustiquement. Le débit est régulier et le moment d'arrivée de chaque syllabe est donc prédictible avec une fréquence syllabique moyenne d'environ 0,65 Hz.
- Les 40 séquences destinées à produire le contexte incohérent consistent en un flux de parole quelconque de durée 4, 7, 10, 13 secondes, se terminant dans la moitié des cas par une séquence « ba » et dans l'autre moitié par une séquence « ga ». Le locuteur devait parler librement sur le sujet de son choix, et au bout d'une durée correspondant à la

condition correspondante (4, 7, 10, 13 secondes), l'indication de la syllabe terminale apparaissait, indiquant au locuteur qu'il devait conclure en prononçant cette syllabe.

• Il y a donc au total 20 séquences cohérentes se terminant par « ba » et 20 par « ga », ainsi que 20 séquences incohérentes se terminant par « ba » et 20 autres par « ga ».

Toutes les séquences ont été prononcées par un locuteur français (J.-L. Schwartz) avec les lèvres maquillées en bleu et enregistrées dans le bloc expérimental (chambre sourde) du département Parole-Cognition de GIPSA-Lab (site Stendhal), spécialement équipé.

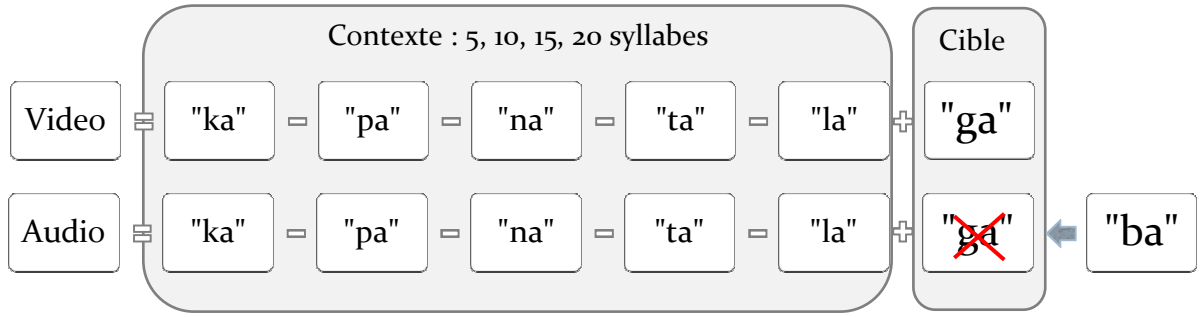


Figure 39 - Création des stimuli "McGurk" en contexte cohérent

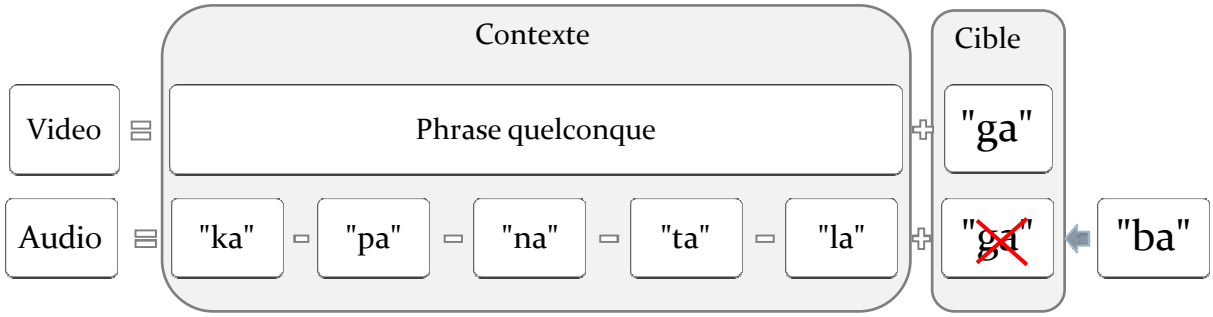


Figure 40- Création des stimuli "McGurk" en contexte incohérent

A partir de ces données, nous avons pu aisément monter tous les stimuli nécessaires, avec des cibles « Ba » ou« McGurk », dans un contexte cohérent ou incohérent, et avec une durée de contexte de 5, 10, 15 ou 20 syllabes.

Pour obtenir l'effet McGurk il s'agit de faire un montage audio en remplaçant le son « ga » par le son « ba », pris dans l'autre groupe des séquences avec « ba » à la fin (Figure 39, Figure 40). Tous les exemplaires de « ba » ont été utilisés pour assurer la variété des stimuli à tester.

5.3.2 Analyse et montage des données audio

Si on veut comparer la perception d'effet McGurk sous les différentes conditions de contexte, il est nécessaire de contrôler différentes propriétés des signaux, qui peuvent influencer sa perception, pour qu'elles soient plus ou moins équivalentes. Il faut aussi analyser les signaux pour faire le montage.

5.3.2.1 *Analyse*

Nous avons veillé à utiliser pour les deux conditions de contexte, cohérent et incohérent, les mêmes séquences syllabiques comme signal auditif.

D'abord nous avons filtré tous les séquences avec un filtre passe haut pour supprimer la composante continue de nos stimuli audio. Le filtrage est fait sous Praat par le filtre « PassHann Band » avec une fréquence de coupure à 20Hz.

Pour remplacer « ba » par « ga » (Figure 39, Figure 40) il nous faut des points de montage précis. Un point de repère crucial est le début de la consonne et plus précisément le burst (« explosion acoustique » correspondant au début du mouvement d'ouverture de la mâchoire, et de la langue pour « ga », ou des lèvres pour « ba ») (Figure 23). Donc nous choisissons cet instant d'apparition du burst comme point de montage pivot.

A l'aide du logiciel MATLAB 7.6.0, sur chaque piste audio on marque manuellement la cible avec 3 marques (Figure 41) :

1. Début acoustique de la cible
2. Burst
3. Fin acoustique de la cible

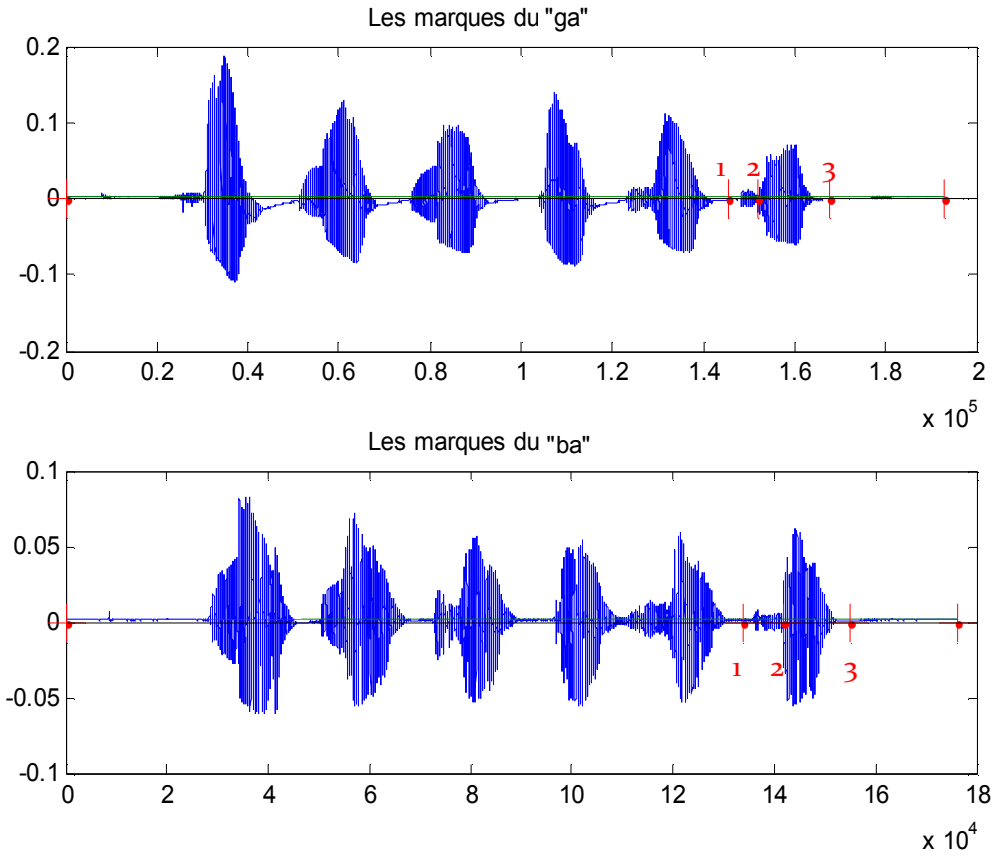


Figure 41 - Le marquage de la cible. 1 – Début ; 2 – Burst ; 3 – Fin

5.3.2.2 *Montage*

A partir de ces 3 marques le montage auditif peut être fait correctement de façon automatique. Le remplacement de la syllabe « ga » par la syllabe « ba » est fait en respectant la coïncidence des marques (2) (instant du burst) de chaque piste (Figure 42). Chaque fois nous faisons le contrôle que le signal d'une cible ne déborde pas sur le signal d'une syllabe précédente.

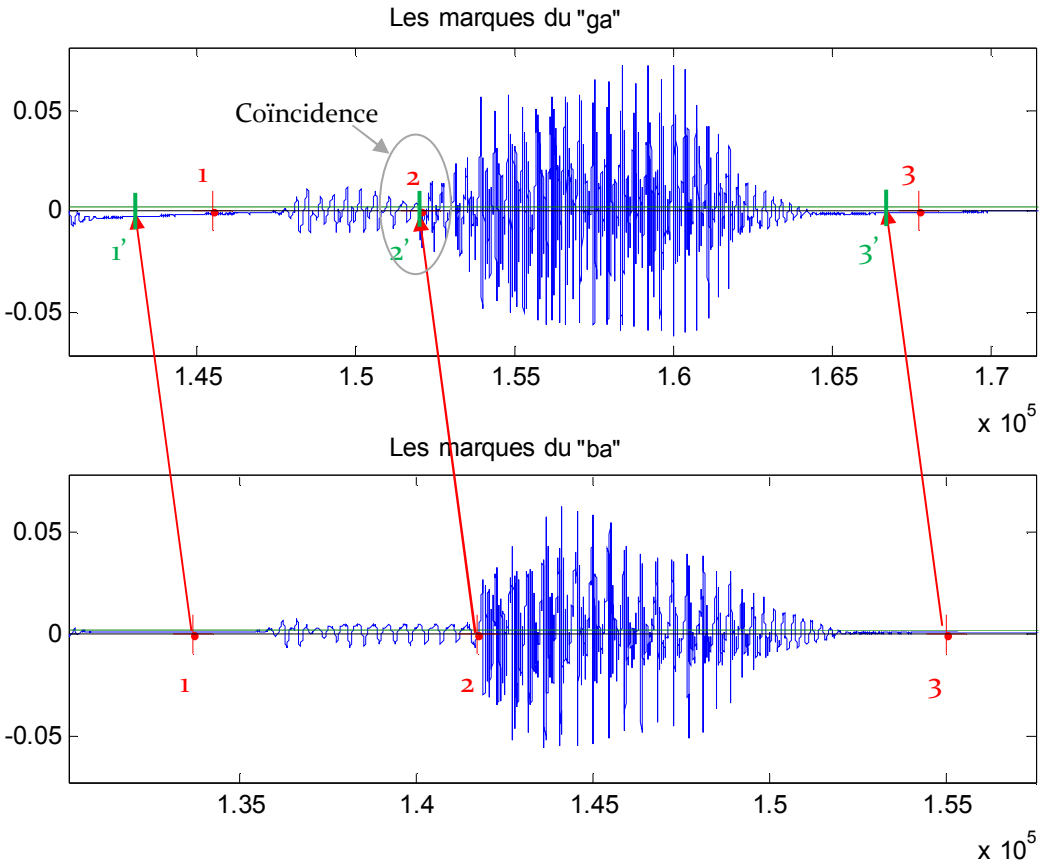


Figure 42 - Remplacement de la syllabe "ga" par "ba"

5.3.2.3 *Normalisation*

Pendant l'enregistrement le locuteur a prononcé les séquences avec des intensités variables. Or, l'effet McGurk est très sensible au volume sonore (Colin et al, 2002).Il est donc nécessaire de contrôler le volume des stimuli cible. Pour cela nous avons fait une normalisation d'intensité sonore du « ba » à partir du « ga » original (Figure 43). La piste audio est ainsi prête.

$$signal\ normalisé = \frac{(std\ d'échantillons\ \text{"Ga"}) \times (tableau\ d'échantillons\ \text{"Ba"})}{std\ d'échantillons\ \text{"Ba"}}$$

L'intensité de la cible est ainsi comparable à celle de son contexte audio, car l'intensité de la dernière syllabe a été contrôlée par le locuteur (il n'y a pas de « chute » à la fin de chaque séquence).

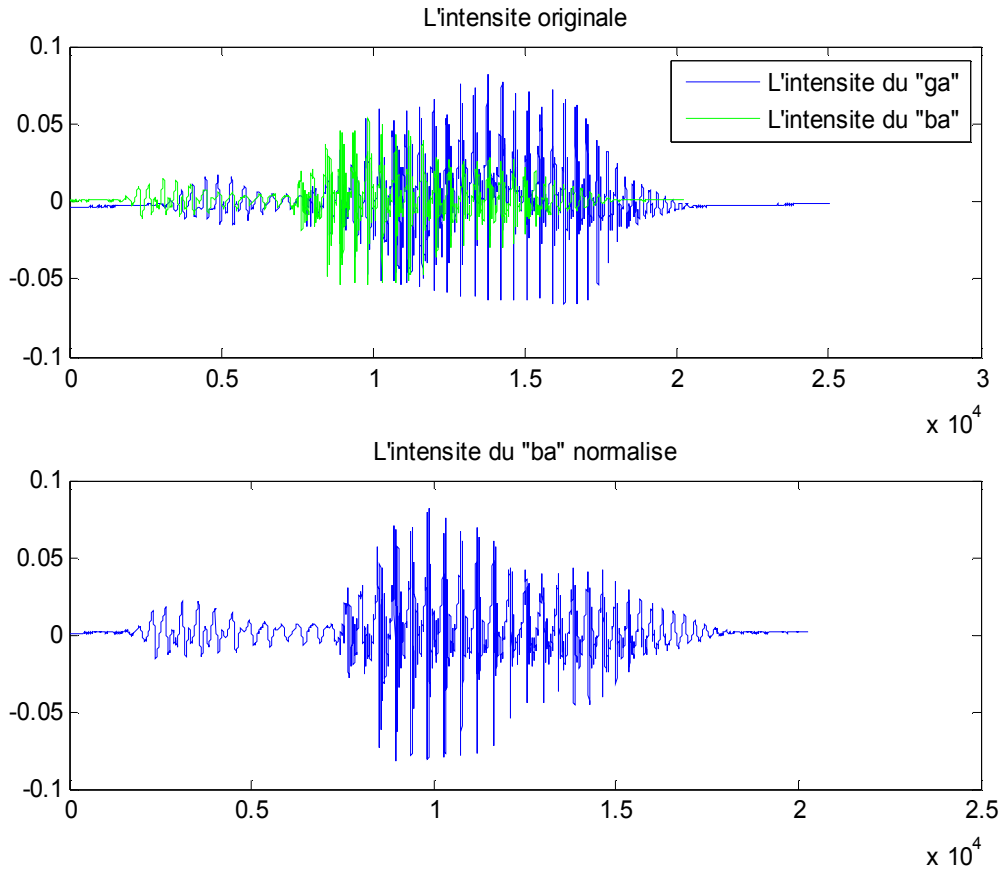


Figure 43 - Normalisation d'intensité de la syllabe "ba" à partir du « ga »

5.3.3 Analyse et montage des données vidéo

Il est important de signaler que toute cette première expérience a été construite dans le but d'éviter tout montage vidéo à l'intérieur des stimuli, donc de ne pas produire de discontinuité perceptive. Ceci a engendré toute une série de complications de montage assez lourdes que nous expliquons systématiquement par la suite.

5.3.3.1 Analyse et contrôle de la dynamique labiale

Comme on a des conditions de production très différentes entre syllabes cibles visuelles en contexte cohérent (syllabes prononcées en fin d'une séquence de syllabes, avec chaque fois le phonème précédent « a », et une élocution très contrôlée et répétitive) et incohérent (syllabes prononcées en fin d'un discours libre, avec contexte antérieur variable, et élocution plus naturelle), il faut faire des mesures physiques sur les signaux, pour choisir des stimuli aux propriétés visuelles comparables. Il est important de noter à ce stade que la volonté de conserver une continuité totale de flux vidéo rend par nature impossible un contrôle parfait des stimuli : la composante visuelle d'une cible en contexte cohérent ne peut pas être la même

que la composante visuelle d'une cible en contexte cohérent : nous ne pouvons que veiller à ce qu'elle en soit aussi proche que possible. Nous reviendrons sur ce point au chapitre 9.

Pour tenter d'obtenir des stimuli aux propriétés visuelles proches d'un contexte à l'autre, il nous faut donc pouvoir mesurer les propriétés visuelles des signaux, et notamment l'amplitude, la durée et la vitesse du mouvement d'ouverture des lèvres.

Pour chaque stimulus, nous avons donc mesuré la dynamique labiale à l'aide du logiciel « Tacle » (Lallouache, 1990), qui permet de traiter des séries d'images pour en extraire les paramètres labiaux. Voici l'algorithme:

1. Application du chroma-key sur l'image, de manière à seuiller en noir les zones teintes en bleues (les lèvres du locuteur)
2. Filtrage des images obtenues via un filtre médian, puis détection de contours sur ces « masses noires » et enfin extraction des paramètres demandés (Figure 44, Figure 45) à partir de ces contours. Dans notre cas on s'intéresse uniquement au paramètre d'ouverture des lèvres B : distance verticale entre les contours internes des lèvres.

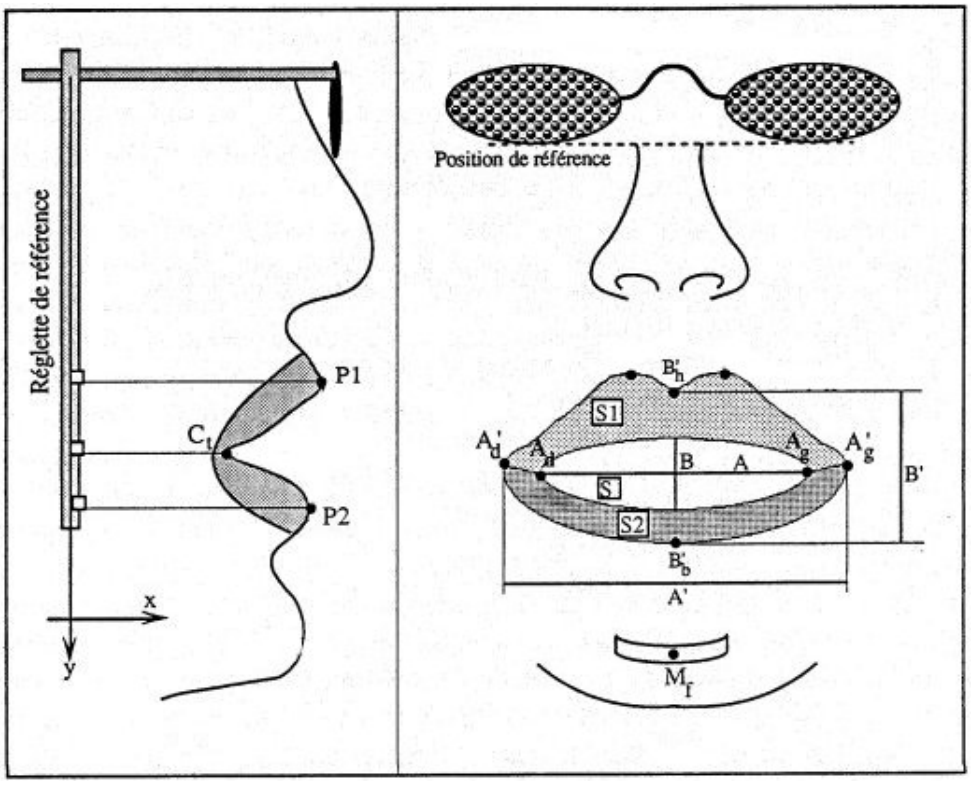


Figure 44 - Paramètres labiaux mesurés par le logiciel Tacle

a)

b)

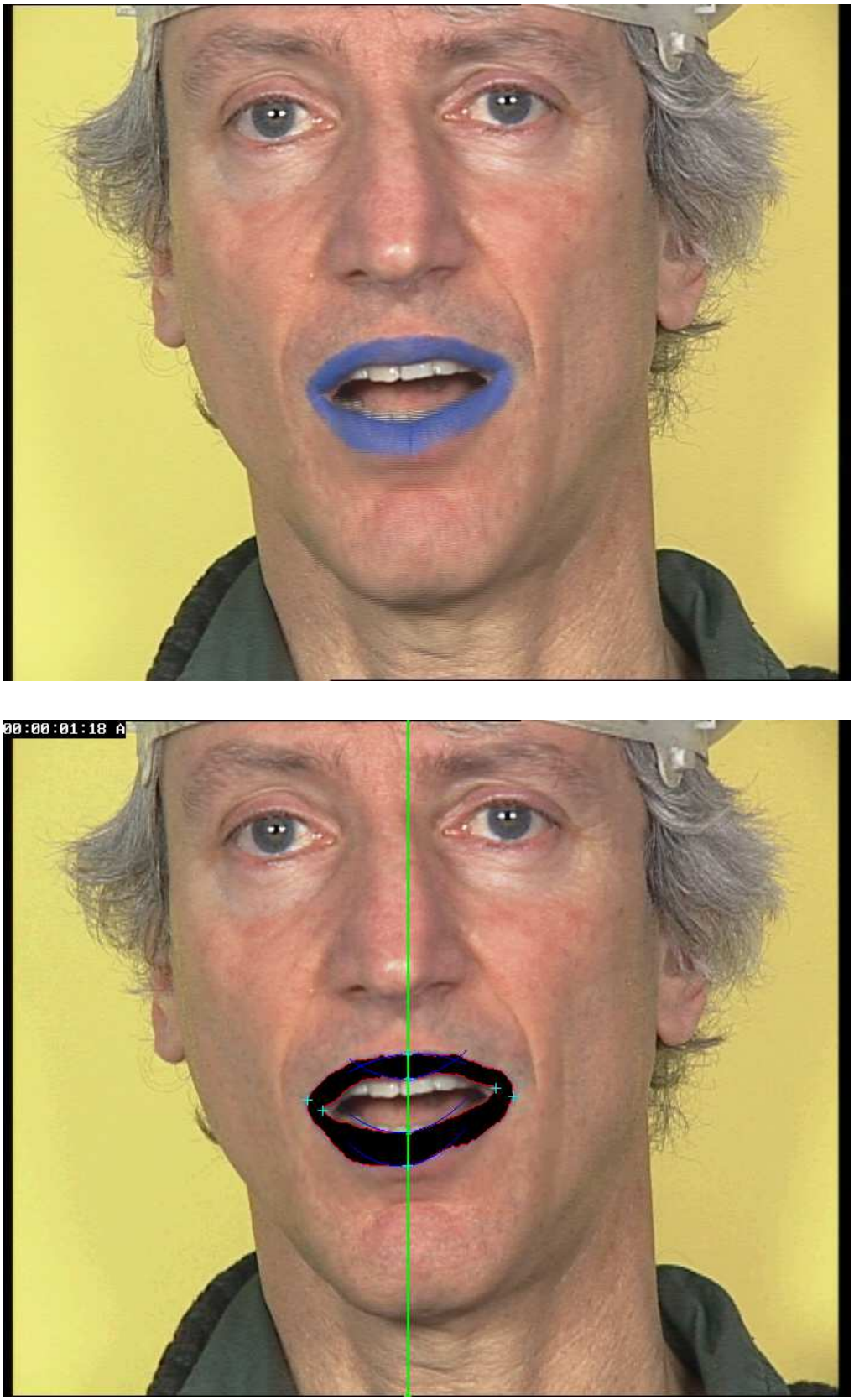
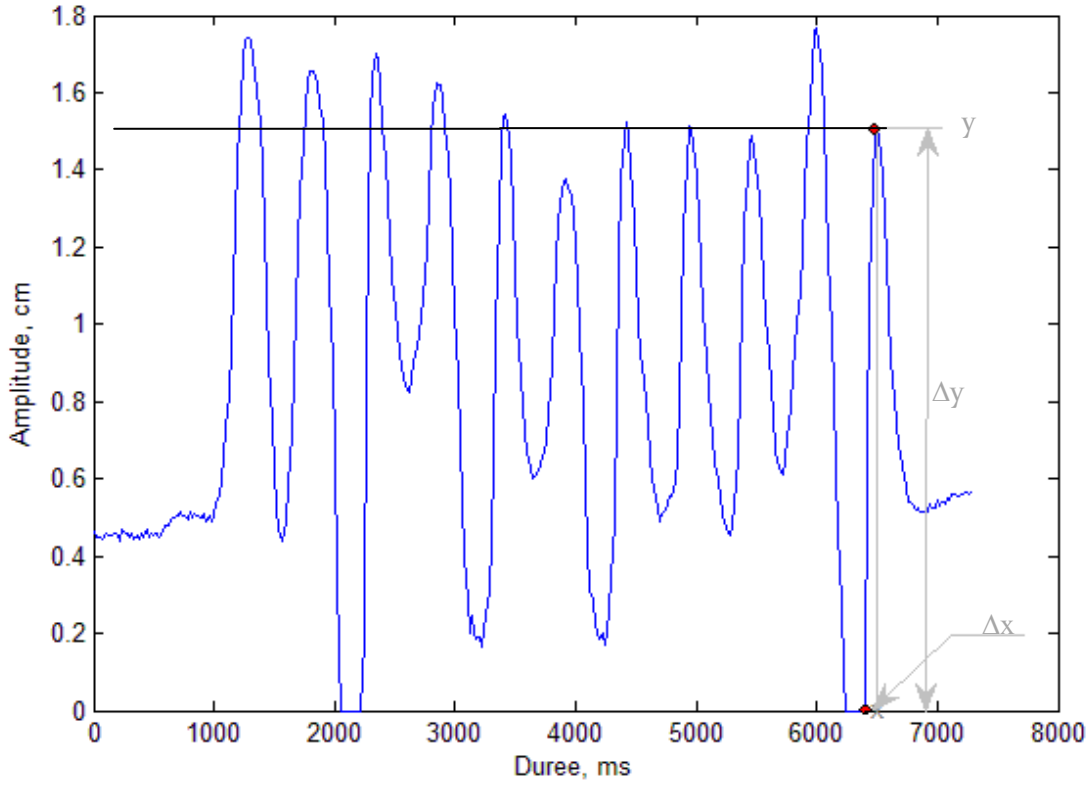


Figure 45 - Mesures d'ouverture des lèvres avec Tacle. a) image originale, b) image traitée.

a)



b)

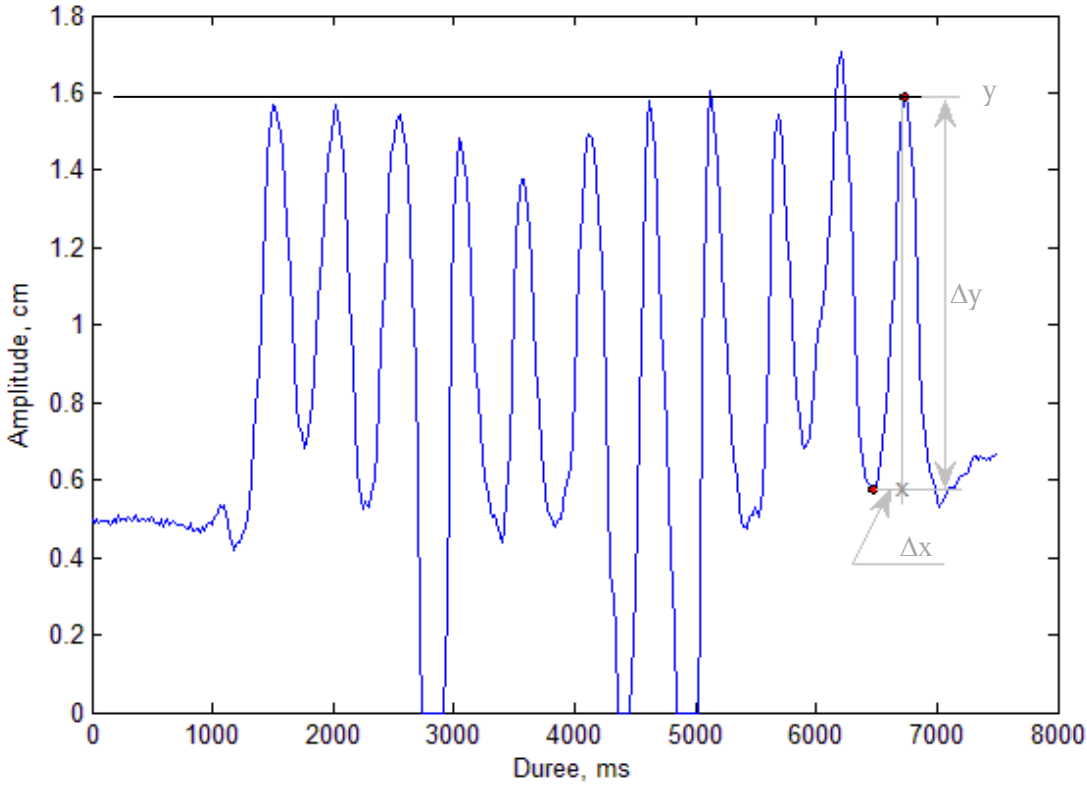


Figure 46 - Mouvement des lèvres pour la durée de contexte 10 syllabes. a) stimulus « ba », b) stimulus « ga »

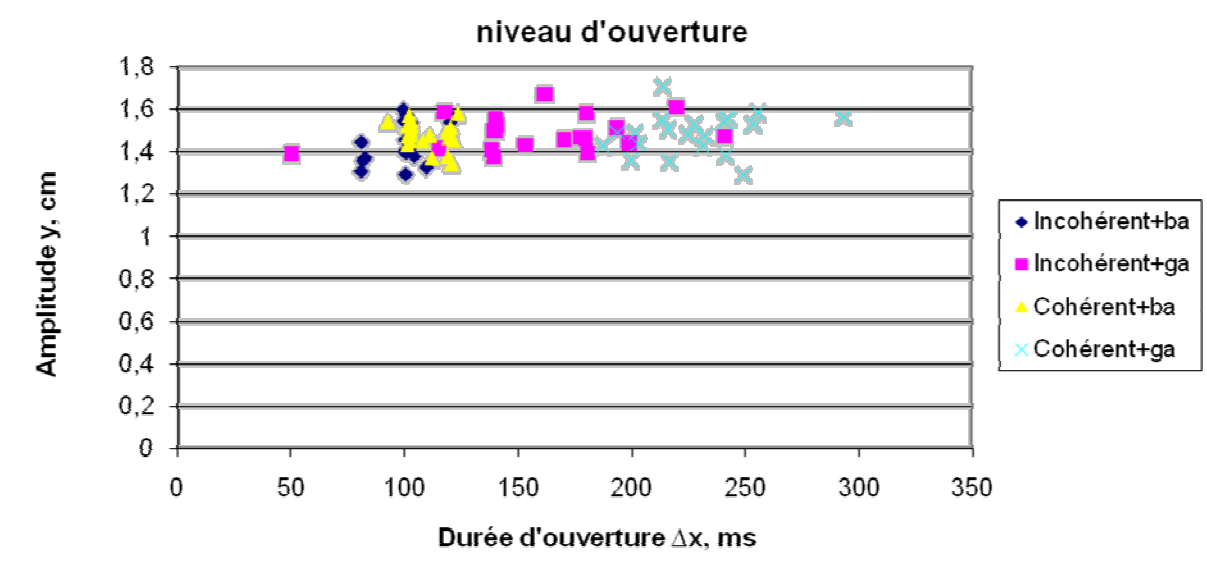


Figure 47 - L'ensemble des valeurs d'ouverture maximale des lèvres en fonction de la durée d'ouverture pour tous les stimuli enregistrés

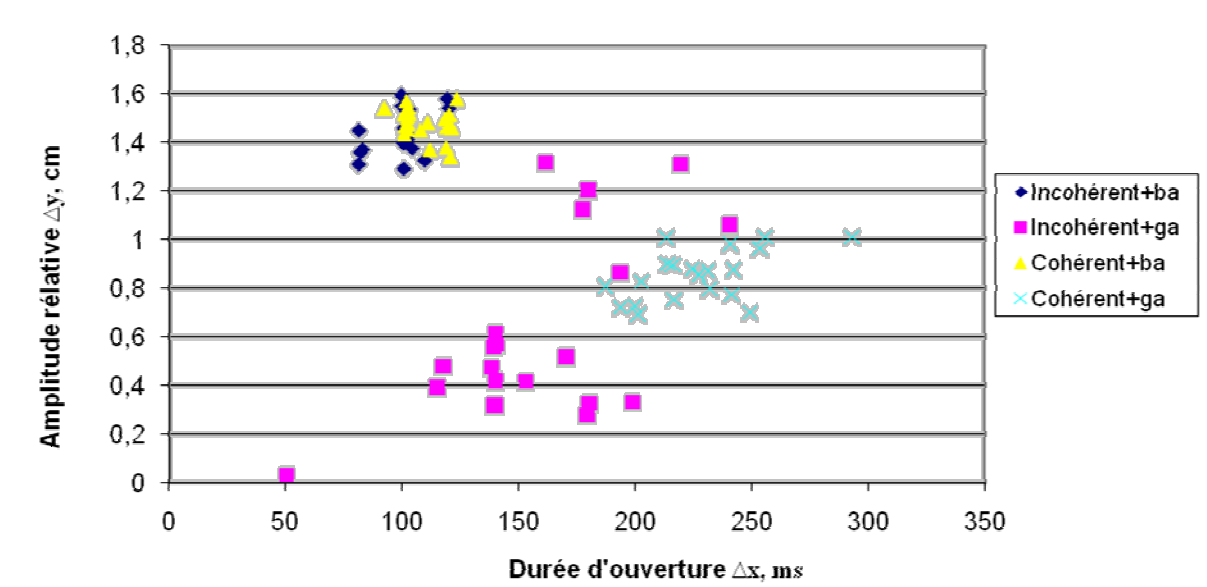


Figure 48 - L'ensemble des valeurs d'ouverture relative des lèvres en fonction de la durée d'ouverture pour tous les stimuli enregistrés

Pour chaque image dans le film nous obtenons ainsi la valeur d'ouverture des lèvres. En combinant ces valeurs avec l'information temporelle on peut suivre la trajectoire d'ouverture labiale : la dernière ascension correspond à la trajectoire d'ouverture pour le stimulus cible « ba » ou « ga » (Figure 46).

Nous avons mesuré trois paramètres sur cette trajectoire : l'amplitude maximale correspondant à la plus grande ouverture des lèvres de la voyelle finale (y), la dynamique d'amplitude (ouverture relative) entre la consonne cible et la voyelle finale (Δy) et la vitesse moyenne (Δy/Δx) d'ouverture des lèvres pendant le temps Δx (Figure 46), et ce pour chaque séquence. L'ensemble de ces valeurs est représenté sur la Figure 47, la Figure 48 et la Figure 49.

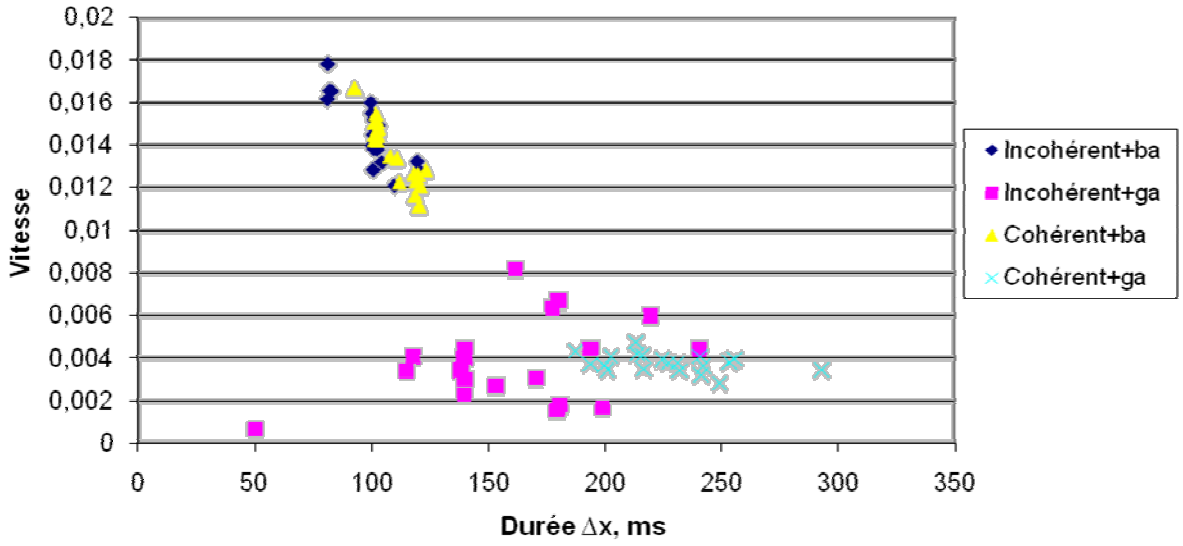


Figure 49 - L'ensemble des valeurs de vitesse d'ouverture des lèvres en fonction de la durée d'ouverture pour tous les stimuli enregistrés

5.3.3.1.1 Sélection des stimuli vidéo

Sur la Figure 47, on voit que l'ouverture complète des lèvres est similaire pour tous les cas. Cela souligne la constance d'une caractéristique de production. L'amplitude relative (Figure 48), c'est-à-dire le mouvement des lèvres, de même que la vitesse moyenne (Figure 49), dépend du stimulus cible (« ba » vs. « ga ») mais aussi du contexte dans le cas de la cible « ga ». La dispersion des données est faible pour les « ba », mais plus forte pour les « ga », surtout en contexte incohérent (carrés roses). Ces derniers sont très dispersés ; dans la majorité des cas, la valeur d'ouverture relative est petite et les lèvres bougent moins vite.

Prioritairement, il nous a semblé important de contrôler le mieux possible les propriétés visuelles des stimuli « ga » en contexte incohérent, de façon à les rendre globalement le plus proche possible de celles des stimuli « ga » en contexte cohérent. Pour ce faire, nous avons cherché à égaliser les moyennes des valeurs d'ouverture relative des lèvres pour le « ga ». Nous avons divisé le groupe des stimuli « ga » incohérents en 2 sous-groupes, dépendants du niveau d'ouverture des lèvres : Δy>0.8cm (groupe incohérent+) et Δy<0.8cm (groupe incohérent-) (Figure 50), cette valeur de 0.8cm étant à peu près la moyenne des valeurs dans le contexte cohérent. En fonction du facteur durée nous avons sélectionné les 6 stimuli du groupe à dynamique forte, et 8 stimuli du groupe à dynamique faible. Les stimuli à éliminer sont marqués par le cercle rouge.

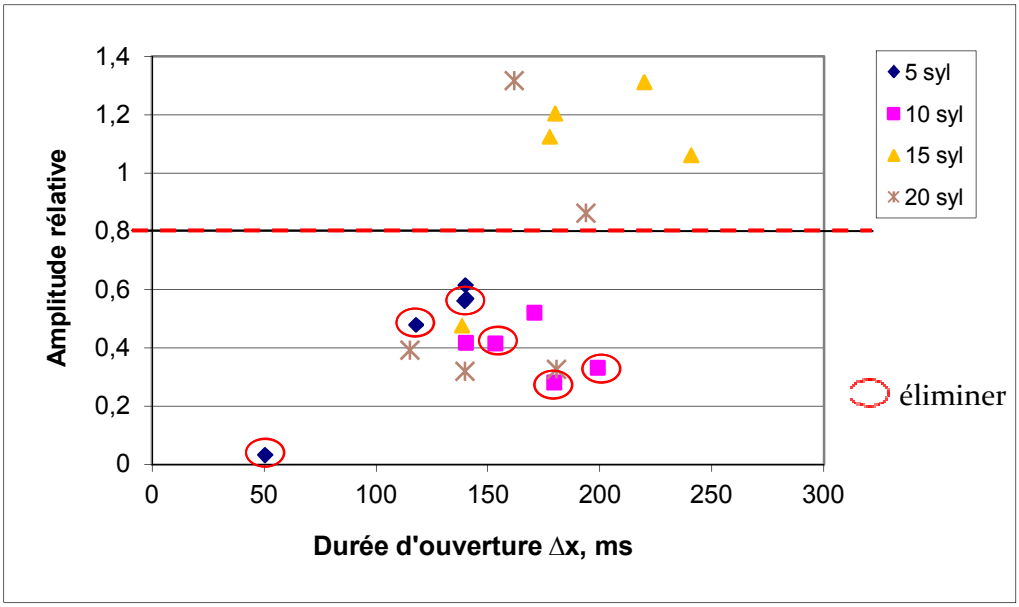


Figure 50 - L'ensemble des valeurs d'ouverture relative des lèvres pour tous les stimuli incohérents (« ga ») en fonction de la durée

Sous - groupe	Durée, syl	Nombre dans sous-groupe	Nombre à obtenir
Δy>0.8cm	5	0	2
	10	0	2
	15	4	2
	20	2	2
Δy<0.8cm	5	5	2
	10	5	2
	15	1	2
	20	3	2

Figure 51 - Regroupement de séquences vidéo à partir d'information d'ouverture des lèvres. (Les nombres indiqués sont les nombres de séquences que l'on va créer à partir des séquences existantes)

On obtient ainsi un ensemble de stimuli « ga » en contexte incohérent pour le sous-groupe incohérent+ avec une durée moyenne de 195 ms (std 29 ms) et une amplitude relative moyenne de 1.15 cm (std 0.17 cm) et le sous-groupe incohérent- avec une durée moyenne de 145 ms (std 21 ms) et une amplitude relative moyenne de 0.45 cm (std 0.11 cm). Globalement, les stimuli « ga » incohérents sont caractérisés par une durée moyenne de 167 ms (std 35 ms) et une amplitude relative moyenne de 0.75 cm (std 0.38 cm) En comparant aux mêmes valeurs pour les stimuli cohérents (durée 227 ms, std 25 ms et amplitude moyenne 0.85 cm, std 0.11 cm) nous obtenons une similitude d'ensemble assez rassurante. De plus, nous disposons parmi les stimuli en contexte incohérent, de stimuli qui devraient avoir des propriétés d'articulation respectivement plus (groupe incohérent+) et moins (groupe incohérent-) marquées que les stimuli en contrôle cohérent, ce qui nous servira pour contrôler a posteriori nos effets.

Nous souhaitons veiller à ne pas trop dupliquer les stimuli pour maintenir une variabilité suffisante des séquences. Pour ce faire, nous utilisons quelques séquences longues que nous redécoupons en leur début pour obtenir des séquences plus courtes et nous réduisons le nombre des stimuli de 20 à 16. Le tableau de toutes ces manipulations est représenté sur la Figure 51.

5.3.4 Montage audiovisuel

Dans notre expérience il faut préparer 4 types de stimuli :

- Cohérent « Ba »[2]
- Cohérent « McGurk »
- Incohérent « Ba »
- Incohérent « McGurk »

Pour les créer, il faut combiner les séquences audio avec des pistes vidéo différentes. Ces correspondances sont présentées dans le Tableau 1.

Tableau 1 - Correspondance entre les séquences audio et vidéo pour chaque type de stimuli

	Vidéo	**Audio**
Cohérent « ba »	Syllabes + « ba »	Syllabes + « ba »
Cohérent « McGurk »	Syllabes + « ga »	Syllabes + « ba » remplacé
Incohérent « ba »	Phrase + « ba »	Syllabes + « ba »
Incohérent « McGurk »	Phrase + « ga »	Syllabes + « ba » remplacé

Nous faisons le montage audiovisuel pour chaque séquence de manière systématique à partir du nom du fichier, qui est décrit selon le principe présenté Figure 52. Par exemple, si on veut créer le stimulus cohérent « McGurk » de durée 10 syllabes, on prend la séquence vidéo nommée syl10_ga_1 qui est assemblée avec la piste audio nommée syl10_ga_1. Le principe est le même pour les séquences incohérentes en combinant les séquences vidéo de durée 4, 7, 10 et 13 secondes avec les séquences audio respectivement de 5, 10, 15, et 20 syllabes.

Les nombres 4, 7, 10, 13 représentent la durée de la séquence contextuelle en secondes. Nous avons enregistré ces séquences avec une durée un peu plus longue que celle des séquences syllabiques, de façon à avoir une durée de la piste vidéo suffisamment longue pour le montage. Pour finaliser, nous découpons le début des séquences vidéo pour synchroniser le début du son avec le début de l'image.

Nos stimuli doivent être présentés en ligne, donc nous rassemblons des séquences différentes dans le film de façon aléatoire.

[2] Dans la suite, nous désignerons par « Ba » (avec une majuscule initiale) les stimuli cohérents « ba », pour les distinguer des réponses « ba » ou « da »

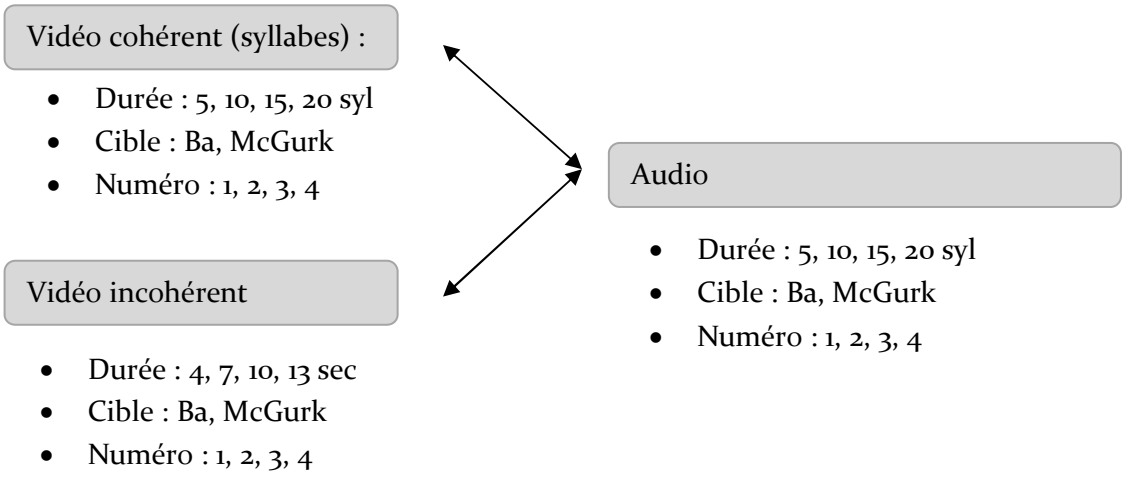


Figure 52 - Organisation des fichiers audio et vidéo et dénomination du fichier résultant dans le montage audiovisuel systématique

Le montage audiovisuel est fait de la manière suivante :
Création de la liste des séquences audio pour toute l'expérience (128 stimuli).

- Randomisation de cette liste.
- Répartition de la liste sur 4 blocs, avec 32 stimuli par bloc.
- Copie des fichiers selon l'ordre de la liste. Pour chaque fichier audio on copie les images vidéo correspondantes. Pendant les premiers essais, nous avons constaté que la pause entre la fin d'un stimulus et le début de la séquence suivante semblait faible et de nature à perturber le sujet au moment de sa réponse. Donc, pour augmenter le temps de la pause après chaque stimulus, nous avons recopié 12 fois la dernière image et ajouté un silence de 480 ms à la fin de chaque séquence.
- Pour un passage continu d'une séquence à l'autre nous avons fusionné 5 images adjacentes entre les deux séquences.
- Le montage final des films a été fait à l'aide du logiciel « Adobe Premiere 6 ».

5.4 Passation de l'expérience

5.4.1 Organisation du test

Nous avons préparé 16 stimuli originaux de chaque type et les avons combinés dans l'expérience avec les proportions : ¾ de « McGurk » versus ¼ de « Ba ». Au total on a présenté 128 stimuli avec le nombre de stimuli de chaque type indiqué dans le (Tableau 2). Les stimuli de type « McGurk« ont été répétés 3 fois.

Tableau 2 - Nombre de stimuli présentés dans l'expérience

A\V	Cohérent « Ba »	Cohérent « McGurk »	Incohérent « Ba »	Incohérent « McGurk »
« ba »	16 (1/8)	48 (3/8)	16 (1/8)	48 (3/8)

Pour avoir un nombre de stimuli suffisant pour l'analyse tout en limitant la fatigue des sujets, nous avons décomposé l'expérience sur 4 blocs avec des pauses entre deux blocs consécutifs. L'ordre de passage des blocs est varié selon les sujets. Chaque bloc consiste en 32 stimuli, choisis de façon aléatoire, parmi tous les stimuli (les blocs contiennent donc à la fois des stimuli à contexte cohérent et incohérent, et de toutes durées de contexte).

En résumé, l'expérience consiste en :

- 128 stimuli en total
- 4 blocs de 32 stimuli, répartis aléatoirement
- 16 stimuli originaux de chaque type
- 4 stimuli par type et par durée
- 2 types de réponses possibles « ba » ou « da » en ligne, choix forcé

5.4.2 Consignes et exécution de l'expérience

Tous les tests ont été effectués dans la chambre sourde du laboratoire. Les sujets étaient assis face à une table avec un moniteur devant eux et ils écoutaient le son au casque. Le niveau sonore était fixé pour l'expérience à une intensité sonore pour une audition confortable (environ 58 dB SPL, d'après une évaluation que nous avons faite a posteriori avec un appareillage (Brüel & Kjaer, 2013).

La présentation d'expérience a été effectuée pour les premières expériences à l'aide d'un logiciel développé sous « JAVA » (Figure 53). Ce logiciel ne permettait pas malheureusement de mesurer des temps de réponse avec une précision très grande. Nous avons changé de système en cours de travail pour passer à une passation sous logiciel Presentation© (Version 0.70, www.neurobs.com), permettant de mesurer des temps de réponse. Nous préciserons à chaque expérience quel système de passation est utilisé.

Avant l'expérience, les sujets lisaient une consigne du type :

« *Avant l'expérience elle-même, vous pouvez passer quelques essais.*

L'expérience consiste en 4 parties, d'environ 5 min chacune. Après chaque partie on fait une pause.

Pendant l'expérience vous ne devez pas modifier la distance à l'écran ni le niveau du volume sonore du casque.

Pendant chaque partie vous devez être très attentif et en permanence regarder le locuteur sur l'écran. Chaque fois que vous entendez le son "ba" ou "da" vous devez appuyer immédiatement sur le bouton correspondant, indiqué par le présentateur au début de l'expérience. »

Il s'agit donc bien d'une tâche de monitoring en ligne, dans laquelle les sujets ne savent pas quand peuvent apparaître les stimuli cible « Ba » ou « McGurk » et peuvent répondre à tout moment à l'intérieur des films.

Au début, le sujet passe 4 essais, consistant en des séquences identiques (cohérentes et incohérentes) de durée différente, mais ne comprenant que des stimuli « ba » à la fin.

Les boutons de réponse sont spécifiés par des marques « ba » et « da ». Pour la moitié des sujets le bouton « ba » est à gauche, sur la touche « Entrée » et le bouton « ga » est à droite, sur la touche « Espace». Pour l'autre moitié les boutons sont inversés. Les mêmes marques sont également affichées sur l'écran, de chaque côté de l'écran au niveau des yeux.

La durée d'expérience est environ 25 minutes. Entre les blocs le sujet peut faire une pause de durée arbitraire.

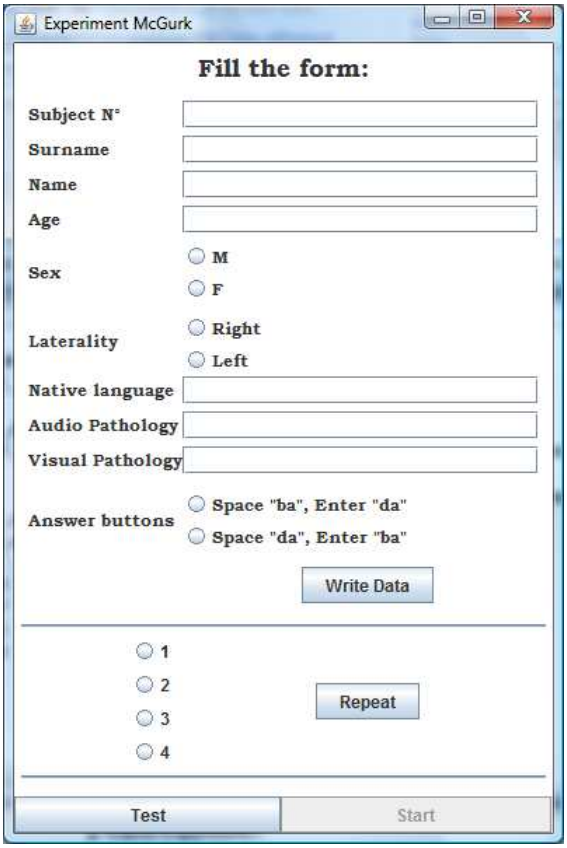


Figure 53 - Logiciel d'exécution d'expérience

5.5 Méthode d'analyse des résultats

5.5.1 Détermination d'une zone de réponses valides

Pendant l'expérience les stimuli sont fournis en ligne, et le sujet peut répondre à chaque instant, qu'il y ait ou non la présence d'une cible perceptive « ba » ou « da ». Il peut donc se produire deux types d'erreurs : la présence d'une réponse « ba » ou « da » en l'absence de cible (stimulus « Ba » ou « McGurk ») ou l'absence de réponse à une cible. Pour traiter correctement les réponses nous avons mis en place la méthodologie suivante.

1. Pour une cible on compte les réponses qui sont apparues après sa présentation, mais avant la cible suivante.
2. On limite la validité temporelle de réponse par un seuil, qui est égal à la durée d'une séquence minimale (3500ms). Ce seuil de 3500 ms a été choisi par inspection de l'histogramme temporel de toutes les réponses (correctes et incorrectes), données par tous les sujets dans l'expérience princeps (Figure 54), sur lequel on voit que la plupart des réponses sont inférieures au seuil.
3. Pour déterminer les réponses incorrectes, nous distinguons 2 types d'erreurs : « Fausses alarmes » et « Absence de réponse » (Figure 55). Toutes les réponses

parvenant à des délais supérieurs au seuil sont considérées comme « fausses alarmes ». S'il n'y a pas de réponse dans l'intervalle entre la cible et le seuil, on compte une « Absence de réponse » pour cette cible. S'il y a plusieurs réponses dans cet intervalle, on fait une vérification de l'identité des réponses. Si elles sont identiques, nous ne prenons que la première d'entre elles et la comptons comme une réponse normale, sinon nous les éliminons toutes, et considérons une « absence de réponse » pour la cible.

Cela nous permet de préserver un maximum de réponses valides et d'éliminer les réponses fortuites.

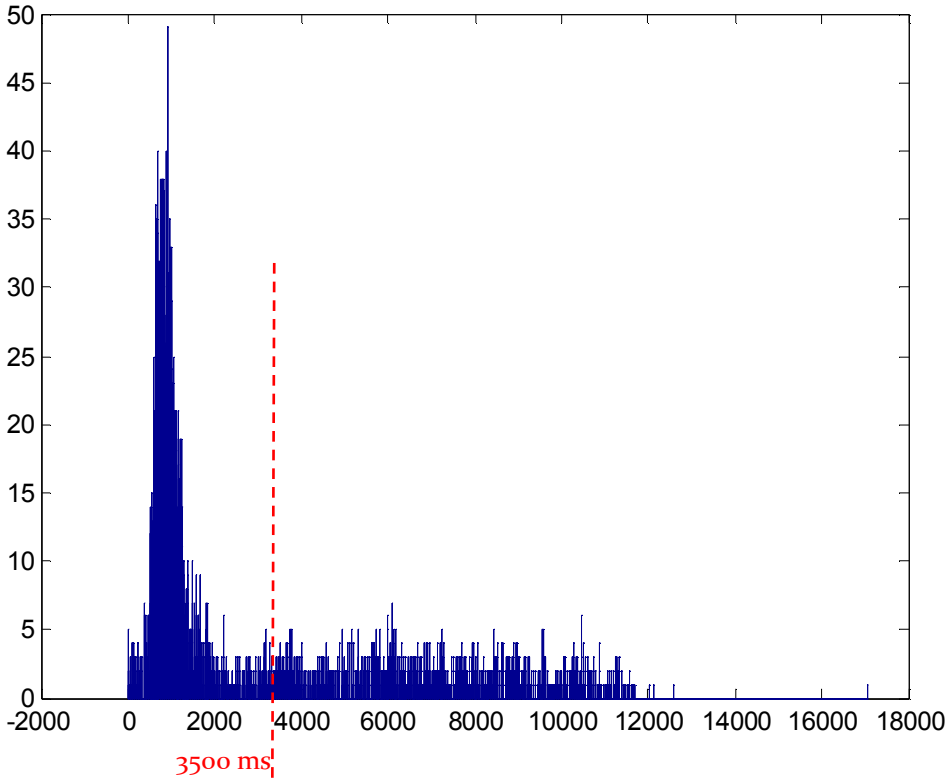


Figure 54 - Histogramme des durées des réponses pour tous les sujets

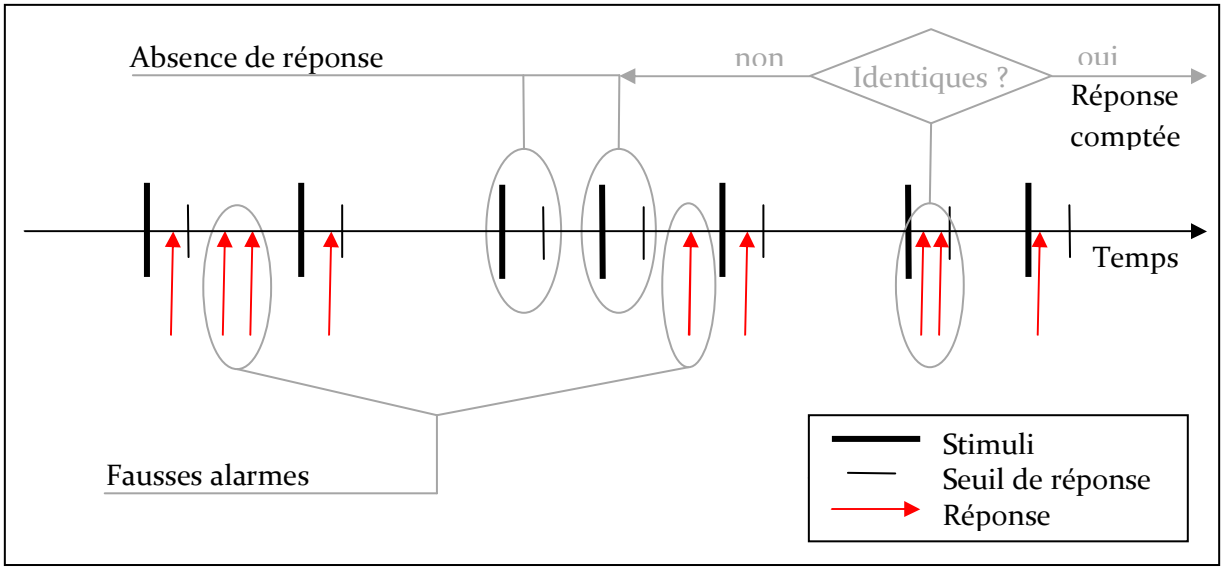


Figure 55 - Traitement des réponses

5.5.2 Analyse des réponses

Les analyses principales concernent les scores de réponses « ba » et « da », en écartant les fausses alarmes et absences de réponse au sens précisé précédemment. L'indicateur principal sur lequel portent nos analyses est le rapport des réponses « ba »/(« ba » + « da »). Pour assurer la gaussianité de cette variable en vue des analyses de variance adéquates nous transformons systématiquement les proportions des réponses « ba »/(« ba » + « da ») en asin(sqrt(proportions)). Les analyses ANOVA à mesures répétées sont faites sur ces données transformées, en utilisant le logiciel *IBM SPSS Statistics 19 © IBM*. En cas de violation de l'hypothèse de sphéricité, nous utilisons une correction de Huynh-Feldt. Les tests Post Hoc avec correction de Bonferroni sont faits sous *Matlab 7.11 © The MathWorks, Inc* (logiciels « *anovan* » puis « *multcompare* ») car les tests Post Hoc dans le cas de l'ANOVA à mesures répétés sous SPSS ne sont pas disponibles sur les termes d'interaction. Dans les analyses faites sous Matlab nous ajoutons systématiquement le facteur sujet comme un facteur aléatoire, ce qui fournit une analyse équivalente à l'ANOVA à mesures répétées.

Les figures avec des données moyennes sont crées en faisant la transformation inverse $(\sin(\text{valeurs moyens}))^2$ des valeurs moyennes estimées et des erreurs standards.

Nous avons également systématiquement analysé les scores « ba »/(toutes réponses), « da »/(toutes réponses) et (réponses absentes)/(toutes réponses). Les analyses font en général ressortir les mêmes résultats sur tous ces scores.

5.5.3 Analyse des temps de réaction

Lorsque nous sommes passés sur un logiciel adéquat (Presentation), nous avons également estimé les temps de réponse (qui étaient estimés avec le premier logiciel avec une précision suffisante pour construire les courbes de la Figure 54, mais pas pour estimer des différences fines de temps de réponse). La mesure de temps de réponse est estimée à partir du

repère temporel du burst de la plosive en début de syllabe (Figure 43, point 2) et déterminée par l'appui du bouton réponse au clavier par le sujet.

Pour assurer la gaussianité de la variable « temps de réponse », en vue des analyses de variance adéquates, nous effectuons une transformation logarithmique.

5.6 Conclusion

La mise en place de l'expérience initiale a demandé, on le voit, de nombreux ajustements et contrôles, de l'enregistrement au découpage et au montage des stimuli puis à la réalisation de films complets pour l'expérience proprement dite. Cette phase, lourde, était nécessaire étant donné le paradigme de monitoring en ligne que nous avons choisi, et en essayant de limiter au maximum les points de montage vidéo, tout en conservant à la fois des stimuli aussi naturels que possible, un niveau raisonnable d'imprédictibilité, et en essayant de contrôler au mieux l'information auditive et visuelle fournie. Nous allons maintenant aborder la première expérience proprement dite. Nous verrons plus tard que la succession des expériences nous conduira à revenir sur ces questions de contrôle des stimuli et de nature des processus de montage.

Chapitre 6. Expérience 1 : première mise en évidence d'un effet de contexte

6.1 Objectifs et hypothèses

Nous l'avons vu précédemment, cette expérience princeps vise à présenter une première mise en évidence claire d'un mécanisme de déliage, en se mettant dans les conditions les plus favorables possibles, c'est-à-dire en contrastant des contextes très cohérents vs. très incohérents. Pour ce faire nous avons enregistré pour un locuteur français des séquences audiovisuelles de syllabes CV et des séquences de parole libre. Le contexte cohérent est fourni par les séquences de syllabes CV cohérentes en audio et en vidéo et le contexte incohérent consiste à monter le contenu audio des séquences CV avec le contenu vidéo de la parole libre.

Dans notre paradigme, décrit en détail dans le chapitre précédent, la cible permettant de mesurer l'effet McGurk est précédée par ces deux types de contexte (Figure 38). Notre hypothèse de base est que nous devrions observer une décroissance de l'effet McGurk (donc une augmentation de réponses « ba ») dans le cas du contexte incohérent. Cette décroissance pourrait indiquer que le processus de fusion audiovisuelle mis en jeu dans l'effet McGurk n'est pas automatique et instantané, comme considéré dans la vision classique, mais au contraire, qu'il existerait un processus préalable de liage susceptible de moduler la fusion.

6.2 Méthodologie

6.2.1 Stimuli

Nous ne décrirons pas à nouveau les stimuli utilisés dans cette expérience, déjà décrits en détail dans le chapitre précédent.

6.2.2 Plan d'expérience

Cette expérience comporte trois variables indépendantes : contexte (cohérent vs incohérent), cible (« Ba » vs « McGurk »), durée du contexte (5, 10, 15 et 20 syllabes) et la variable dépendante qui est le taux de perception d'effet McGurk. Dans cette première expérience, nous l'avons dit, le logiciel utilisé ne nous permettait pas de mesurer avec la précision nécessaire le temps de réponse.

Au total nous avons présenté 128 stimuli, dont 16 stimuli « Ba » et 48 stimuli « McGurk » dans le contexte cohérent et autant dans le contexte incohérent. Nous avons 16 stimuli audio originaux (contexte + cible) pour l'ensemble des 4 durées avec 4 stimuli par durée. Ces stimuli audio sont utilisés dans tous les films et répétés 3 fois dans le cas de cibles « McGurk ». Pour les contextes cohérents ces stimuli audio étaient associés à leur composante vidéo d'origine et pour les contextes incohérents avec les pistes vidéo de la parole libre.

Tous les stimuli ont été randomisés et séparés en 4 blocs avec 32 stimuli par bloc. La durée d'expérience est environ 25 minutes.

Dans nos expériences nous travaillons avec des stimuli naturels et donc il peut y avoir des différences entre cibles conduisant à des variations d'amplitude de l'effet McGurk. Dans notre paradigme expérimental le facteur « durée du contexte » correspondant à des cibles différentes, nous ne présenterons pas l'analyse de durée du contexte, mais nous y reviendrons dans le chapitre 9.

6.2.3 Sujets

19 sujets français ont participé à l'expérience (8 femmes et 11 hommes, entre 22 et 51 ans avec 30,3 ans en moyenne, 17 droitiers et 2 gauchers). Ils avaient tous une vision et audition normale ou corrigée. Tous les sujets ont donné un consentement éclairé pour participer à l'expérience et ils n'étaient pas au courant du but de l'étude.

6.3 Résultats

6.3.1 Scores bruts

Nous présentons le détail des réponses des sujets, regroupées dans une matrice de confusion (Tableau 3), et présentées graphiquement dans la Figure 56.

6.3.1.1 Taux de réponses manquantes

Examinons d'abord les réponses manquantes, correspondant à deux types de situation : absence de réponse ou plusieurs réponses contradictoires pendant la fenêtre temporelle de 3.5s. Le taux global de réponses manquantes, regroupant ces deux situations, est de 12,5 %. Ce taux important peut s'expliquer d'une part par la tâche qui implique un choix forcé « ba » vs. « da » sur des stimuli McGurk qui peuvent être perçus de manière variable et éventuellement autrement que « ba » ou « da » (« tha », par exemple) et d'autre part par l'incertitude temporelle de l'arrivée des cibles et le temps assez court (3.5 s) mis à dispositions des sujets pour répondre (rappelons qu'il s'agit d'une tâche de monitoring en ligne). Nous verrons cependant que dans les expériences suivantes, menées sur des sujets plus jeunes, les taux de non réponse sont sensiblement plus faibles.

Une analyse des réponses plus détaillée montre que si le nombre de cas avec plusieurs réponses différentes est assez semblable selon les conditions, le nombre de cas de réponses manquantes est plus élevé pour les cibles « McGurk ».

Tableau 3 – Matrice de confusion (le cas « plusieurs réponses « ba » est ensuite inclus dans les réponses « ba », idem pour les réponses «da » ; le cas « absence de réponse » signifie qu'aucune réponse n'a été fournie pendant la période disponible de 3.5 s, le cas « absence / plusieurs réponses » signifie que 2 réponses différentes au moins ont été proposées dans la période de 3.5 s, ce cas est ensuite inclus dans le score total de non réponse)

Stimuli		Stimuli présentés	Réponse « ba »		Réponse « da »		Plusieurs réponses « ba »		Plusieurs réponses « da »		Absence de réponse		Absence plusieurs réponses	
Cohérent	Ba	304	252	(83 %)	8	(3 %)	9	(3 %)	0	(0 %)	27	(9 %)	8	(3 %)
	McG	912	447	(49 %)	317	(35 %)	19	(2 %)	5	(1 %)	101	(11 %)	23	(3 %)
Incohérent	Ba	304	259	(85 %)	9	(3 %)	8	(3 %)	1	(0 %)	13	(4 %)	14	(5 %)
	McG	912	708	(78 %)	51	(6 %)	35	(4 %)	0	(0 %)	94	(10 %)	24	(3 %)

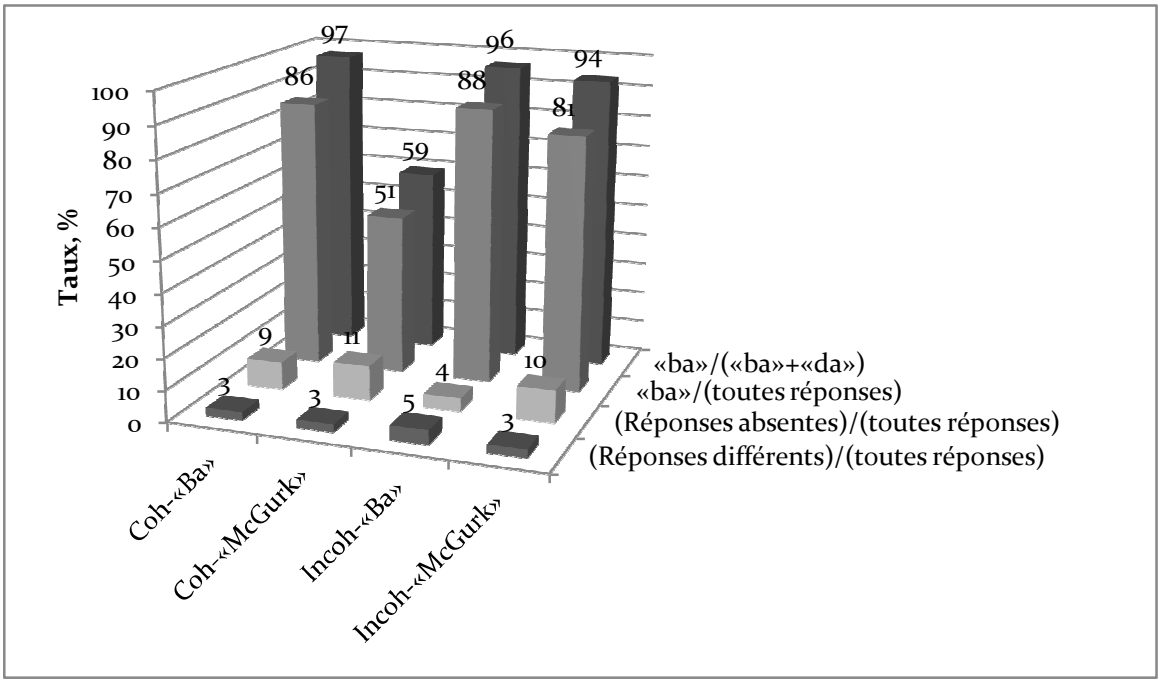


Figure 56–Expérience 1, données brutes, %.

6.3.1.2 Taux d'effet McGurk

Examinons maintenant les taux de réponses « ba » (en regroupant le cas d'une seule réponse et de plusieurs réponses « ba ») et « da » (en regroupant de même le cas d'une seule réponse et de plusieurs réponses « da »). On observe sur la Figure 56 que si les cibles « Ba » fournissent pour les deux contextes une quasi unanimité de réponses « ba », les cibles « McGurk » fournissent un score beaucoup plus faible de réponses « ba » en contexte cohérent, mais pas en contexte incohérent. Ceci apparaît que l'on s'intéresse au score « ba »/(« ba »+« da ») ou « ba »/« total des réponses ».

Dans la figure suivante (Figure 57) nous analysons les comportements individuels des participants. Regardons d'abord la perception d'effet McGurk dans la condition favorable, c'est-à-dire dans le contexte cohérent (Figure 57, traits sombre). Notons que dans cette figure comme dans la plupart des autres nous présentons le score « ba »/(« ba »+« da »), et donc pour calculer le taux de réponses « da », qui correspond à l'effet McGurk, il faut faire la différence (100% - taux présenté). Sur cette figure nous voyons que le taux de perception d'effet McGurk est différent selon les sujets. Les larges différences interindividuelles sont en accord avec la littérature (Schwartz, 2010). Par contre, dans le contexte incohérent nous observons que le taux d'effet McGurk diminue considérablement pour tous les sujets (les scores de réponses « ba » augmentent) et les deux courbes ne se croisent quasiment jamais : l'amplitude d'effet McGurk est plus élevée en contexte cohérent pour tous les sujets qui présentent un effet McGurk (c'est-à-dire qui répondent à moins de 95% « Ba » pour une cible McGurk). L'homogénéité de réponse des sujets au contexte est frappante.

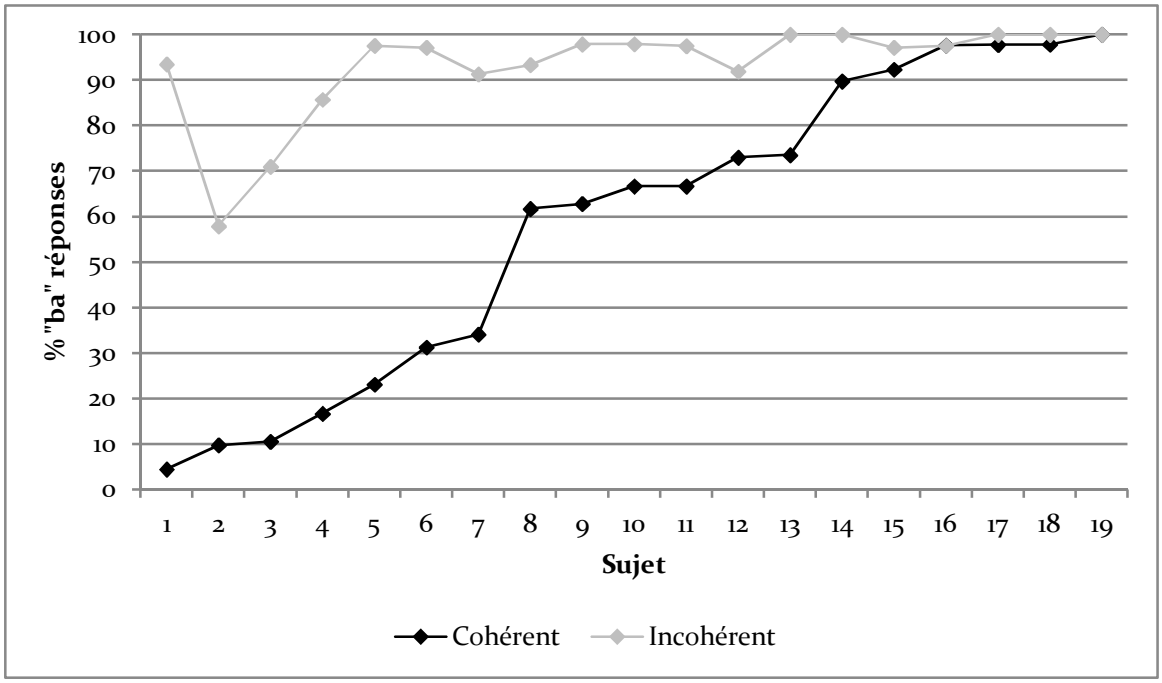


Figure 57 - Perception d'effet McGurk par sujet (« ba »/(« ba » + « da »)).

6.3.2 Analyses statistiques des pourcentages de réponse

6.3.2.1 Effets de la cible et du contexte

6.3.2.1.1 Taux des réponses « ba »/(« ba » + « da »).

Regardons maintenant les pourcentages globaux de réponse aux cibles « Ba » et « McGurk » (Figure 58). Nous l'avons vu, les cibles « Ba » sont presque toujours identifiées correctement (réponse « ba ») alors que les cibles « McGurk » sont identifiées en contexte cohérent, à 62% comme « ba », donc à 38% comme « da ». On retrouve les scores classiques d'effet McGurk en français (Cathiard et al, 2001). Il faut aussi mentionner les études sur l'effet McGurk en français de Colin et al. (Colin et al, 2002), qui précisent l'influence de l'intensité sonore sur la perception d'effet McGurk. Avec un niveau confortable à 70 dB l'effet McGurk y est très faible, tandis qu'une réduction d'intensité jusqu'à 40 dB favorise la fusion audiovisuelle. Rappelons que dans nos expériences le niveau d'intensité sonore est à 58 dB. Dans l'étude de Colin et al. le taux de fusion audiovisuelle était très faible (autour de 5%), alors que le taux de combinaisons des réponses était assez élevé (entre 40 et 70% selon l'intensité sonore). Cependant le paradigme expérimental de Colin et al. est différent, avec des stimuli présentés un par un de façon très prédictible, et un choix de nombre de catégories de réponses plus élevé, tandis que dans notre expérience le sujet doit identifier les cibles en ligne avec un choix forcé entre « ba » et « da ». Globalement, on sait que l'effet McGurk classique en anglais se traduit plus souvent par la perception de « tha », non phonologique en français, que de « da », ce qui peut expliquer le niveau moins élevé d'effet McGurk en français.

En contexte incohérent, l'effet McGurk est quasiment supprimé : les cibles « McGurk » sont identifiées comme « ba » dans plus de 95% des cas. Une ANOVA à mesures répétées sur les facteurs contexte et cible (Tableau 4) montre un effet significatif de deux facteurs (contexte

[F(1,18)=24.23, P<0.001], cible [F(1,18)=24.99, P<0.001]), ainsi que de leur interaction [F(1,18)=45.76, P<0.001]. Une analyse post hoc confirme que le taux de réponses « ba » pour les cibles McGurk est significativement différent entre les contextes cohérent et incohérent (P<0.001). Les analyses statistiques confirment qu'il y a bien un effet McGurk, mais qu'il est fortement modulé (et même quasiment supprimé) par l'effet du contexte.

L'effet sujet est également significatif [F(1,18)=985.96, P<0.001], ce qui confirme que la perception d'effet McGurk est variable selon les sujets (Figure 57).

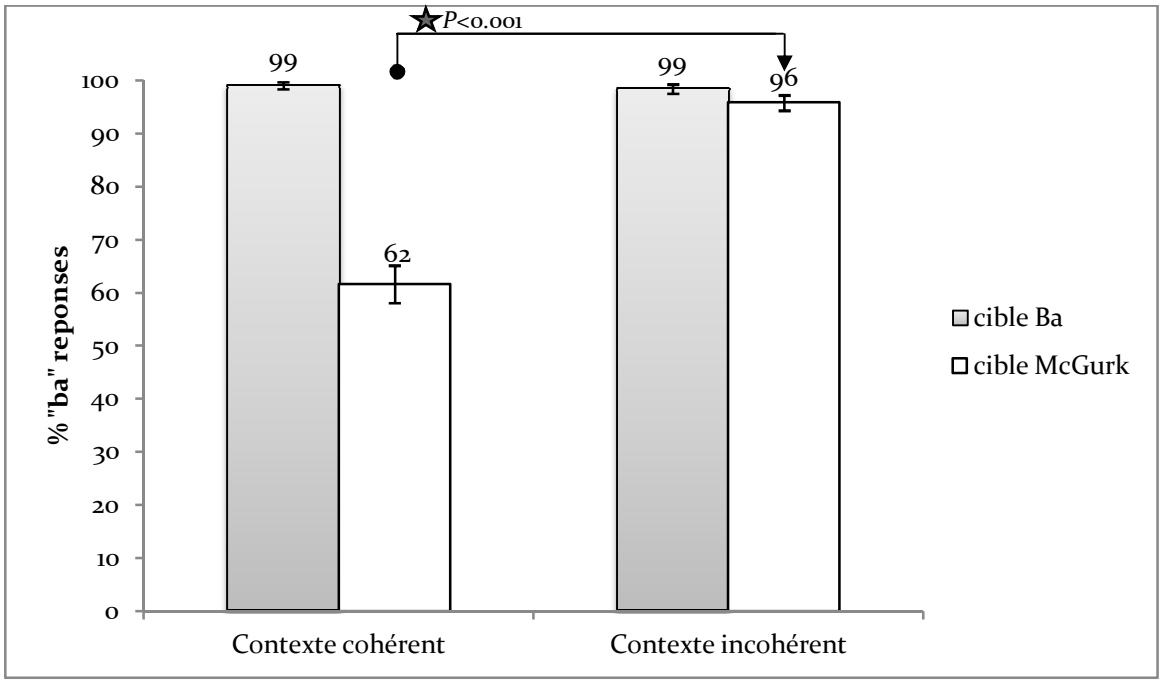


Figure 58 - Taux de réponses (« ba »/(« ba » + « da ») dans l'Expérience 1.

Tableau 4– ANOVA à mesures répétées: cible, contexte.

Source		Somme des carrés de type III	ddl	Moyenne des carrés	D	Sig.
Contexte : coh, incoh	Sphéricité supposée	,899	1	,899	24,232	,000
Erreur(contexte)	Sphéricité supposée	,668	18	,037		
Cible : Ba, McGurk	Sphéricité supposée	2,058	1	2,058	24,986	,000
Erreur(cible)	Sphéricité supposée	1,482	18	,082		
*Contexte * cible*	Sphéricité supposée	1,159	1	1,159	45,761	,000
*Erreur(contexte*cible)*	Sphéricité supposée	,456	18	,025		

6.3.2.1.2 Taux des réponses« ba »/(toutes réponses), « da »/(toutes réponses), (réponses absentes/(toutes réponses)

Nous avons également contrôlé les effets sur les scores « ba »/« total des réponses », « da »/« total » et « erreur »/ « total »afin de vérifier qu'ils sont similaires à ceux des scores « ba »/(« ba »+« da »). Nous présentons dans cette expérience princeps un exemple de ces analyses en large partie redondantes, mais dans les expériences suivantes, nous n'évoquerons ces analyses complémentaires que si elles produisent des résultats différents.

L'analyse sur les scores « ba »/« total des réponses » est présentée dans la Figure 59. Les cibles « Ba » sont perçues comme « ba » dans 90% des cas dans les deux contextes. Les stimuli « McGurk » sont perçus comme « ba » dans 52% des cas dans le contexte cohérent et 85% des cas dans le contexte incohérent. L'ANOVA à mesures répétées (Tableau 5) donne des résultats significatifs pour les facteurs contexte [$F(1,18)=32.3$, $P<0.001$] et cible [$F(1,18)=20.51$, $P<0.001$], ainsi que pour leur interaction [$F(1,18)=12.22$, $P<0.003$]. Cette analyse confirme notre analyse principale.

Les scores « da »/(toutes réponses) sont cohérents avec les analyses précédents (Figure 59, Tableau 6). Les sujets ont répondu « da » dans 30% des cas avec les stimuli « McGurk »en contexte cohérent et dans 3% des cas en contexte incohérent. L'ANOVA à mesures répétées est également significative pour les facteurs contexte [$F(1,18)=24.96$, $P<0.001$] et cible [$F(1,18)=25.06$, $P<0.001$] et pour leur interaction [$F(1,18)=46.38$, $P<0.001$].

Par contre l'analyse des taux de réponses manquantes (Figure 61, Tableau 7) donne un résultat non significatif pour le facteur contexte [$F(1,18)=2.72$, $P=0.12$] et l'interaction contexte*cible [$F(1,18)=0.77$, $P=0.39$], mais un résultat significatif pour le facteur cible [$F(1,18)=7.01$, $P<0.016$]. Cette analyse confirme notre présentation initiale de la Figure 56 : les cibles « McGurk » mettent plus souvent les sujets en difficulté pour répondre. L'effet non significatif du facteur contexte, seul ou en interaction avec la cible, suggère que le contexte ne module pas les scores d'erreur.

Globalement néanmoins, les scores d'erreur ne conduisent pas à des différences de portrait entre les analyses, que l'on s'intéresse aux scores « ba » / (« ba » + « da »), « ba » / total ou « da » / total, ce qui est rassurant sur la portée de nos analyses.

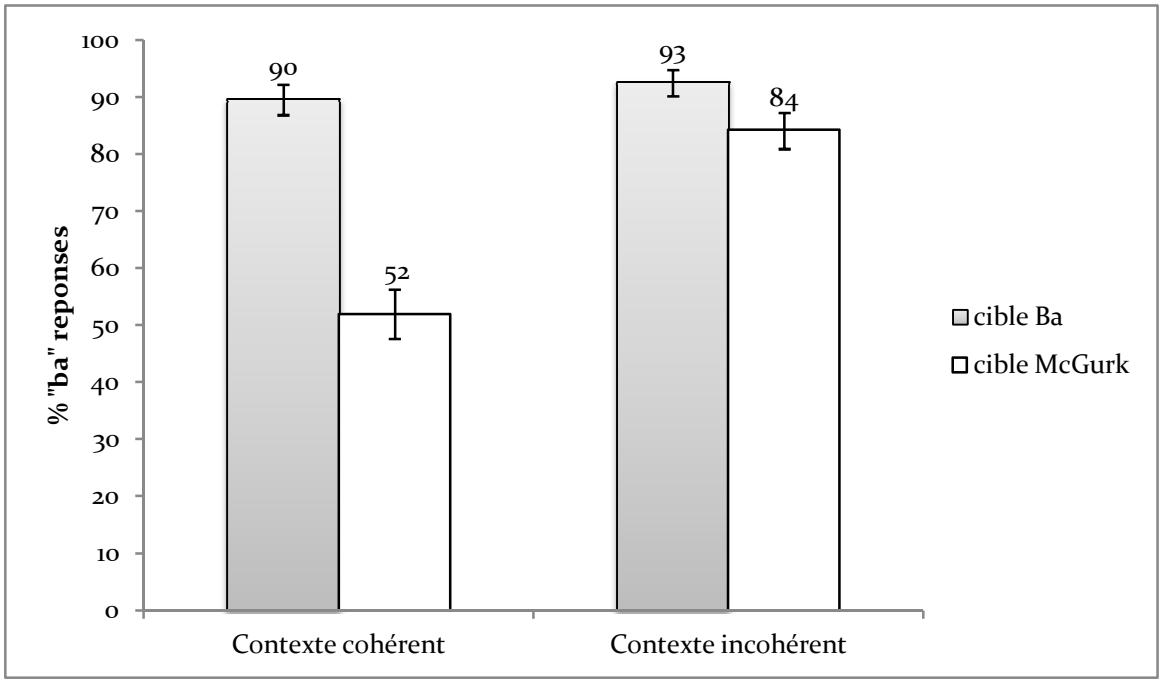


Figure 59 - Taux des réponses (« ba »/(toutes réponses)).

Tableau 5– ANOVA à mesures répétées: cible, contexte sur le taux « ba »/(toutes réponses).

Source		Somme des carrés de type III	ddl	Moyenne des carrés	D	Sig.
Contexte : coh, incoh	Sphéricité supposée	,798	1	,798	32,298	,000
Erreur(contexte)	Sphéricité supposée	,445	18	,025		
Cible : Ba, McGurk	Sphéricité supposée	1,556	1	1,556	20,514	,000
Erreur(cible)	Sphéricité supposée	1,365	18	,076		
*contexte * cible*	Sphéricité supposée	,441	1	,441	12,222	,003
*Erreur(contexte*cible)*	Sphéricité supposée	,650	18	,036		

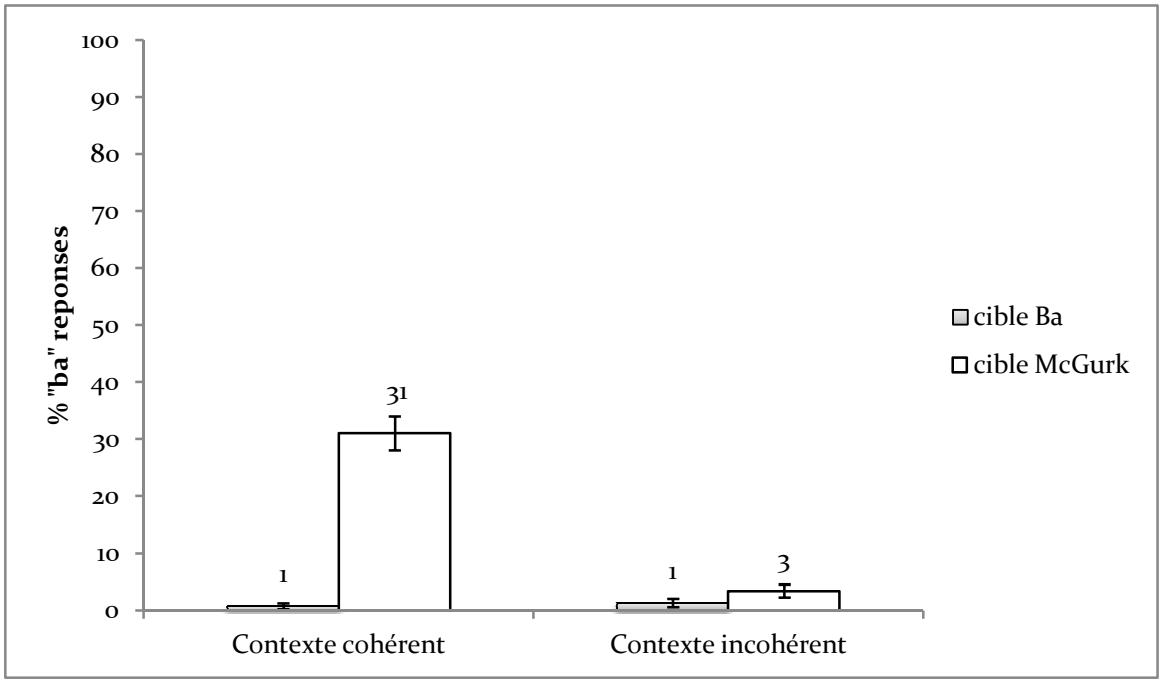


Figure 60 - Taux des réponses (« da »/(toutes réponses)).

Tableau 6 – ANOVA à mesures répétées: cible, contexte sur le taux « da »/(toutes réponses).

Source		Somme des carrés de type III	ddl	Moyenne des carrés	D	Sig.
Contexte : coh, incoh	Sphéricité supposée	,683	1	,683	24,964	,000
Erreur(contexte)	Sphéricité supposée	,493	18	,027		
Cible : Ba, McGurk	Sphéricité supposée	1,607	1	1,607	25,060	,000
Erreur(cible)	Sphéricité supposée	1,154	18	,064		
*contexte * cible*	Sphéricité supposée	,898	1	,898	46,383	,000
*Erreur(contexte*cible)*	Sphéricité supposée	,348	18	,019		

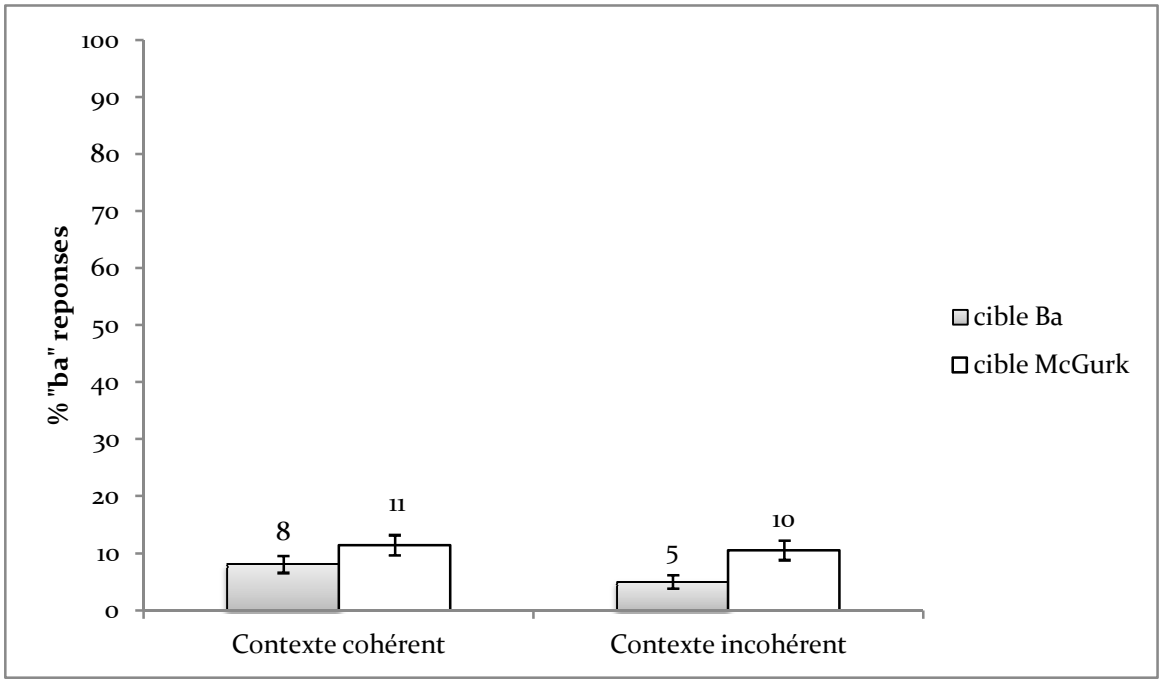


Figure 61 - Taux des réponses (« Réponses absentes»/(toutes réponses)).

Tableau 7 – ANOVA à mesures répétées: cible, contexte sur le taux «absence des réponses »/(toutes réponses).

Source		Somme des carrés de type III	ddl	Moyenne des carrés	D	Sig.
Contexte : coh, incoh	Sphéricité supposée	,029	1	,029	2,715	,117
Erreur(contexte)	Sphéricité supposée	,191	18	,011		
Cible : Ba, McGurk	Sphéricité supposée	,125	1	,125	7,006	,016
Erreur(cible)	Sphéricité supposée	,322	18	,018		
*Contexte * cible*	Sphéricité supposée	,011	1	,011	,765	,393
*Erreur(contexte*cible)*	Sphéricité supposée	,263	18	,015		

6.3.2.2 *L'effet durée*

Le facteur « durée » était un facteur important dans cette expérience. Cependant, comme mentionné précédemment, des analyses a posteriori à la lumière de l'Expérience 4 que nous présentons plus loin, nous ont montré que des différences de contenu des stimuli cible pouvaient brouiller l'interprétation des résultats. Nous ne commenterons donc pas ces analyses ici.

6.3.2.3 *L'effet sous-groupe des cibles*

Nous avons mentionné en détail au chapitre précédent comment nous avons cherché à contrôler aussi précisément que possible le contenu visuel des cibles, en sélectionnant dans le contexte incohérent des stimuli au sein de deux groupes, selon l'amplitude relative d'ouverture de la bouche (Figure 50) (§5.3.3). Nous testons dans cette section l'influence du contenu visuel des stimuli incohérents, en comparant les résultats obtenus pour les deux sous groupes, le sous groupe « incohérent-» d'ouverture relative plus faible, et le sous-groupe « incohérent+» d'ouverture relative plus grande par rapport aux cibles cohérentes.

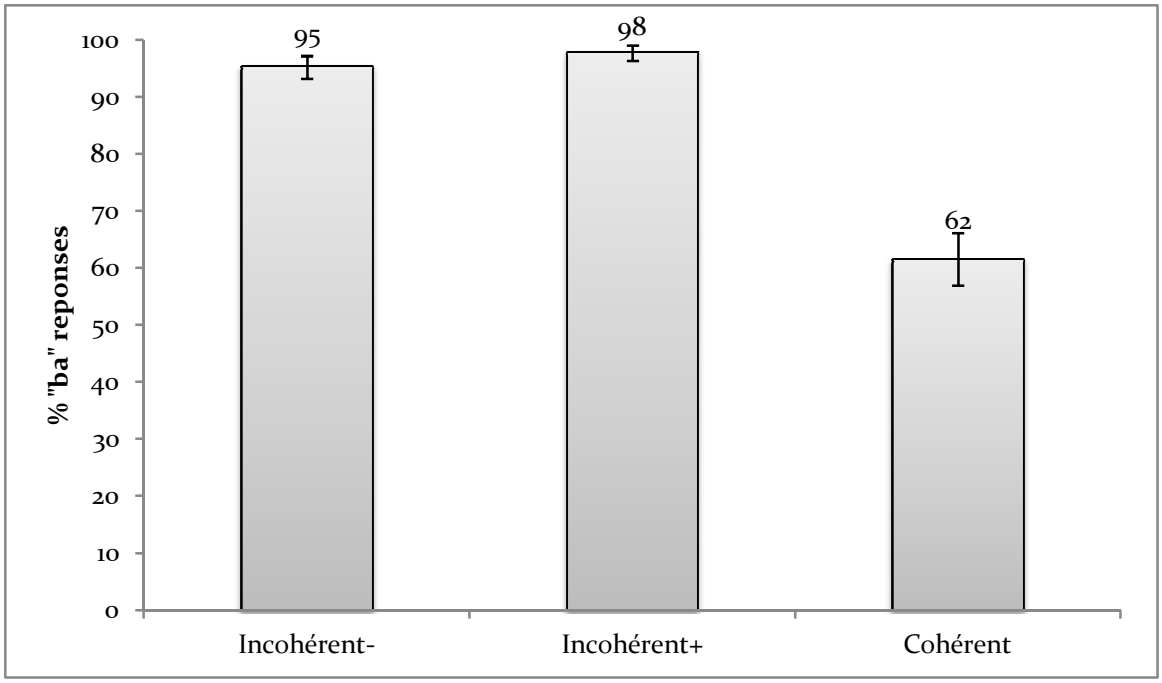


Figure 62 Taux de réponses (« ba »/(« ba » + « da »)) pour les deux sous-groupes en fonction de la valeur d'ouverture des lèvres (amplitude relative Δy>0.8 et Δy<0.8)

Sur la Figure 62 nous constatons que les cibles « McGurk » dans le groupe « incohérent-» sont perçues comme « ba » dans 95% des cas et dans le groupe « incohérent+» dans 97% des cas. Par comparaison, dans le contexte cohérent les sujets ont répondu « ba » dans 60% des cas. Ainsi, l'effet de contexte apparaît très clairement dans les deux sous-groupes. Cela se confirme dans une ANOVA à mesures répétées conduite sur les trois sous groupes (cohérent et deux sous-groupes incohérents) [$F(1.26,22.64)=35.5$, $P<0.001$] (Tableau 8). L'analyse post hoc($P<0.001$) indique que cet effet significatif est du à une différence du sous-

groupe cohérent par rapport aux sous-groupes « incohérent-» et « incohérent+», qui sont perçus de la même façon. Cela confirme que c'est bien le contexte qui supprime l'effet McGurk dans cette expérience et non le contenu visible des stimuli.

Tableau 8- ANOVA à mesures répétées sur les réponses aux cibles McGurk pour trois conditions de contexte en introduisant un facteur sous-groupe

Source		Somme des carrés de type III	ddl	Moyenne des carrés	D	Sig.
sous-groupe: coh, incoh-, incoh+	Huynh-Feldt	3,057	1,308	2,337	35,499	,000
Erreur(sous-groupe)	Huynh-Feldt	1,550	23,552	,066		

6.3.3 Temps de réponse

Le logiciel utilisé ne nous permet pas de faire une analyse précise de temps de réponse.

6.4 Conclusion

Les résultats obtenus ont mis en évidence que l'effet McGurk dépend du contexte préalable. La suppression d'effet McGurk signifie que l'on peut bloquer la fusion audio-visuelle. Dans le cas d'un contexte cohérent les flux auditif et visuel se combinent au sein de la cible dans un même percept. Dans le cas incohérent ce ne semble pas être le cas et la réponse est essentiellement gérée par l'information auditive. Cela signifie que la fusion audiovisuelle n'est pas automatique et inconditionnelle. Ce résultat va dans le sens de notre hypothèse d'existence d'un processus de liage qui module/supprime la fusion audiovisuelle.

L'analyse des deux sous-groupes de stimuli a permis de vérifier que la différence de perception d'effet McGurk observé semble réellement due aux variations du contexte et non à une différence de contenu visuel des cibles. Les analyses complémentaires sur les différents scores de réponse ont confirmé que l'effet n'est pas « pollué » par des taux d'absence de réponse.

Cette expérience a également confirmé que l'effet McGurk varie fortement selon les sujets avec un faible taux moyen d'effet McGurk en accord avec la littérature sur la perception d'effet McGurk chez les sujets français (Cathiard et al, 2001). Cependant, l'effet du contexte est, lui, présent de manière remarquablement homogène chez tous les sujets à part bien sûr ceux qui ne présentent rigoureusement aucun effet McGurk.

Par contre, pour des raisons de contrôle des stimuli sur lesquelles nous reviendrons, nous n'avons pas pu analyser le facteur « durée », qui était pourtant un des enjeux de cette expérience. Plus tard, dans le chapitre 9 (Expérience 4), nous donnerons des éléments de réponse à cette question.

Nous allons, dans la prochaine expérience, tenter de conforter ce premier résultat, qui apparaît très fort, en étudiant sa résistance à un facteur attentionnel local.

Chapitre 7. Expérience 2 : est-ce qu'un stimulus d'alerte temporelle influence le liage ?

7.1 Objectifs et hypothèses

Les résultats de l'expérience précédente ont montré l'influence de la cohérence du contexte sur la perception d'effet McGurk, donc sur la fusion audiovisuelle de la parole. Etant donnée la forte capacité de ce stimulus contexte à éliminer presque totalement l'effet McGurk, nous étudions dans ce chapitre s'il est possible d'éliminer l'effet de déliage du contexte et de « reconstruire » l'effet McGurk en tentant de focaliser temporellement l'attention du sujet sur la cible McGurk.

Il est connu qu'une courte alerte peut améliorer la performance dans un certain nombre de tâches, telles que la détection d'une sonde à l'intérieur d'une fenêtre temporelle, ou un jugement d'ordre temporel (Coull & Nobre, 1998), (Correa et al, 2006). Dans l'étude de Van der Burg et al. (Van der Burg et al, 2008), une courte alerte auditive permet à l'inverse de focaliser l'attention visuelle dans une tâche de détection de cible au milieu d'un ensemble de distracteurs, en permettant au sujet, si l'alerte est synchrone avec un changement de couleur de la cible, de la détecter plus rapidement.

Nous nous demandons ici si une courte alerte positionnée juste avant la cible McGurk pourrait rehausser la perception d'effet McGurk, malgré le déliage dû à un contexte incohérent préalable ou si, au contraire, le déliage audiovisuel résiste à une telle alerte. L'expérience est de type runtime, avec nécessité de détecter la cible en ligne, donc on peut supposer que le fait de mettre une alerte avant la cible favorise la détection et ainsi permet aux sujets de focaliser leur attention sur les cibles d'intérêt, ce qui pourrait renforcer l'effet McGurk en cas de contexte incohérent. Ainsi notre question expérimentale est de tester si une courte alerte pourrait relier les deux flux auditif et visuel qui étaient précédemment déliés.

Pour maximiser les chances d'observer cet effet, nous avons choisi une alerte audiovisuelle, impliquant à la fois les deux modalités.

7.2 Méthodologie

7.2.1 Principe

Pour tester cette hypothèse nous avons repris intégralement dans cette expérience le paradigme expérimental de l'Expérience 1 (Figure 38). Dans cette nouvelle expérience nous avons rajouté une condition où la cible est systématiquement précédée par une alerte audiovisuelle (Figure 63).

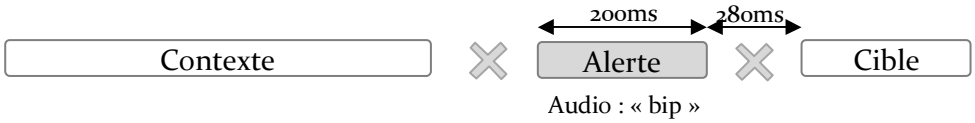


Figure 63 - Paradigme expérimental

L'alerte consiste en un stimulus audiovisuel de 200 ms constitué d'un « bip » audio et d'un cadre périphérique noir. Le bip est un son pur à 1 kHz modulé par une enveloppe de type fenêtre de Hamming. Le composant vidéo est un cadre d'une épaisseur de 150x40 pixels

entourant l'image (qui a une taille de 720 x 576 pixels), permettant d'attirer l'attention des sujets par la vision périphérique sans cacher le visage du locuteur, et ainsi sans produire d'effet de masquage visuel. L'alerte est placée 280ms avant le burst d'une cible, durée que nous avons choisie courte mais suffisante pour éviter un effet de « masquage postérieur » (forward masking), qui aurait pu se produire si l'alerte était placée moins de 200 ms avant la cible (Moore, 2003). La durée du stimulus d'alerte est 200 ms. L'alerte est temporellement concomitante avec la dernière syllabe du contexte.

7.2.2 Stimuli

Les stimuli utilisés dans cette expérience sont les mêmes que ceux de l'Expérience 1, décrits en détail aux paragraphes §5.2 et le §5.3. A partir de ces stimuli nous avons créé le deuxième groupe de stimuli avec une alerte audiovisuelle.

7.2.3 Plan d'expérience

Dans cette expérience nous avons les variables indépendantes : contexte (cohérent *vs.* incohérent), alerte (sans alerte *vs.* avec alerte), cible (« Ba » *vs.* « McGurk »), et durée du contexte (5, 10, 15 et 20 syllabes). La variable dépendante est le taux de perception d'effet McGurk. Malheureusement le logiciel utilisé ne nous permettait pas de mesurer avec la précision nécessaire le temps de réponse.

Dans chaque groupe de stimuli (sans alerte et avec alerte) nous avons préparé 16 stimuli originaux « McGurk » (chacun répété deux fois) et 10 stimuli « Ba » précédés par les contextes cohérent et incohérent, soit 42 stimuli originaux au total sans alerte et 42 stimuli avec alerte (Tableau 9).

Comme dans l'expérience précédente pour préserver un nombre suffisant de stimuli d'intérêt avec les cibles « McGurk » pendant l'analyse statistique, nous avons gardé des proportions de l'ordre de ¾ de stimuli avec les cibles « McGurk » versus ¼ de stimuli avec cibles « Ba », tout en diminuant le nombre total de stimuli pour conserver un temps raisonnable à cette expérience qui implique deux fois plus de conditions. La répartition exacte du nombre de stimuli de chaque type est indiquée dans le (Tableau 9). Au total dans cette expérience nous avons présenté 168 stimuli.

Tableau 9 - Répartition des stimuli présentés dans l'Expérience 2

Absence d'alerte				**Alerte AV**			
cohérent		**incohérent**		**cohérent**		**incohérent**	
« Ba »	**« McGurk »**	**« Ba »**	**« McGurk »**	**« Ba »**	**« McGurk »**	**« Ba »**	**« McGurk »**
10	32	10	32	10	32	10	32

Une première hypothèse dans cette expérience était que l'alerte pourrait diminuer le rôle des mécanismes de déliage et de réduction de l'effet McGurk en contexte incohérent. Une seconde hypothèse était que l'alerte permettrait progressivement de focaliser l'attention des sujets vers les portions finales de chaque film (rappelons que dans cette tâche de monitoring, les sujets ne savent pas quand peuvent apparaître des « ba » ou des « da »). Pour tester cette seconde hypothèse, nous avons décidé de présenter ici les stimuli en 2 blocs, un bloc avec

alerte et un bloc sans alerte, et d'opérer sur deux groupes de sujets, l'un passant les stimuli avec alerte en second, et l'autre en premier : le second groupe étant susceptible de démontrer des effets d'apprentissage.

L'expérience consiste ainsi en 4 blocs, dont 2 blocs de stimuli sans alerte et 2 blocs de stimuli avec alerte, avec des pauses entre les blocs. La durée des pauses est décidée par le sujet. L'ordre du passage des blocs est varié selon les sujets, la moitié des sujets ont commencé par des blocs sans alerte, et l'autre moitié par des blocs avec alerte. Chaque bloc consiste en 42 stimuli, choisis de façon aléatoire, parmi les stimuli soit avec alerte, soit sans alerte.

Dans cette expérience, comme dans la précédente, le facteur « durée du contexte » n'est pas contrôlé de manière adéquate pour permettre l'analyse de son effet, nous y reviendrons dans le chapitre 9.

En résumé, l'expérience consiste en :

- 168 stimuli en total
- 4 blocs de 42 stimuli, répartis aléatoirement. Deux blocs avec alerte, deux blocs sans alerte, présentés selon les sujets dans un ordre (alerte d'abord) ou l'autre (alerte ensuite).
- 2 types de réponses possibles « ba » ou « da » en ligne, choix forcé

Les conditions de passation d'expérience, la présentation des stimuli, les consignes, les types de réponses sont les mêmes que dans l'Expérience 1 (Paragraphe 5.4.2).

7.2.4 Sujets

20 sujets français (10 femmes et 10 hommes, entre 21 et 29 ans avec 24,3 ans en moyenne, 19 droitiers et 1 gaucher) ont participé à l'expérience, présentant tous une vision et audition normale ou corrigée. Tous les sujets ont donné un consentement éclairé à participer à l'expérience et n'étaient pas au courant du but de l'étude.

7.3 Résultats

7.3.1 Scores bruts

Nous suivons le même décours de présentation des résultats que dans l'expérience précédente. L'ensemble des réponses est présenté dans une matrice de confusion (Tableau 10). Le taux d'erreur global est de **10,45 %**.Ce taux reste donc important. Cependant, par rapport à l'expérience précédente, où ce taux élevé provenait surtout de cas d'absence de réponses, les erreurs dues à plusieurs réponses différentes sont ici plus nombreuses (Figure 64).

Nous centrons maintenant l'analyse sur le score de réponses « ba »/« ba »+« da ».

Tableau 10 - Matrice de confusion

Stimuli			Stimuli présentés	Réponse « ba »		Réponse « da »		Plusieurs réponses « ba »		Plusieurs réponses « da »		Absence de réponses		Absence plusieurs réponses	
Sans alerte	Cohérent	Ba	200	165	(83%)	9	(5%)	15	(8%)	0	(0%)	0	(0%)	11	(6%)
		McG	640	271	(42%)	260	(41%)	29	(5%)	9	(1%)	27	(4%)	44	(7%)
	Incohérent	Ba	200	166	(83%)	7	(4%)	12	(6%)	1	(1%)	1	(1%)	13	(7%)
		McG	640	508	(79%)	38	(6%)	35	(6%)	4	(1%)	22	(3%)	33	(5%)
Avec alerte	Cohérent	Ba	200	164	(82%)	1	(1%)	22	(11%)	0	(0%)	7	(4%)	6	(3%)
		McG	640	261	(41%)	242	(38%)	39	(6%)	9	(1%)	48	(8%)	41	(6%)
	Incohérent	Ba	200	164	(82%)	3	(2%)	19	(10%)	1	(1%)	4	(2%)	9	(5%)
		McG	640	439	(69%)	40	(6%)	66	(10%)	5	(1%)	31	(5%)	59	(9%)

Sur la Figure 65 nous présentons la perception d'effet McGurk en fonction du contexte et de l'alerte. Nous l'avons organisé en fonction d'une condition contexte cohérent sans alerte (scores croissant du sujet 1 au sujet 20). Nous observons que, comme dans l'expérience précédente, le taux de perception d'effet McGurk, très variable entre les sujets, varie par contre de façon remarquablement homogène en fonction du contexte, l'effet McGurk diminuant en contexte incohérent pour tous les sujets, sauf le sujet 14. Au contraire, l'effet d'alerte (traits pointillés) produit des effets plus faibles et très inhomogènes, certains sujets percevant plus d'effet McGurk, certains moins.

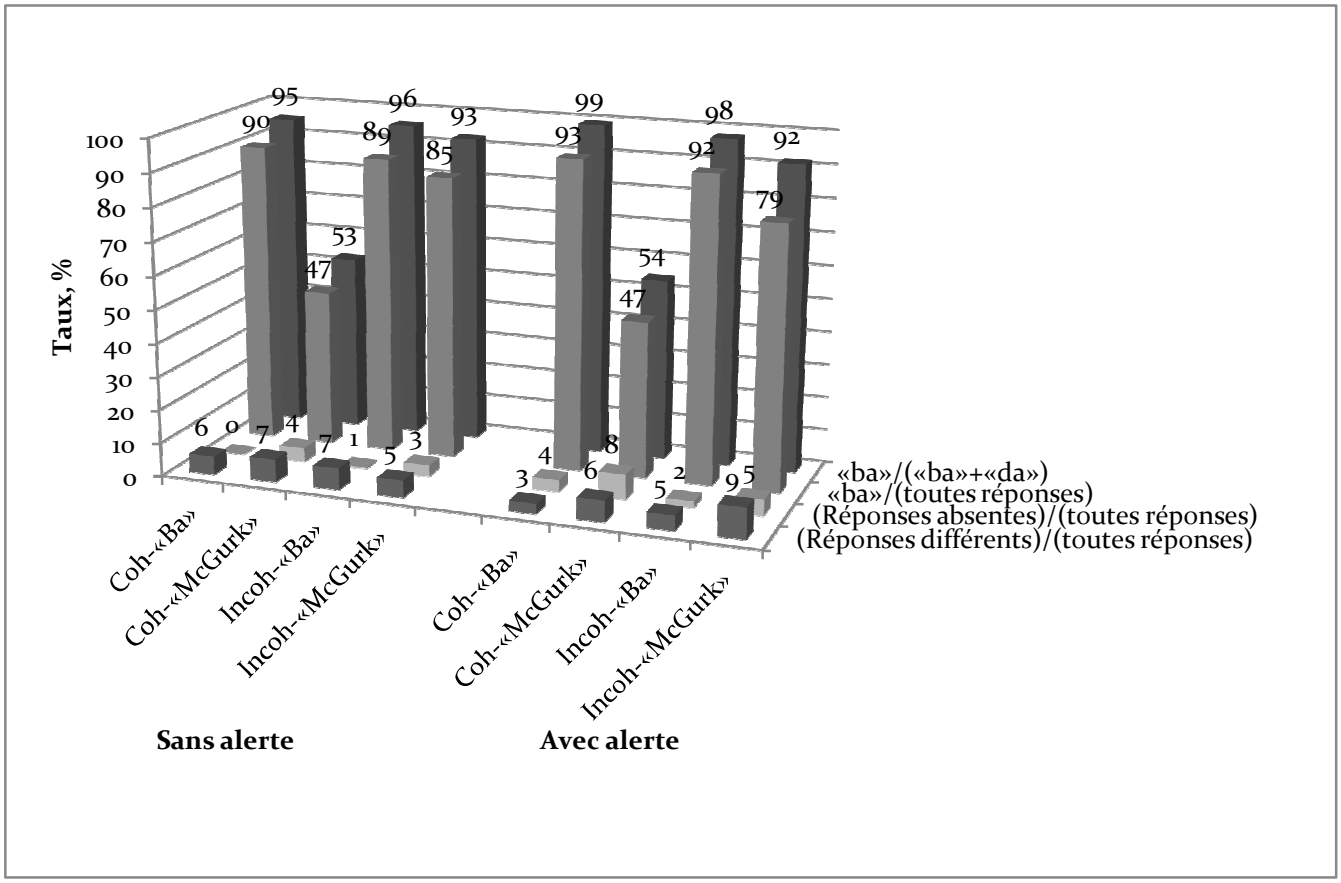


Figure 64–Expérience 2, données brutes, %

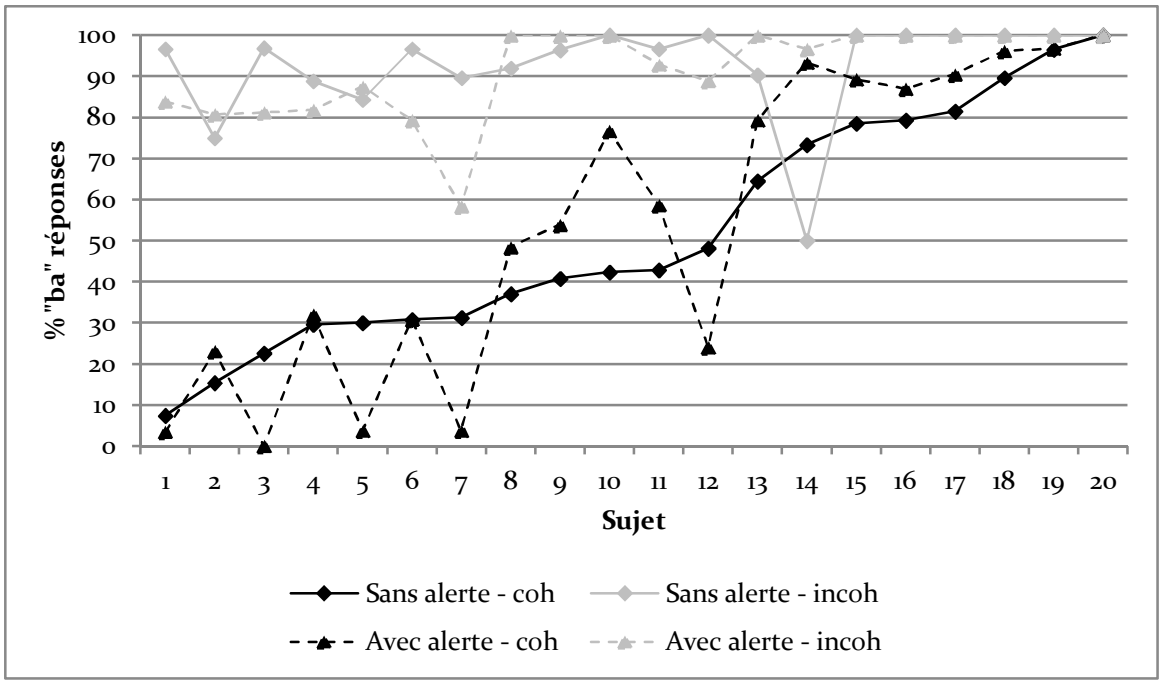


Figure 65 - Perception d'effet McGurk par sujet (« ba »/(« ba » + « da »)).

7.3.2 Analyses statistiques des pourcentages de réponse

7.3.2.1 *Effets de la cible, du contexte et de l'alerte.*

Passons maintenant aux pourcentages globaux de réponse aux cibles « Ba » et « McGurk » (Figure 66). Les cibles « Ba » restent presque toujours identifiées correctement. Les cibles « McGurk » sont identifiées en contexte cohérent, à 55% comme « ba » et à 45% comme « da », conformément aux scores classiques d'effet McGurk en français (Cathiard et al, 2001). En contexte incohérent, l'effet McGurk est encore une fois quasiment supprimé : les cibles McGurk sont identifiées comme « ba » dans plus de 95% des cas. Par contre, la présence d'alerte ne semble pas influencer la perception d'effet McGurk. Une ANOVA à mesures répétées à trois facteurs (contexte, cible, alerte) confirme l'effet significatif des facteurs contexte [$F(1,19)=61.65$, $P<0.001$] et cible [$F(1,19)=42.47$, $P<0.001$], ainsi que leur interaction [$F(1,19)=50.70$, $P<0.001$] et montre l'absence d'effet alerte, seul [$F(1,19)=0.25$, $P=0.64$] ou en interaction (Tableau 11).Une analyse post hoc confirme l'augmentation de taux des réponses « ba » pour les cibles « McGurk » du contexte cohérent au contexte incohérent ($P<0.001$).

L'effet sujet est significatif [$F(1,19)=1391.8$, $P<0.001$], confirmant que la perception d'effet McGurk est variable selon les sujets (Figure 65).

Ainsi cette expérience confirme les résultats de l'expérience précédente en montrant que le contexte incohérent diminue la perception d'effet McGurk. Par contre l'alerte n'influence pas sur la perception d'effet McGurk.

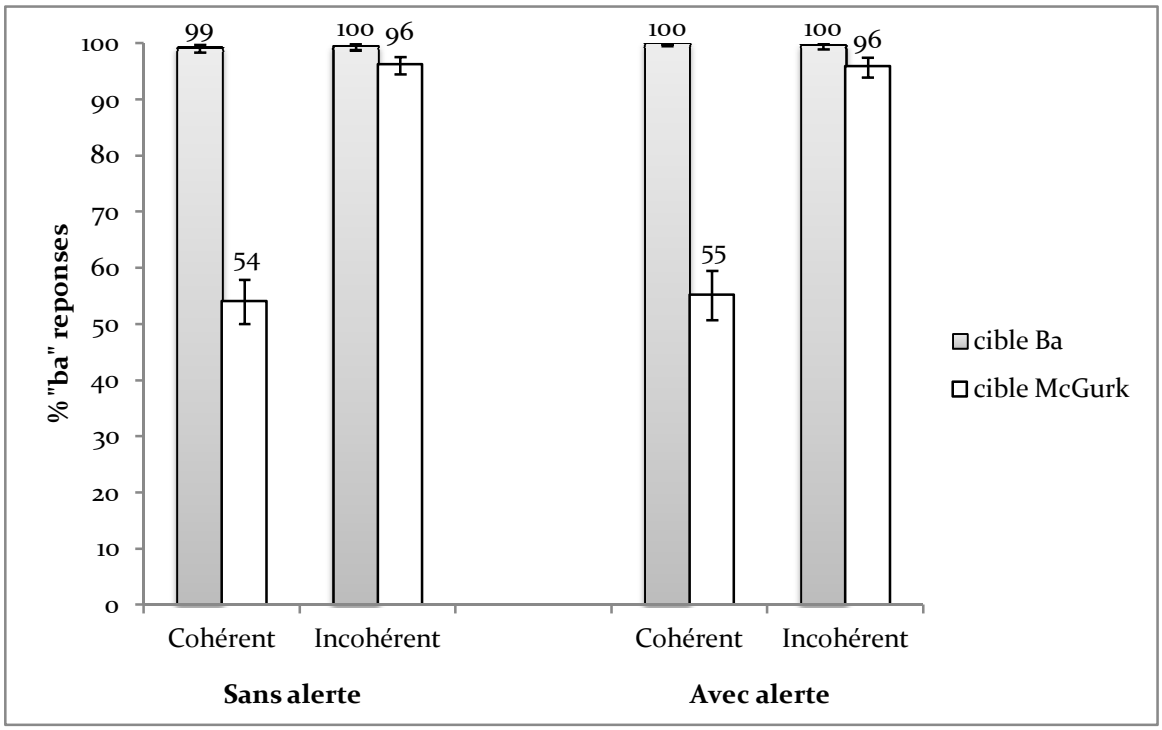


Figure 66 - Taux des réponses (« ba »/(« ba » + « da »)) dans l'Expérience 2.

Tableau 11– ANOVA à mesures répétées: alerte, contexte, cible.

Source		Somme des carrés de type III	ddl	Moyenne des carrés	D	Sig.
Alerte	Sphéricité supposée	,019	1	,019	,226	,640
Erreur (alerte)	Sphéricité supposée	1,557	19	,082		
Contexte : coh, incoh	Sphéricité supposée	2,802	1	2,802	61,652	,000
Erreur (contexte)	Sphéricité supposée	,863	19	,045		
Cible : Ba, McGurk	Sphéricité supposée	6,787	1	6,787	42,471	,000
Erreur (cible)	Sphéricité supposée	3,036	19	,160		
*Alerte*contexte*	Sphéricité supposée	,015	1	,015	2,802	,111
*Erreur (alerte*contexte)*	Sphéricité supposée	,105	19	,006		
*Alerte*Cible*	Sphéricité supposée	,017	1	,017	,681	,420
*Erreur (alerte*cible)*	Sphéricité supposée	,461	19	,024		
*Contexte * cible*	Sphéricité supposée	3,020	1	3,020	50,703	,000
*Erreur (contexte*cible)*	Sphéricité supposée	1,132	19	,060		
*Alerte*contexte*cible*	Sphéricité supposée	,004	1	,004	,317	,580
*Erreur (alerte*contexte*cible)*	Sphéricité supposée	,215	19	,011		

7.3.2.2 L'effet d'ordre de présentation des blocs

Pour savoir s'il y a une différence de perception de l'alerte entre les deux groupes de sujets en fonction de l'ordre de passage des blocs nous avons effectué une seconde ANOVA à mesures répétées, centrée sur les cibles « McGurk », avec deux facteurs intrasujets : alerte (avec alerte *vs*. sans alerte) et contexte (cohérent *vs*. incohérent) ; et un facteur intersujet, l'ordre de passage des blocs (sujets commençant l'expérience par des blocs sans alerte *vs*. par des blocs avec alerte) (Tableau 12). L'effet contexte est bien significatif [$F(1,18)=74.546$, $P<0.001$], mais ni l'alerte [$F(1,18)=0.001$, $P=0.98$] ni l'ordre des blocs [$F(1,18)=3.86$, $P=0.065$] ne le sont, ni aucune des interactions.

Ainsi, la présence d'un stimulus d'alerte temporelle n'influe pas sur le mécanisme de déliage par présentation d'un contexte incohérent, que l'on commence l'expérience par les

stimuli avec alerte ou sans alerte. Le mécanisme, supposé, de déliage audiovisuel résiste à cette alerte temporelle qui ne suffit donc pas à « re-lier » les flux auditif et visuel.

Là encore, nous ne présentons pas les résultats sur le facteur « durée de contexte » pour des raisons de contrôle sur lesquelles nous reviendrons.

Tableau 12 - ANOVA à mesures répétées sur les cibles « McGurk », avec deux facteurs intrasujets, contexte et alerte, et un facteur intersujet, l'ordre des blocs

Tests des effets intra-sujets

Source		Somme des carrés de type III	ddl	Moyenne des carrés	D	Sig.
Alerte	Sphéricité supposée	2,824E-05	1	2,824E-05	,001	,980
*Alerte*Ordre de blocs*	Sphéricité supposée	,046	1	,046	1,035	,322
Erreur (alerte)	Sphéricité supposée	,792	18	,044		
Contexte	Sphéricité supposée	5,820	1	5,820	74,546	,000
*Contexte*Ordre de blocs*	Sphéricité supposée	,267	1	,267	3,426	,081
Erreur (contexte)	Sphéricité supposée	1,405	18	,078		
*Alerte*contexte*	Sphéricité supposée	,002	1	,002	,204	,657
*Alerte*Contexte*Ordre de blocs*	Sphéricité supposée	,008	1	,008	,783	,388
*Erreur (alerte*contexte)*	Sphéricité supposée	,184	18	,010		

Tests des effets inter-sujets

Source	Somme des carrés de type III	ddl	Moyenne des carrés	D	Sig.
Ordonnée à l'origine (Sujet)	97,037	1	97,037	376,135	,000
Ordre des blocs	,995	1	,995	3,858	,065
Erreur	4,644	18	,258		

7.4 Discussion

Cette expérience confirme d'abord les résultats de l'expérience précédente. Les taux de perception d'effet McGurk dans les deux contextes sont similaires à ceux que nous avons observés dans l'Expérience 1, confirmant la stabilité des processus de déliage.

Elle montre également qu'une courte alerte audiovisuelle ne modifie pas la perception d'effet McGurk ni dans le contexte cohérent ni dans le contexte incohérent. Ainsi, bien que focalisant l'attention des sujets sur les cibles, cette alerte ne permet pas de « re-lier » les flux auditif et visuel, supposés partiellement ou totalement déliés par le contexte incohérent.

On peut se demander si cette alerte arrive trop tard par rapport à la cible, pour permettre un « re-liage » efficace. Pour le tester, il faudrait varier la distance temporelle entre alerte et cible.

D'autre part, le stimulus d'alerte, au contenu clairement disjoint du flux de parole (alerte composée, rappelons-le, d'un son pur et d'un cadre noir) a pu être exclu de la scène de parole et être considéré comme un événement différent et parallèle. Ce type d'effet est régulièrement noté par Bregman dans le cas de groupement des scènes auditives par primitives (Bregman, 1990). Dans notre cas la scène est audiovisuelle et non auditive, mais il est probable que le flux de parole était perçu comme un objet central et un alerte comme un deuxième objet de la scène, susceptible de ne pas interférer sur ce premier objet.

Ainsi, cette expérience produit un résultat négatif – l'absence d'effet d'alerte – qui est à la fois un appel à chercher d'autres moyens de « re-lier » les flux auditif et visuel (nous y reviendrons dans la Partie III) mais aussi une mise en évidence de la solidité du mécanisme de déliage, capable de résister ainsi à un effet de focalisation temporelle par l'application d'une discontinuité audiovisuelle juste avant la cible. Ceci nous sera particulièrement utile dans le cadre de l'Expérience 4 qui conclura la Partie II.

Chapitre 8. Expérience 3 : évaluation perceptive des cibles isolées

8.1 Objectifs et hypothèses

Les résultats des deux expériences précédentes semblent indiquer une claire dépendance de la fusion audio-visuelle vis-à-vis de son contexte préalable. Nous avons décrit dans le chapitre 5 les précautions prises pour mesurer et contrôler la dynamique labiale et les propriétés acoustiques de nos stimuli. Néanmoins ce contrôle n'est pas parfait, comme le montrent, sur la Figure 48, les différences des mesures physiques entre cibles extraites des contextes cohérents et incohérents, et comme le suggèrent, sur la Figure 62, les légères différences perceptives en terme d'effet McGurk des stimuli complets (contexte plus cible). Cette situation, inévitable pour préserver une continuité complète entre contexte et cible (empêchant de prendre les mêmes stimuli cibles pour les différents contextes) n'est évidemment pas optimale, et laisse un doute sur les résultats des deux expériences précédentes : est-ce que les différences de cible pourraient expliquer la chute d'effet McGurk des stimuli en contexte incohérent, chute que nous interprétons comme un effet de contexte ? L'objectif de cette troisième expérience est de tester directement comment sont perçues les cibles « McGurk » isolées, sans contexte, afin de mettre en évidence d'éventuelles différences de perception entre les cibles enregistrées après le contexte cohérent (production d'une séquence de syllabes se terminant par un « ba » ou un « ga » et celles qui suivent le contexte incohérent (production de phrases libres se terminant par un « ba » ou un « ga »).

8.2 Méthodologie

8.2.1 Principe

Le but de cette expérience est de tester la perception des cibles « McGurk » *vs.* cibles « ba » isolées, enregistrées dans des contextes différents et précédemment utilisées dans les Expériences 1 et 2. Notons que ceci diffère du paradigme classique de l'effet McGurk (McGurk & MacDonald, 1976). Dans notre expérience nous ne présentons que des cibles « Ba » et « McGurk », tandis que l'expérience classique contient en général également des cibles « Da ». De plus, nous maintenons un paradigme de choix forcé entre réponses « ba » et « da ». Nous maintenons également un paradigme de monitoring avec présentation des stimuli en continu, tandis que le paradigme classique est de type forme stimuli-réponse.

8.2.2 Stimuli

Les stimuli utilisés dans cette expérience sont les mêmes que ceux utilisés dans l'Expérience 1 et décrits en détail dans le §5.2 et le §5.3. Pour extraire les cibles de leur contexte et minimiser les effets de coarticulation persévératrice provenant de la fin du contexte précédent, les cibles sont coupées 280 ms avant le burst (plus un délai non contrôlable entre 0 et 40 ms, la durée d'une image), en maintenant la cohérence audio-visuelle. Nous conservons la proportion de ¾ de cibles « McGurk » *vs.* ¼ de cibles « Ba », avec choix forcé en ligne « ba » *vs.* « da ».

Nous obtenons ainsi deux groupes de cibles : cibles enregistrées dans le contexte cohérent et cibles enregistrées dans le contexte incohérent.

8.2.3 Plan d'expérience

Les variables indépendantes de cette expérience sont le contexte d'origine (cohérent *vs.* incohérent) et la cible (« Ba » *vs.* « McGurk »). La variable dépendante est le taux de perception d'effet McGurk. Encore une fois, malheureusement le logiciel utilisé ne nous a pas permis de mesurer avec la précision nécessaire le temps de réponse.

Nous avons préparé 16 stimuli originaux de chaque type et les avons combinés dans l'expérience avec les proportions : ¾ de « McGurk » versus ¼ de « ba ». Pour obtenir cette proportion nous avons répété 3 fois nos stimuli « McGurk. » Au total on a présenté 128 stimuli avec le nombre de stimuli de chaque type indiqué dans le (Tableau 13).

Tableau 13 - Le nombre des stimuli présentés dans l'expérience

A\V	**Cohérent « Ba »**	**Cohérent « McGurk »**	**Incohérent « Ba »**	**Incohérent « McGurk »**
« ba »	16 (1/8)	48 (3/8)	16 (1/8)	48 (3/8)

L'expérience consiste en un seul bloc de durée 11 minutes. Les cibles sont mélangées aléatoirement et séparées par une pause de durée 3500 ms avec un écran noir. Les conditions de passation d'expérience, la présentation des stimuli et les consignes sont les mêmes que dans l'Expérience 1 (Paragraphe 5.4.2).

8.2.4 Sujets

12 sujets français ont participé à l'expérience, ayant tous une vision et audition normale ou corrigée (3 femmes et 9 hommes, entre 22 et 59 ans avec 30,3 ans en moyenne, tous sujets droitiers). Tous les sujets ont donné un consentement éclairé à participer à l'expérience et n'étaient pas au courant du but de l'étude.

8.3 Résultats

8.3.1 Scores bruts

Nous présentons le détail des réponses des sujets dans la matrice de confusion (Tableau 14). Le taux d'erreur (réponses absentes) est **0,12** % (Figure 67). Ce taux est extrêmement faible, car en réalité, puisqu'il n'y a que des cibles, notre protocole se rapproche d'une tâche de décision plutôt que de détection de cibles.

Les réponses des sujets aux cibles « McGurk »en fonction du facteur contexte sont présentées sur la Figure 68. Nous constatons pour tous les sujets une différence de traitement perceptif entre les deux groupes de cibles. Le groupe de cibles enregistrées dans le contexte incohérent produit plus de réponses « ba »pour tous les sujets.

Tableau 14 – Matrice de confusion

Stimuli		Stimuli présentés	Réponse « ba »		Réponse « da »		Plusieurs réponses « ba »		Plusieurs réponses « da »		Absence de réponses		Absence plusieurs réponses	
Cohérent	Ba	192	190	(99%)	0	(0%)	1	(1%)	0	(0%)	0	(0%)	1	(1%)
	McG	576	131	(23%)	433	(75%)	0	(0%)	3	(1%)	6	(1%)	3	(1%)
Incohérent	Ba	192	186	(97%)	3	(2%)	1	(1%)	0	(0%)	0	(0%)	2	(1%)
	McG	576	306	(53%)	261	(45%)	2	(0%)	5	(1)%	1	(0%)	1	(0%)

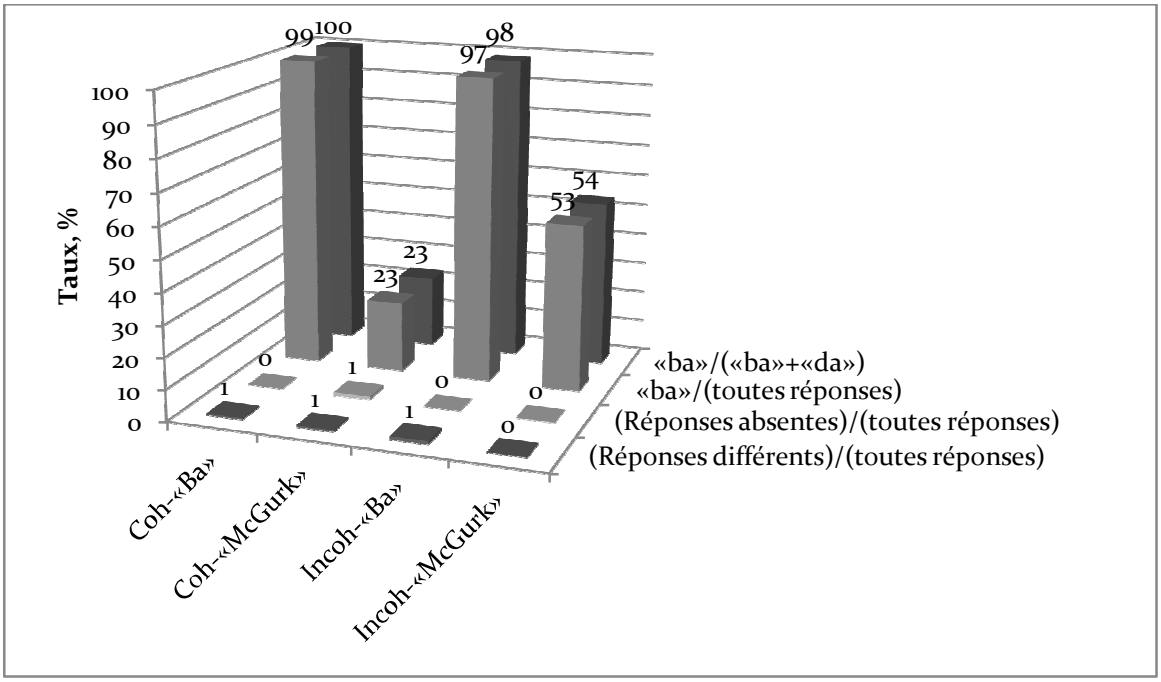


Figure 67 – Expérience 3, données brutes, %.

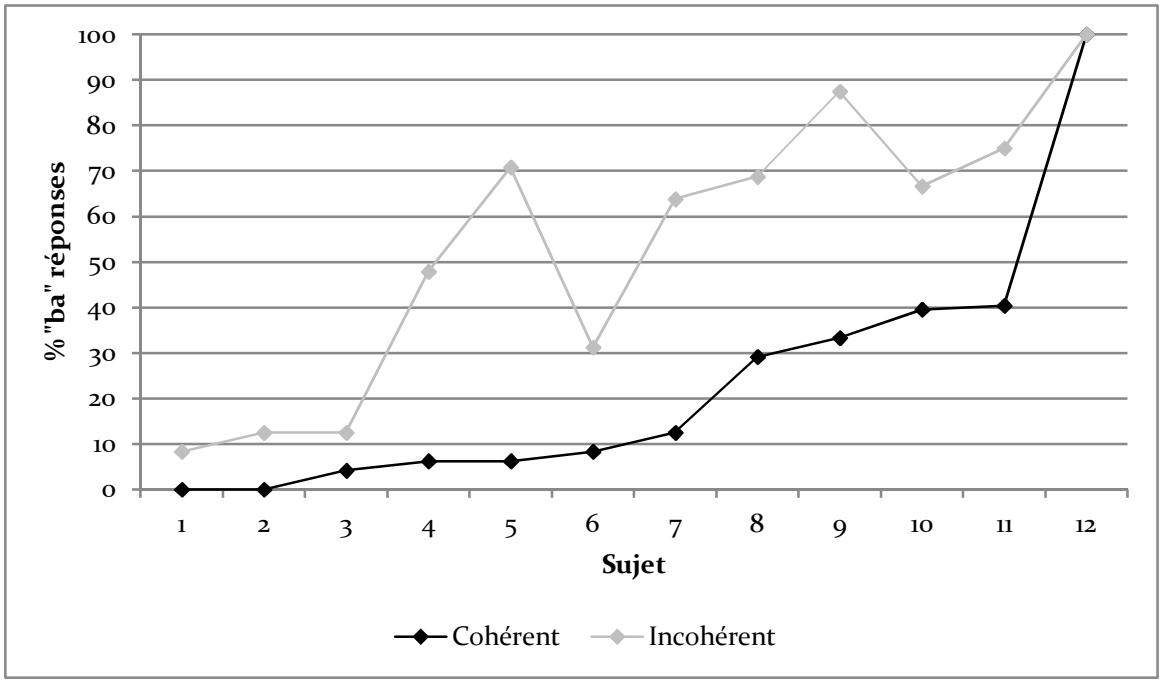


Figure 68 - Perception d'effet McGurk par sujet (« ba »/(« ba » + « da »)).

8.3.2 Analyses statistiques des pourcentages de réponse

8.3.2.1 Effets du contexte et de la cible

Les pourcentages globaux de réponse aux cibles « Ba » et « McGurk » en fonction du contexte d'enregistrement sont présentés Figure 69. Les cibles « Ba » sont toujours identifiées correctement (réponse « ba ») alors que les cibles « McGurk » sont identifiées, à 20% comme « ba », donc à 80% comme « da » en contexte cohérent et à 55% comme « ba », donc à 45% comme « da » en contexte incohérent. Ces taux de réponses « da » sont élevés par rapport aux scores classiques d'effet McGurk en français (Cathiard et al, 2001). Cependant les conditions, nous l'avons dit, sont différentes de celles d'une expérience classique sur l'effet McGurk. L'absence de stimuli « da » fait que notre tâche, qui n'implique en réalité que deux types de stimuli et deux réponses possibles, peut être interprétée comme une tâche de discrimination. C'est dans ce contexte qu'il conviendra d'interpréter les résultats.

Une ANOVA à mesures répétées sur les facteurs « contexte » et « cible » (Tableau 15) montre un effet significatif des deux facteurs (contexte [$F(1,11)=24.27$, $P<0.001$], cible [$F(1,11)=58.69$, $P<0.001$]), ainsi que de leur interaction [$F(1,11)=39.84$, $P<0.001$]. Un analyse post hoc confirme que l'augmentation du taux des réponses « ba » pour les cibles « McGurk » est significativement différent entre les contextes « cohérent » et « incohérent » ($P<0.001$).
Une nouvelle fois, l'effet sujet est significatif [$F(1,11)=383.21$, $P<0.001$] ce qui confirme que la perception d'effet McGurk est variable selon les sujets (Figure 68).

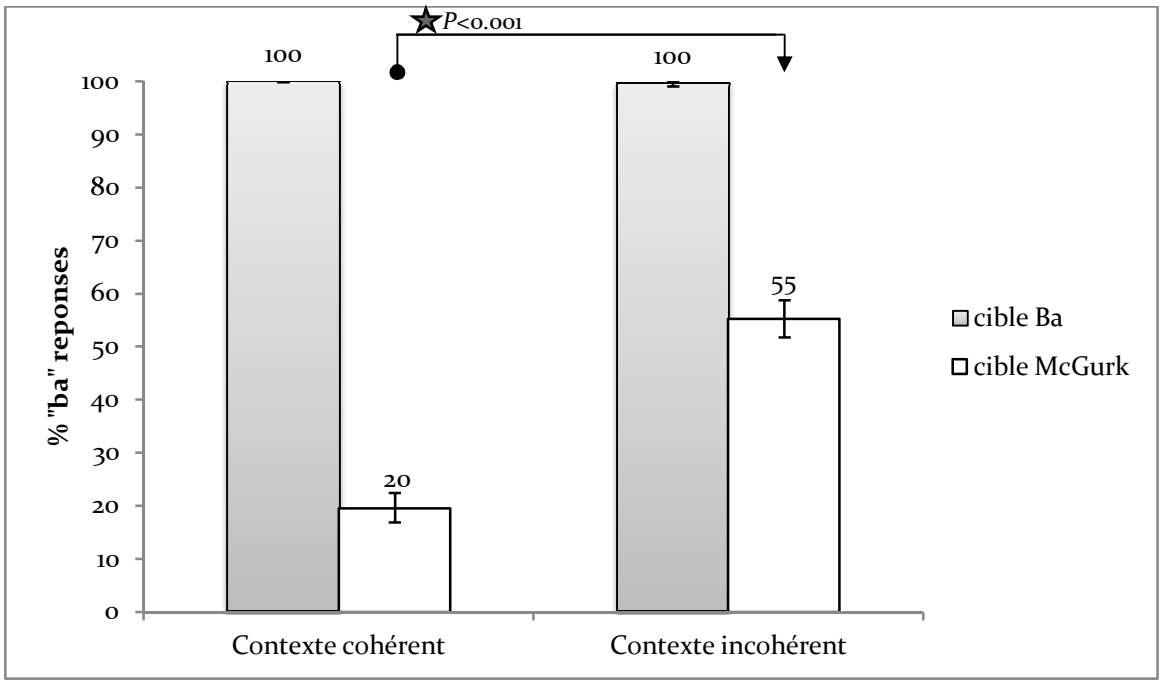


Figure 69 - Taux des réponses (« ba »/(« ba » + « da »)) dans l'Expérience 3.

Tableau 15– ANOVA à mesures répétées sur les facteurs « cible » et « contexte ».

Source		Somme des carrés de type III	ddl	Moyenne des carrés	D	Sig.
Contexte : coh, incoh	Sphéricité supposée	,301	1	,301	24,266	,000
Erreur (contexte)	Sphéricité supposée	,136	11	,012		
Cible : Ba, McGurk	Sphéricité supposée	9,524	1	9,524	58,689	,000
Erreur (cible)	Sphéricité supposée	1,785	11	,162		
*Contexte * cible*	Sphéricité supposée	,593	1	,593	39,837	,000
*Erreur (contexte*cible)*	Sphéricité supposée	,164	11	,015		

8.3.2.2 Effet du sous-groupe des cibles en contexte « incohérent »

Nous avons ensuite testé l'influence du contenu visuel des stimuli incohérents, en comparant les résultats obtenus pour les deux sous groupes, le sous groupe « incohérent- » d'ouverture relative plus faible, et le sous-groupe « incohérent+ » d'ouverture relative plus grande (Paragraphe 5.3.3, Figure 50). Une ANOVA à mesures répétées centrée sur les cibles « McGurk » pour les sous-groupes « incohérent- », « incohérent+ » et « cohérent » (Tableau 16) montre un effet significatif du facteur sous groupe [$F(2,22)=23.185$, $P<0.001$]). Une analyse posthoc montre que cet effet significatif est lié à la différence entre le sous-groupe « cohérent » et les deux sous-groupes « incohérent- » et « incohérent+ ». Ainsi, il n'y a pas de différence de perception entre les sous-groupes « incohérent- » et « incohérent+ ».

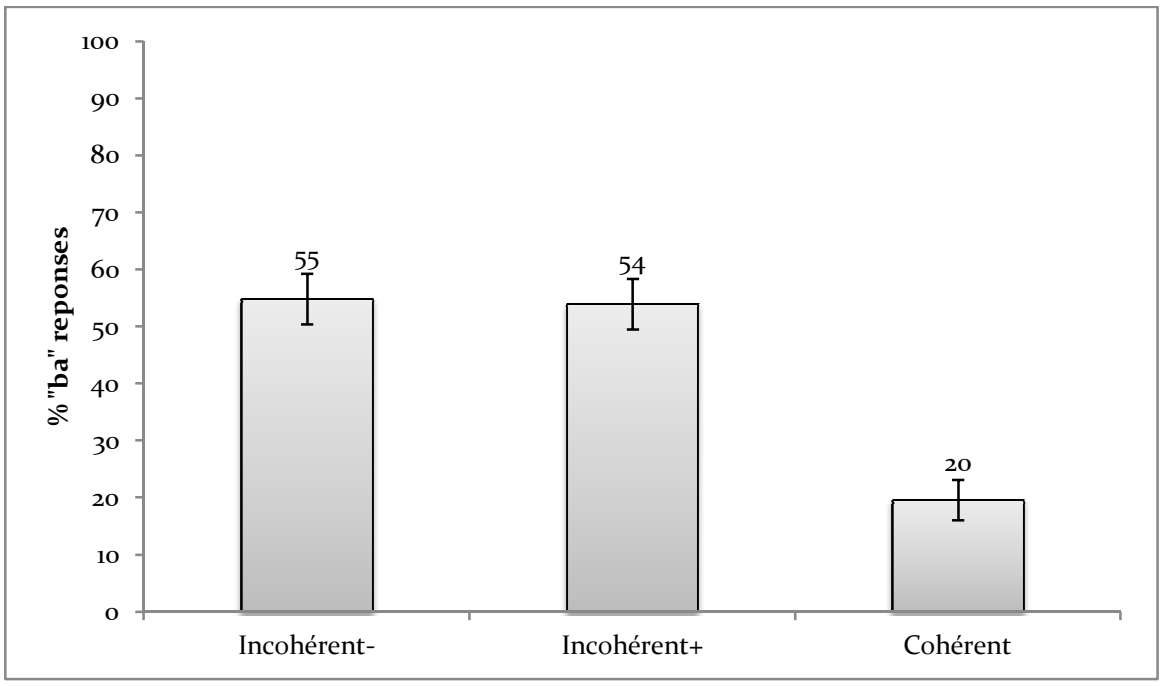


Figure 70 - Taux de réponses (« ba »/(« ba » + « da »)) pour les deux sous-groupes en fonction de la valeur d'ouverture des lèvres (amplitude relative Δy>0.8 et Δy<0.8)

Tableau 16– ANOVA à mesures répétées: cible, contexte.

Source		Somme des carrés de type III	ddl	Moyenne des carrés	D	Sig.
Sous-groupe : incoh-, incoh+, coh	Sphéricité supposée	1,106	2,000	,553	23,185	,000
Erreur (Sous-groupe)	Sphéricité supposée	,525	22,000	,024		

8.4 Conclusion

Les résultats de cette expérience montrent la différence significative de perception des cibles« McGurk » isolées, selon qu'elles ont été enregistrées en contexte antérieur cohérent ou incohérent. L'analyse des sous-groupes de cibles isolées nous permet d'évaluer la différence entre les cibles. Malgré les contrôles et mesures sur les cibles effectués précédemment, les deux sous-groupes de cibles extraites du contexte incohérent donnent un percept différent de celui des cibles enregistrées en contexte cohérent. Ce résultat peut remettre en cause notre interprétation des résultats des Expériences 1 et 2, en termes d'influence du contexte antérieur.

Evidemment, la tâche proposée n'est pas réellement, nous l'avons dit, une tâche de type « mise en évidence de l'effet McGurk » : le contexte simplificateur incluant seulement deux types de stimuli et de réponses a probablement induit une tâche de type discrimination entre deux ensembles de stimuli. On peut ainsi se demander si les sujets ne s'appuient pas uniquement sur la vue de la bouche ouverte *vs*. fermée au démarrage pour répondre et si un effet McGurk est réellement perçu ici (il s'agirait alors d'une réponse uniquement visuelle et non auditive). Néanmoins les résultats montrent que les sujets discriminent clairement les cibles « McGurk » selon leur contexte, et ce de manière homogène à l'intérieur du contexte « incohérent ». Notons qu'à l'inverse, la Figure 69 montre aussi que les cibles « McGurk », même en contexte incohérent, restent très clairement différentes des cibles « Ba », alors que les résultats des Expériences 1 et 2 montrent que l'effet McGurk en contexte incohérent (donc avec ces cibles) est quasiment supprimé (voir par exemple Figure 66). Ceci ne peut s'expliquer sans invoquer un effet contextuel.

Reste que, globalement, il apparaît que la différence entre les groupes de cibles dans cette expérience peut partialement remettre en question notre interprétation de l'effet contexte sur la perception d'effet McGurk dans les deux premières expériences. La prochaine expérience vise à lever ce biais possible et à tenter de mettre en évidence de manière incontestable l'effet du contexte sur la perception d'effet McGurk.

Chapitre 9. Expérience 4 : Validation de l'effet contexte

9.1 Objectifs et hypothèses

L'expérience précédente a mis en évidence des différences de taux d'effet McGurk entre cibles isolées selon les conditions contextuelles de production, remettant ainsi en cause potentiellement notre interprétation des résultats des Expériences 1 et 2 en termes d'effet du contexte antérieur sur la fusion audiovisuelle. Le but de cette nouvelle expérience est de proposer un paradigme adéquat, incluant un contrôle parfait des cibles, permettant d'évaluer sans ambigüité et le cas échéant de démontrer l'effet du contexte préalable et l'existence de processus du liage en évitant les biais provoqués par les différents groupes de cibles, selon qu'elles sont enregistrées en contexte cohérent ou incohérent.

9.2 Méthodologie

9.2.1 Principe

Pour éviter le biais provoqué par les différents groupes des cibles, il est nécessaire d'utiliser les mêmes cibles dans les deux conditions de contexte. Ceci implique d'abandonner notre exigence initiale de continuité visuelle parfaite entre le contexte et la cible. La rupture de continuité risquait selon nous, a priori, de fournir aux sujets une indication temporelle et une focalisation attentionnelle sur l'instant d'arrivée de la cible et ainsi de détruire un éventuel effet de déliage.

Notre interprétation des deux premières expériences est que, même si les différences de cibles peuvent expliquer pour partir l'effet de contexte, cet effet est trop fort pour ne pas impliquer, au moins partiellement, un mécanisme de déliage. Nous avons donné en conclusion du chapitre précédent un élément de raisonnement en faveur de cette interprétation – même si le doute subsiste, et implique la mise en place d'un nouveau paradigme. Dans ce cadre, le résultat phare de l'Expérience 2 est précieux : il suggère qu'une alerte audiovisuelle fournie peu avant la cible ne semble pas produire pas d'effet sur les résultats, et maintient la suppression quasi totale d'effet McGurk en contexte incohérent. Nous allons utiliser cette propriété pour créer une transition entre le contexte et la cible. Cette transition nous permettra de conserver un même groupe de cibles dans les deux conditions de contexte.

Ainsi, nous modifions légèrement le paradigme expérimental de l'Expérience 1 (Figure 38) de façon à ajouter une transition entre le contexte et la cible (Figure 71). Les autres conditions restent inchangées : 2 types de contexte (cohérent vs incohérent), 2 types de cibles (« Ba » vs « McGurk »), 4 durées de contexte (5, 10, 15, 20 syllabes).

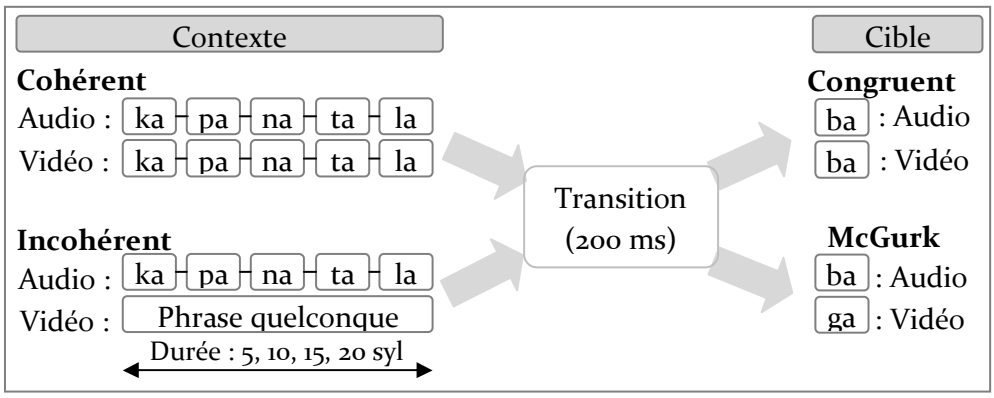


Figure 71 - Schéma de l'Expérience 4

9.2.2 Stimuli

9.2.2.1 *Séparation contexte - cible*

Les stimuli utilisés dans cette expérience sont issus du corpus de l'Expérience 1, et sont décrits en détail dans le §5.2 et le §5.3. Dans notre corpus les séquences sont composées à la fois des contextes et des cibles, enregistrés de façon continue. Pour pouvoir utiliser le même groupe de cibles dans les deux conditions, nous avons séparé les séquences sur deux parties : contexte et cible. L'instant de coupure est situé 240 ms avant le burst d'une cible. Ensuite la durée de chaque cible est égalisée à une valeur minimale de 960 ms (Figure 72). Pour enchaîner le contexte avec la cible nous appliquons une transition selon le principe décrit dans une section ultérieure (§9.2.2.4).

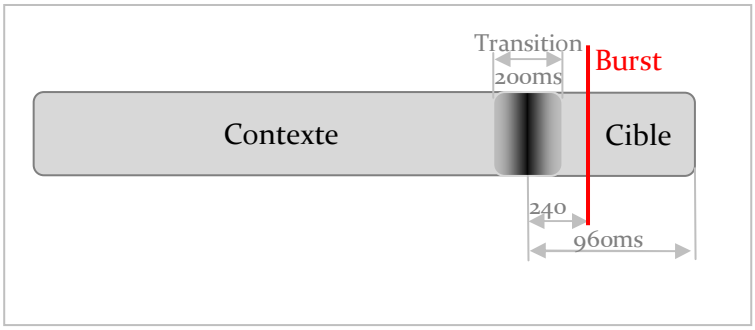


Figure 72 – Modèle d'un stimulus

9.2.2.2 *Sélection des cibles*

Nous devons choisir un groupe de cibles à tester dans les deux conditions de contexte. Selon les résultats de l'expérience 3, les cibles en contexte « cohérent » produisent isolément plus d'effet McGurk (moins de réponses « ba ») que les cibles produites en contexte « incohérent », nous les avons donc sélectionnées. Parmi les 16 cibles dont nous disposons nous avons choisi les 12 cibles qui produisaient dans l'Expérience 3 le moins de réponses « ba » (le plus d'effet McGurk). Les cibles « Ba » étaient toutes à peu près équivalentes, nous en avons sélectionné 4, produites également en contexte « cohérent ».

9.2.2.3 *Traitement*

Pour enchainer de manière adéquate des cibles et des contextes parfois différents, nous avons dû homogénéiser l'intensité auditive des contextes et des cibles. Pour cela nous avons aligné les intensités sonores de chaque stimulus à la moyenne des intensités de tous les contextes utilisés. Nous avons effectué la même opération pour toutes les cibles. Pour calculer la puissance nous avons appliqué la formule :

$$power = 20 \times \log_{10}(std(audio))$$

Et pour recalculer l'intensité :

$$audio' = audio \times 10^{\frac{power' - power}{20}},$$

où *power*' est la moyenne de tous les *power*.

9.2.2.4 *Montage des stimuli complets*

Avant de procéder à la génération des stimuli complets (contexte plus cible), nous avons généré le tableau des correspondances entre le contexte et la cible, en appliquant certains règles de contrôle. Pour les deux contextes nous procédons de la même façon décrite ci-dessous.

Les 12 cibles « McGurk » ont été associées à chaque condition de durée 5, 10, 15 et 20 syllabes, de façon à pouvoir (enfin !) tester l'effet de durée du contexte. Nous avons 4 exemplaires pour chaque durée de contexte, ainsi les 12 cibles sont distribuées aléatoirement entre ces 4 exemplaires, soit 3 cibles « McGurk » par exemplaire.

Les 4 cibles « Ba » sont de même affectées aléatoirement aux 4 exemplaires de chaque valeur de durée, soit une cible Ba par exemplaire d'un contexte dans une durée donnée.

Afin de créer un stimulus complet (contexte plus cible) il nous faut monter les parties « contexte » et « cibles » issues de séquences différentes. Notons que, comme le contexte incohérent produira par définition une discontinuité (puisque les cibles sont extraites de contextes « cohérents »), dans le cas de contextes cohérents nous avons également imposé systématiquement une discontinuité en veillant à ce qu'une cible ne soit jamais associée à un contexte qui était son contexte d'origine. Pour limiter l'effet visuel de ces discontinuités et rendre le passage le plus lisse possible nous avons appliqué systématiquement une transition progressive en passant par une image noire sur 200 ms, soit un « fading » sur 5 images avec un niveau du fond noir évoluant selon une dynamique 1/3-2/3-1-2/3-1/3 (Figure 73) :

$$X' = PN + (1 - P)X,$$

où X est une image d'origine, N est une image noire, P la proportion.

Figure 73 - Fondu noir progressif sur 5 images

9.2.3 Plan d'expérience

Comme dans l'Expérience 1, nous avons les variables indépendantes : contexte (cohérent vs incohérent), cible (« Ba » vs « McGurk »), durée du contexte (5, 10, 15 et 20 syllabes) et une première variable dépendante qui est le taux de perception d'effet McGurk. C'est à partir de cette expérience (à l'unique exception de la suivante) que nous avons mis en place le passage sur le logiciel Presentation® software (Version 0.70, www.neurobs.com), qui nous a permis de mesurer précisément les temps de réponse, qui seront donc analysés à partir de maintenant.

Nous conservons les proportions ¾ des cibles « McGurk » et ¼ des cibles « Ba ». Chaque stimulus complet (contexte = cible) est présenté une seul fois, soit au total 128 stimuli avec une répartition du nombre de stimuli de chaque type indiquée dans le Tableau 17.

Tableau 17 - Nombre de stimuli présentés dans l'Expérience 4

A\V	Cohérent « Ba »	Cohérent « McGurk »	Incohérent « Ba »	Incohérent « McGurk »
« ba »	16 (1/8)	48 (3/8)	16 (1/8)	48 (3/8)

Nous avons décomposé l'expérience sur 4 blocs de 4 minutes environ avec des pauses entre les blocs. Chaque bloc consiste en 32 stimuli, choisis de façon aléatoire, parmi tous les stimuli. L'ordre du passage des blocs est varié selon les sujets. La tâche, comme précédemment, consiste en une réponse « ba » ou « da » en ligne.

En résumé, l'expérience consiste en :

- 128 stimuli en total
- 4 blocs de 32 stimuli, répartis aléatoirement
- 2 types de réponse possibles « ba » ou « da » en ligne, choix forcé

9.2.4 Sujets

19 sujets français ont participé à l'expérience avec vision et audition normale ou corrigée (6 femmes et 13 hommes, entre 20 et 27 ans avec 23,5 ans en moyenne, 17 droitiers et 2 gauchers). Tous les sujets ont donné un consentement éclairé à participer à l'expérience et n'étaient pas au courant du but de l'étude.

9.3 Résultats

9.3.1 Scores bruts

Nous présentons les réponses des sujets dans la matrice de confusion (Tableau 18). Le taux d'erreur (réponses absentes) est de **6,37 %**. Cette diminution du taux d'erreur par rapport aux Expériences 1 et 2 peut être expliquée par différents facteurs. D'abord il faut prendre en compte que nous avons rajouté une alerte temporelle créée par la transition contexte-cible, qui facilite la tâche de détection de la cible. Ainsi, le nombre de réponses absentes diminue par rapport à celui des Expériences 1 et 2, tandis que le nombre des réponses multiples reste au même niveau (Tableau 18, Figure 74). Ensuite il faut mentionner que pour cette expérience nous avons choisi les meilleures cibles, qui produisent le plus d'effet McGurk, ce qui peut également avoir facilité le traitement perceptif des cibles.

Tableau 18 – Matrice de confusion

Stimuli		Stimuli présentés	Réponse « ba »		Réponse « da »		Plusieurs réponses « ba »		Plusieurs réponses « da »		Absence de réponses		Absence plusieurs réponses	
Cohérent	Ba	304	251	(83%)	2	(1%)	20	(7%)	0	(0%)	2	(1%)	29	(10%)
	McG	912	432	(47%)	362	(40%)	40	(4%)	15	(2%)	14	(2%)	49	(5%)
Incohérent	Ba	304	256	(84%)	3	(1%)	29	(10%)	0	(0%)	3	(1%)	13	(4%)
	McG	912	560	(61%)	238	(26%)	69	(8%)	0	(0%)	7	(1%)	38	(4%)

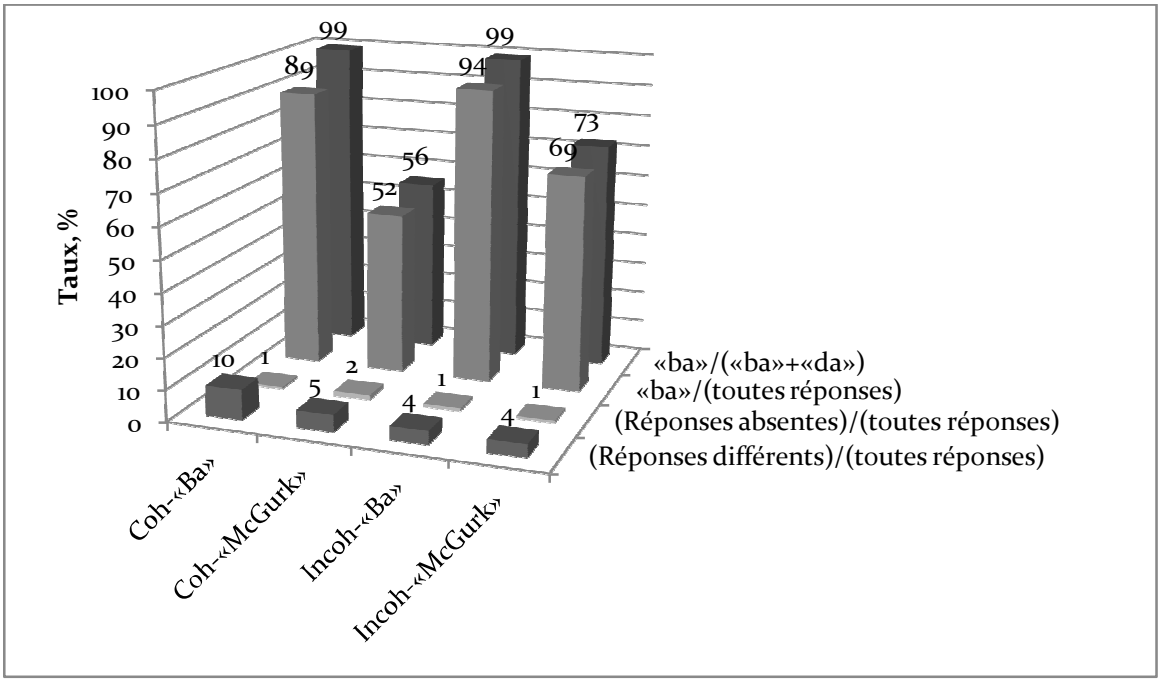


Figure 74 – Expérience 4, données brutes, %.

L'analyse des réponses des participants aux cibles McGurk selon le contexte est présentée sur la Figure 75, en triant les participants selon leur score de réponses aux cibles en contexte cohérent. Malgré les différences fortes de l'effet McGurk selon les sujets, une fois encore en accord avec la littérature (Schwartz, 2010), il apparaît que dans le contexte incohérent le taux d'effet McGurk diminue pour tous les sujets par rapport au contexte cohérent (sauf pour quelques sujets qui présentent un score à 0% ou à 100% dans les deux contextes).

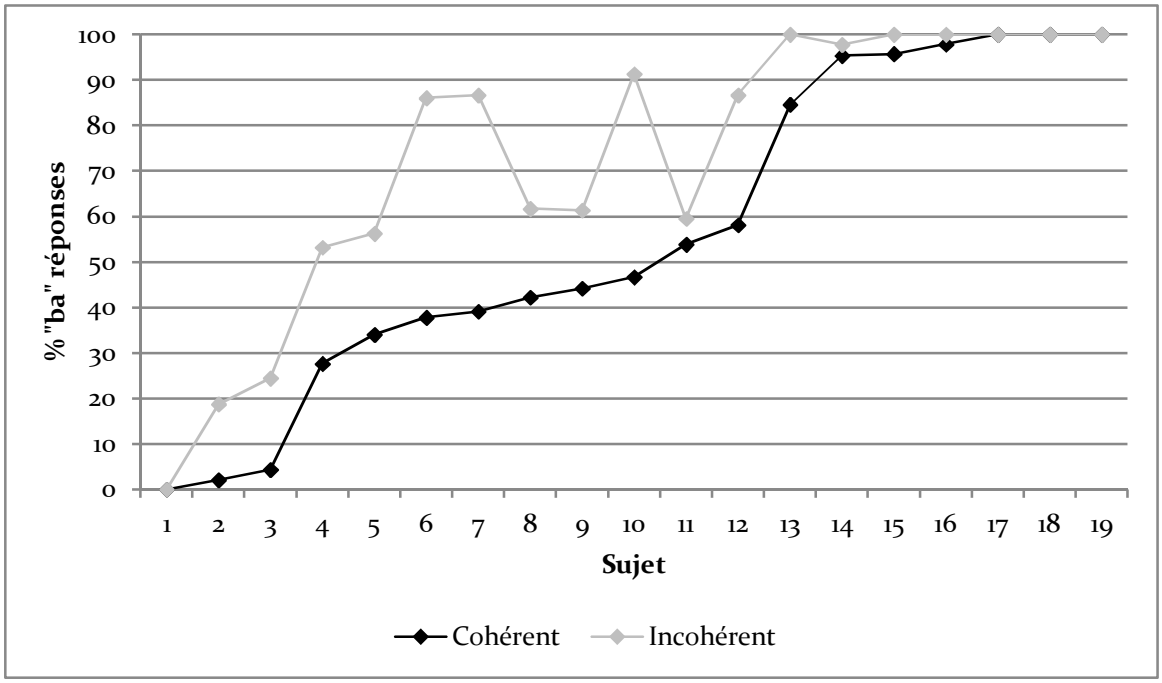


Figure 75 - Perception d'effet McGurk par sujet (« ba »/(« ba » + « da »)).

9.3.2 Analyses statistiques des pourcentages de réponse

9.3.2.1 Effet de la cible et du contexte

Regardons les pourcentages globaux de réponse aux cibles « Ba » et « McGurk » (Figure 76). Il apparaît que les cibles « Ba » sont presque toujours identifiées correctement (réponse « ba ») alors que les cibles McGurk sont identifiées en contexte cohérent, à 60% comme « ba », donc à 40% comme « da ». On retrouve une fois encore les scores classiques d'effet McGurk en français (Cathiard et al, 2001).

En contexte incohérent, le score de réponses « ba » augmente à 80% (la perception d'effet McGurk diminue à 80%). Une ANOVA à mesures répétées sur les facteurs « contexte » et « cible » (Tableau 19) montre un effet significatif des deux facteurs (contexte [$F(1,18)=20.47$, $P<0.001$], cible [$F(1,18)=26.5$, $P<0.001$]), ainsi que de leur interaction [$F(1,18)=22.83$, $P<0.001$]. Un analyse post hoc confirme que l'augmentation du taux de réponses « ba » pour les cibles « McGurk » du contexte cohérent au contexte incohérent est significative ($P<0.001$). Cette analyse nous confirme – cette fois avec des cibles parfaitement contrôlées - que la perception d'effet McGurk dépend du contexte préalable et nous permet ainsi de valider notre hypothèse d'existence d'un processus de liage audiovisuel modulant la fusion.

L'effet sujet est également significatif [$F(1,18)=516.21$, $P<0.001$] (Figure 75).

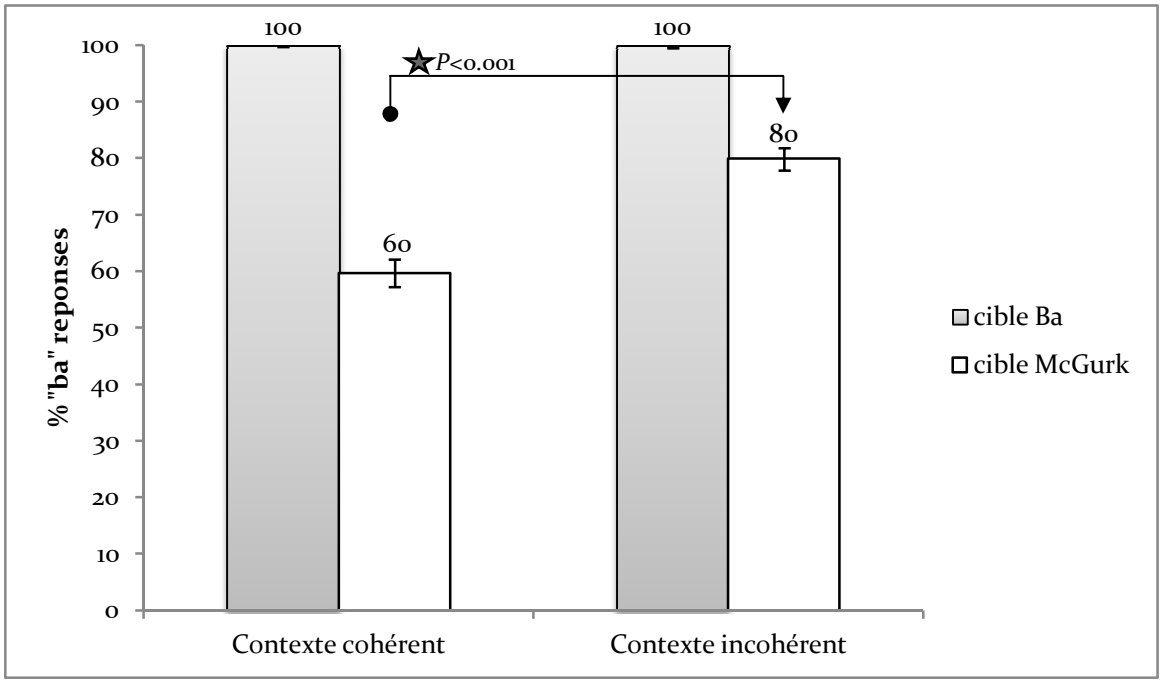


Figure 76 - Taux des réponses (« ba »/(« ba » + « da »)) dans l'Expérience 4

Tableau 19 – ANOVA à mesures répétées : cible, contexte.

Source		Somme des carrés de type III	ddl	Moyenne des carrés	D	Sig.
Contexte : coh, incoh	Sphéricité supposée	,209	1	,209	20,472	,000
Erreur (contexte)	Sphéricité supposée	,184	18	,010		
Cible : Ba, McGurk	Sphéricité supposée	5,563	1	5,563	26,495	,000
Erreur (cible)	Sphéricité supposée	3,779	18	,210		
*Contexte * cible*	Sphéricité supposée	,267	1	,267	22,837	,000
*Erreur (contexte*cible)*	Sphéricité supposée	,211	18	,012		

9.3.2.2 *Effet de la durée du contexte*

Comme nous l'avons mentionné au chapitre précédent, les variations – aléatoires – des cibles en fonction de la durée du contexte dans le paradigme précédent nous avaient empêché jusqu'à présent de tester l'effet de la durée du contexte sur la modulation de l'effet McGurk. Dans cette expérience nous avons pris toutes les précautions pour pouvoir enfin tester cet effet. Sur la Figure 77 nous présentons la perception d'effet McGurk selon la durée du contexte préalable. Une ANOVA à mesures répétées centrée sur les cibles « McGurk » confirme l'effet significatif du facteur contexte [$F(1,18)=25.21$, $P<0.001$] et montre aussi un effet en limite de significativité du facteur durée du contexte [$F(3,54)=2.78$, $P=0.049$]. L'interaction des deux facteurs n'est pas significative [$F(3,54)=.26$, $P=0.3$]. L'analyse post hoc ne fournit d'effet significatif sur aucune des paires. Les variations selon le contexte apparaissent selon la Figure 77 suivre une courbe en U, similaire pour les deux contextes, et dont l'interprétation n'est pas claire. Etant donnée la faiblesse de l'effet, en limite de significativité, nous en attendons confirmation avant de proposer une interprétation théorique.

Par contre, un point important à noter est l'absence de différence entre scores de réponses en contexte incohérent des durées courtes (5 syllabes) aux durées longues (20 syllabes), ce qui montre que la modulation de l'effet McGurk attribuable dans notre raisonnement à un mécanisme de déliage en contexte incohérent apparaît dès les durées les plus courtes. Ainsi, 5 syllabes (soient 3 secondes environ d'incohérence) sont suffisantes pour délier significativement les flux auditifs et visuels.

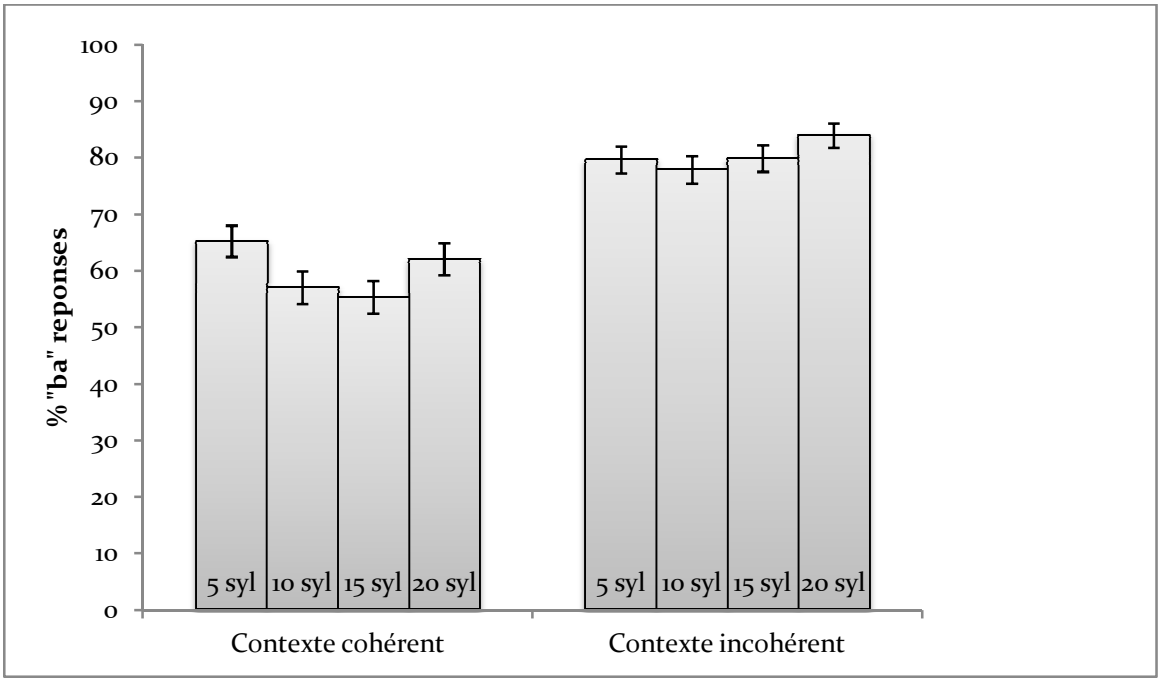


Figure 77 - Taux de réponses (« ba »/(« ba » + « da »)) selon le contexte et la durée dans l'Expérience 4.

Tableau 20 - ANOVA à facteurs multiples: durée, contexte, sujet

Source		Somme des carrés de type III	ddl	Moyenne des carrés	D	Sig.
Contexte : coh, incoh	Sphéricité supposée	1,959	1	1,959	25,211	,000
Erreur (contexte)	Sphéricité supposée	1,398	18	,078		
Durée : 5, 10, 15, 20 syl	Sphéricité supposée	,126	3	,042	2,785	,049
Erreur (cible)	Sphéricité supposée	,814	54	,015		
*Contexte * durée*	Sphéricité supposée	,061	3	,020	1,258	,298
*Erreur (contexte*cible)*	Sphéricité supposée	,874	54	,016		

9.3.3 Analyse des temps de réponses

Le passage à l'utilisation du logicielle Presentation® software (Version 0.70, www.neurobs.com) nous permet dans cette expérience d'analyser les temps de réponse. Rappelons, qu'avant d'effectuer l'analyse de la variance nous effectuons une transformation logarithmique pour assurer la gaussianité des données.

Sur la Figure 78 on observe une variation du temps de réponse selon les cibles, mais pas selon le contexte. Ainsi les cibles McGurk induisent une augmentation de temps de réponse d'environ 25-30 ms et ce de manière similaire dans les deux contextes. Une ANOVA à mesures répétées (Tableau 21) donne des résultats non significatifs pour le facteur contexte [$F(1,18)=0.77$, $P=0.39$] et pour le facteur cible [$F(1,18)=2.91$, $P=0.11$], ainsi que pour leur interaction [$F(1,18)=0.1$, $P=0.76$]. Cependant, une analyse complémentaire que nous avons menée en considérant le facteur sujets non comme aléatoire mais comme fixe fournit un effet « cible » significatif ([$F(1,18)=6.35$, $P=0.021$], ce qui suggère qu'il y a effectivement une possible différence de temps de réponse entre cibles congruentes Ba et incongruentes McGurk.

Cet effet, qui demande confirmation – et qui sera effectivement confirmé dans la Partie III – fournit un premier élément de réponse à une question que l'on pourrait se poser sur nos résultats. Dans la tâche de monitoring en ligne, nous ne contrôlons pas si le sujet regarde le locuteur. Nos consignes sont évidemment de maintenir le regard fixé sur l'écran en permanence. On peut néanmoins se demander si les sujets n'écartent pas leur regard de la cible en contexte incohérent, ayant noté que l'image ne correspond pas au son. Cette hypothèse est très peu probable étant donné le paradigme dans lequel tous les stimuli sont mélangés, avec une alternance aléatoire de stimuli à contexte incohérent et cohérent. Cependant, l'absence d'utilisation d'un système de contrôle de la direction du regard (eye tracker) ne permet pas de valider cette assertion de manière irréfutable. Néanmoins, le fait que l'on observe une augmentation du temps de réponse pour les cibles « McGurk » indépendamment du contexte suggère que les sujets perçoivent bien une incongruence et donc

qu'ils continuent à fixer la cible avec autant d'attention quel que soit le contexte. Ce résultat est également en soi surprenant : le déliage semble moduler la réponse des sujets aux cibles « McGurk », mais pas leur temps de traitement. Nous reviendrons sur ce point une fois qu'il sera confirmé expérimentalement dans la Partie III.

Notons pour finir, une fois encore, la large variation de comportement des sujets [$F(1,18)=15937.22$, $P<0.001$].

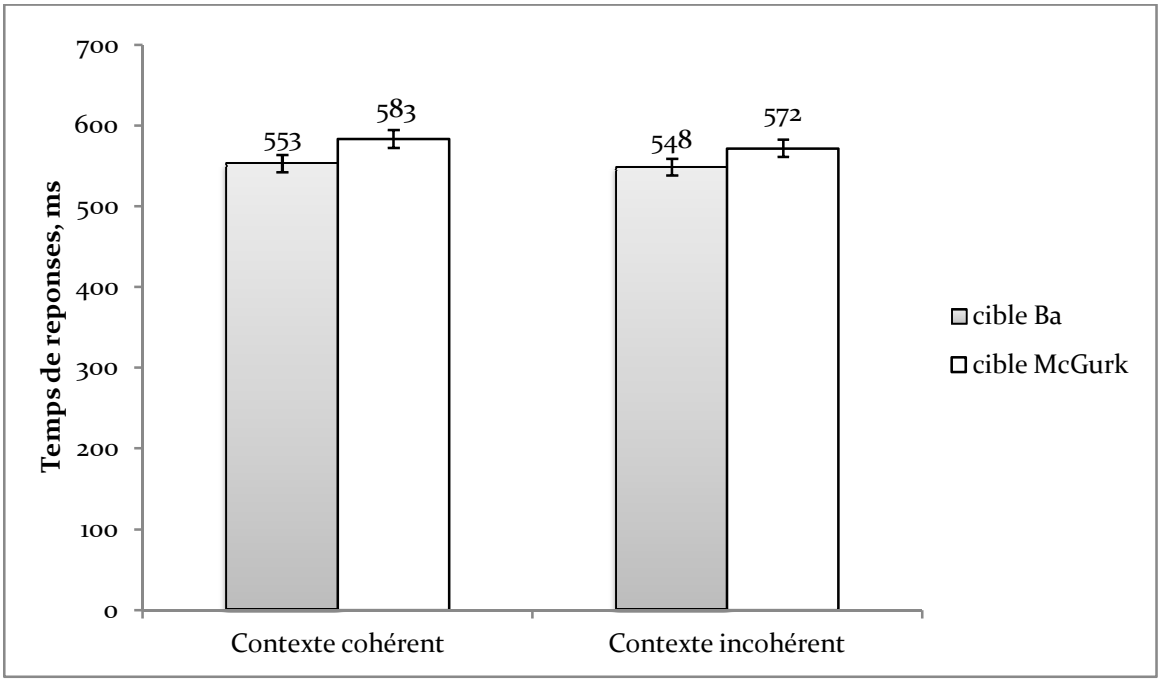


Figure 78 - Temps de réponses (en ms) selon la cible et le contexte dans l'Expérience 4

Tableau 21 - ANOVA sur les temps de réponses selon les facteurs : cible, contexte, sujet

Source		Somme des carrés de type III	ddl	Moyenne des carrés	D	Sig.
Contexte : coh, incoh	Sphéricité supposée	,004	1	,004	,767	,393
Erreur (contexte)	Sphéricité supposée	,087	18	,005		
Cible : Ba, McGurk	Sphéricité supposée	,043	1	,043	2,913	,105
Erreur (cible)	Sphéricité supposée	,269	18	,015		
*Contexte * cible*	Sphéricité supposée	,001	1	,001	,098	,758
*Erreur (contexte*cible)*	Sphéricité supposée	,123	18	,007		

9.4 Discussion

Les résultats de cette expérience montrent que nous avons bien réussi à moduler la perception d'effet McGurk par la différence entre les contextes cohérent et incohérent. L'effet de modulation est moins élevé que dans les Expériences 1 et 2. En effet, le score de réponses « ba » pour les cibles « McGurk » en contexte incohérent a baissé de 95% dans les Expériences 1 et 2 à 80% dans l'Expérience 4, et l'écart par rapport au score de base en contexte cohérent a été réduit presque de moitié. L'origine de cette diminution de l'effet peut être due soit au choix des cibles (nous avons choisi des cibles qui donnent le plus d'effet McGurk), soit au contrôle des variations entre cibles selon le contexte (confirmant ainsi a posteriori que nous avons eu raison d'effectuer cette expérience contrôle avec un paradigme permettant de maîtriser parfaitement les cibles). Il n'est pas impossible non plus que l'ajout d'une transition ait fourni une information visuelle qui limite les effets de déliage. Reste que cette diminution de modulation ne remet pas en cause l'existence du processus du liage, car la différence entre les deux contextes est bel et bien significative. Ainsi nous pouvons valider notre hypothèse de base. Un point très important est du coup que nous disposons maintenant d'un paradigme souple et efficace, qui permet de monter n'importe quel contexte et n'importe quelle cible, et de produire des effets potentiels de modulation contextuelle malgré la petite discontinuité visuelle imposée par le montage.

Les variations de durée de contexte entre 3 et 10 s produit un effet en limite de significativité, que nous ne savons interpréter pour l'instant et qui, devant sa faiblesse, demande confirmation. Un résultat important est que l'effet de déliage est déjà maximal pour la durée du contexte la plus courte, montrant que 3 s ou 5 syllabes de contexte incohérent sont suffisantes pour délier les deux flux. Pour étudier le liage et déliage sur des durées de contexte plus courtes que 3 s ou 5 syllabes il nous faudra mettre en place une nouvelle expérience.

Les temps de réponse ne semblent pas dépendre du contexte, mais peut-être de la cible, avec un temps de réponse plus élevé des cibles « McGurk » incongruentes, indépendamment du contexte. Ce résultat, surprenant et très intéressant, devra être confirmé dans les expériences suivantes, nous y reviendrons.

Disposant ainsi d'une mise en évidence claire et qui nous semble convaincante des effets de modulation contextuelle de l'effet McGurk, que nous interprétons en termes de déliage audiovisuel, et disposant également d'un paradigme d'étude bien contrôlé et apparemment efficace, nous allons maintenant, dans la Partie suivante, analyser quantitativement plus finement les mécanismes de liage-déliage.

Partie III
Caractérisation du processus du liage

Dans les expériences de la Partie précédente, nous avons progressivement affiné le paradigme expérimental qui nous a permis de montrer que la fusion audio-visuelle n'est pas automatique et dépend du contexte préalable. Nous avons ainsi mis en évidence ce que nous considérons être un « processus de liage audiovisuel » conditionnant la fusion et modulant l'effet McGurk. Pour ce faire, nous avons testé introduit deux types de contexte, l'un cohérent et l'autre incohérent avec un niveau d'incohérence maximale, et l'un et l'autre de durées relativement longues (de 5 à 20 syllabes). Nous avons obtenu des effets relativement forts, et ce dès les premières durées utilisées (5 syllabes). Dans les expériences de cette Partie III, nous allons nous attacher à étudier et décrire le fonctionnement de ce processus de liage plus en détail, et ce en trois temps. D'abord, nous proposerons des niveaux d'incohérence moins élevés avec des types d'incohérence variés (Expérience 5). Puis nous testerons des durées de contexte plus courtes, afin d'explorer ce qui se passe en dessous de 5 syllabes. Enfin, nous étudierons la possibilité de « relier » les flux auditif et visuel après un contexte incohérent susceptible de produire un premier effet de déliage comme dans la Partie II.

Chapitre 10. Expérience 5 : Décomposition de l'incohérence sur les dimensions phonétique et temporelle

10.1 Objectifs et hypothèses

La première question que nous nous posons dans cette troisième partie est celle des composantes principales d'une paire de signaux auditif et visuel qui permettent de déterminer jusqu'à quel point les deux flux sont cohérents ou non. Dans la Partie précédente, pour maximiser les chances de démontrer un effet de déliage dû au contexte, nous nous sommes toujours placés dans des conditions d'incohérence maximale entre les flux auditif et visuel, avec dans le flux audio une séquence régulière de syllabes, et dans le flux vidéo une séquence libre de phrases improvisées. L'incohérence est alors située à tous les niveaux d'information : phonétique, lexical, sémantique.

Dans cette expérience nous souhaitons étudier spécifiquement quelles composantes des signaux acoustique et visuel définissent la cohérence. Pour cela, nous allons définir deux composantes principales susceptibles de participer aux processus de liage/déliage à bas niveau. La première est la cohérence des modulations temporelles. Nous avons vu dans la Section 3.3 du chapitre 3 le rôle a priori essentiel des corrélations audiovisuelles, typiquement entre les variations de la position de la mâchoire ou de l'ouverture des lèvres, et les variations de l'intensité globale ou dans une bande de fréquence donnée (par exemple autour du second formant) au cours du temps. C'est cette première composante que nous manipulons dans cette expérience. En regard, nous nous intéressons à une composante de cohérence phonétique.

Dans cette expérience nous allons donc décomposer l'incohérence selon deux dimensions, l'incohérence phonétique et temporelle, par exemple en présentant des flux parfaitement synchrones temporellement, mais incohérents phonétiquement, ou identiques du point de vue phonique mais décalés légèrement dans le temps pour rompre la cohérence temporelle. Notre hypothèse est que chacune de ces dimensions d'incohérence pourrait produire un effet de déliage, que nous souhaitons évaluer plus précisément ici

10.2 Méthodologie

10.2.1 Principe

Nous avons repris le paradigme général de l'Expérience 1 (Figure 38), où nos stimuli consistent en contexte suivi d'une cible, avec deux types de cible: « Ba » et « McGurk ». Pour définir le contexte, nous mettons en œuvre plusieurs types d'incohérence : phonétique, temporel et phonético-temporel, comparés à un contexte cohérent qui a fourni le matériau de base pour la préparation des stimuli (Figure 79). Donc au total nous avons 4 types de contexte :

- Cohérent (C) qui consiste en une séquence de syllabes CV (en français), C étant une plosive ou une fricative, à l'exclusion des syllabes « ba », « da » et « ga », soit 13 syllabes possibles (« pa », « ta », « va », « fa », « za », « sa », « ka », « ra », « la », « ja », « cha », « ma », « na »). Le contexte cohérent, comme tous les contextes incohérents qui suivent, a une durée de 5, 10, 15 ou 20 syllabes.
- Incohérent phonétique (P) qui consiste en une séquence de syllabes synchrones mais incongruentes entre les flux audio et vidéo. Nous les avons créées en partant du flux vidéo du contexte cohérent, et en introduisant à la même position temporelle le son d'une autre syllabe. Pour assurer une incohérence perceptive claire, nous avons déterminé les groupes de syllabes correspondant à un même visème: visème bilabial [pa, ma], labiodental [fa, va], dental [ta, na, sa, za], palatal [cha, ja], palato-vélaire [ka, la, ra, ga] et nous avons appliqué une règle qui interdit les permutations à l'intérieur de chaque groupe. Ainsi, ce que nous appelons par commodité incohérence phonétique est en réalité une incohérence visémique.
- Incohérent temporel (T) qui consiste en une séquence de syllabes congruentes phonétiquement mais qui ne sont plus synchrones. Pour ce faire, nous avons décalé les syllabes audio par rapport aux syllabes vidéo sur l'axe temporel dans l'ensemble des valeurs [-30; 20; 70; 120; 170] ms. Pour des durées de contexte de 5, 10, 15 et 20 syllabes, nous tirons aléatoirement dans cet ensemble de 5 décalages possibles, de façon à appliquer une permutation complète des 5 valeurs sur 5 syllabes consécutives. . Van Wassenhove et al. (van Wassenhove et al, 2007) ont montré que pour un stimulus McGurk isolé la fenêtre de décalage temporel [-30, 170] ms fournit un plateau de perception stable, correspondant à une fenêtre temporelle pour l'intégration audiovisuelle. Avec ce « contexte temporellement incohérent », nous cherchons à déterminer si le liage résiste à des fluctuations aléatoires en dépit de la robustesse de l'effet McGurk à des asynchronies temporelles à l'intérieur de cette fenêtre pour un stimulus isolé.
- Incohérent phonético-temporel (PT) qui combine les incohérences temporelles et phonétiques.

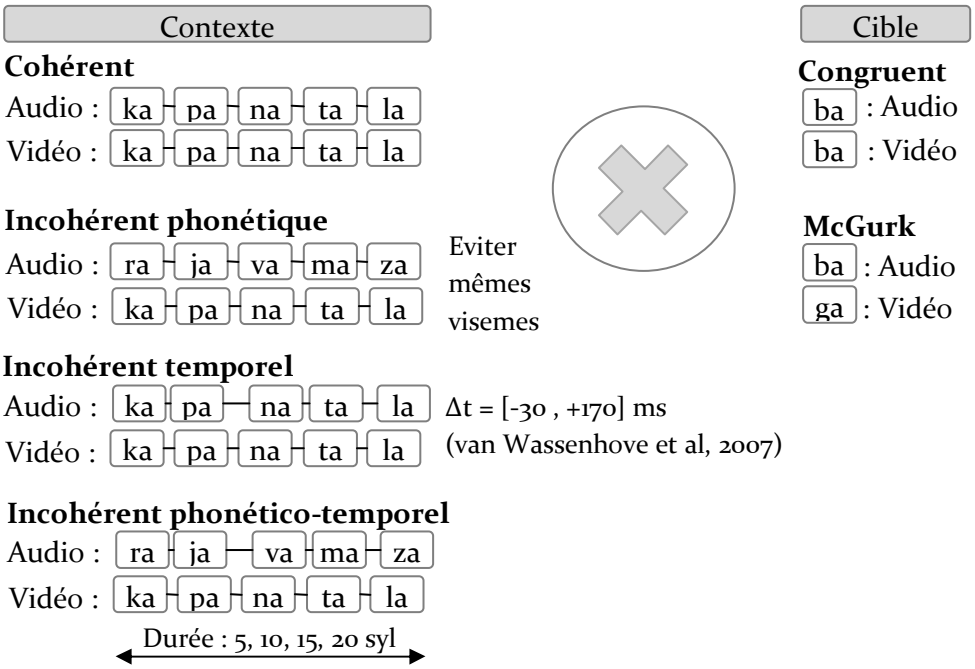


Figure 79 - Principe expérimental

Cette étude a été effectuée avant les Expériences 3 et 4, et ainsi elle contient certaines imperfections mentionnées et corrigées dans l'Expérience 4. Ainsi, dans cette expérience, nous utilisons des stimuli en continuité visuelle, comme dans l'Expérience 1. Contrairement à l'Expérience 1 cependant, les cibles sont exactement les mêmes pour les 4 contextes, et donc les résultats sur les effets du contexte sont fiables, avec l'intérêt d'avoir ici une expérience sans aucune discontinuité entre contexte et cible, car sans aucun montage vidéo. Cependant, les cibles sont par contre différentes d'une durée de contexte à une autre. Ainsi, à la lumière de résultats de l'Expérience 3, nous avons dû constater que le contrôle de l'effet durée n'est pas suffisant, et nous ne le discuterons donc pas dans la partie résultats (§ 10.3).

10.2.2 Stimuli

10.2.2.1 Choix des cibles

Pour préparer nos stimuli nous avons utilisé le même corpus, les détails d'enregistrement sont décrits au paragraphe 5.3.1. Dans ce corpus nous n'avons choisi que 16 séquences avec contexte cohérent et cible McGurk (celles qui étaient testées précédemment dans l'Expérience 2) et 8 séquences avec contexte cohérent et cibles « Ba ». Ainsi, nous n'utilisons que des cibles enregistrées dans le contexte cohérent, ce qui permet d'échapper au problème de différence perceptive entre les deux groupes de cibles, mis en évidence dans l'Expérience 3.

10.2.2.2 Marquage des syllabes

Nous avons marqué toutes les syllabes de ces séquences avec 3 marques respectivement en début de prévoisement, burst, fin de syllabe (Figure 80).

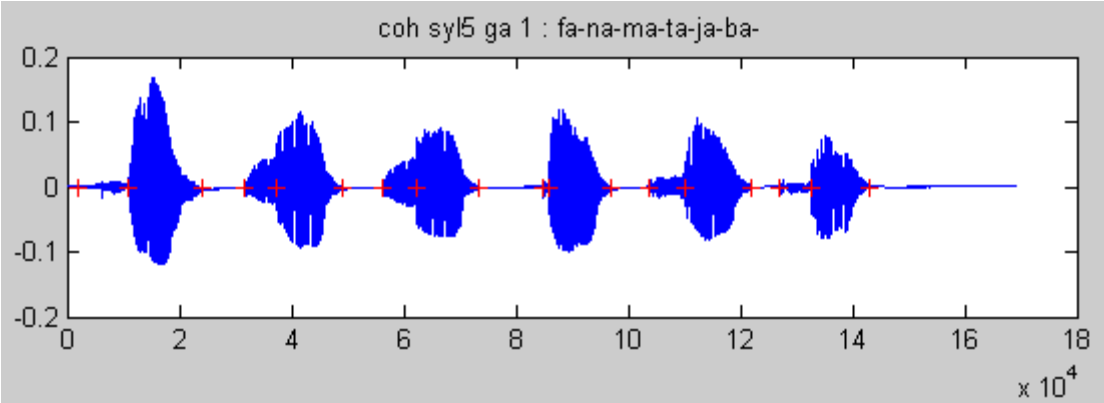


Figure 80 - Marquage des pistes audio avec 3 marques : début de prévoisement, burst, fin de syllable

10.2.2.3 Traitement

Pour augmenter la marge de manœuvre sur les décalages temporels nous avons systématiquement coupé les dernières 50 ms de chaque syllabe du contexte (mais pas de la cible). Pour rendre cette coupure inaudible, nous avons appliqué une fenêtre d'enveloppe de type « Hamming » (sous Matlab) sur les dernières 15 ms. Cette coupure, qui n'a produit aucune incohérence audiovisuelle, permettait d'avoir des plages temporelles plus larges pour appliquer des délais.

Puis nous avons procédé aux permutations et décalages temporels selon les règles décrites dans le paragraphe 10.2. Le point de repère pour appliquer permutations ou décalages était toujours le début du burst.

10.2.3 Plan d'expérience

Au total nous avons préparé 16 stimuli originaux avec cible « McGurk » à la fin et 10 stimuli avec cible « Ba » pour chaque type de contexte. Nous avons dont 4 variations pour chaque durée d'un contexte de 5, 10, 15 et 20 syllabes pour les cibles McGurk et 168 stimuli au total.

Nous avons conservé des proportions proches de ¾ de cibles « McGurk » et ¼ de cibles « Ba », avec au total 168 stimuli selon la répartition indiquée dans le Tableau 22.

Tableau 22 - Nombre de stimuli présentés dans l'expérience

Cohérent		**Incohérent phonétique**		**Incohérent temporel**		**Incohérent phonético-temporel**	
« Ba »	**« McGurk »**	**« Ba »**	**« McGurk »**	**« Ba »**	**« McGurk »**	**« Ba »**	**« McGurk »**
10 (1/16)	32 (3/16)	10 (1/16)	32 (3/16)	10 (1/16)	32 (3/16)	10 (1/16)	32 (3/16)

L'expérience était décomposée en 4 blocs avec des pauses entre les blocs pour ne pas fatiguer les sujets. L'ordre du passage des blocs était varié selon les sujets. Chaque bloc consistait donc en 42 stimuli, choisis de façon aléatoire, parmi tous les stimuli. Les conditions

de passation d'expérience, la présentation des stimuli, les consignes, les types de réponses (2 types de réponse possibles « ba » ou « da » en ligne, choix forcé) étaient les mêmes que dans l'Expérience 1 (Paragraphe 5.4.2). Pour la dernière fois de cette thèse, l'expérience était effectuée avec un logiciel écrit sous JAVA (§5.4.2) ne permettant pas de mesurer des temps de réponse avec une précision adéquate. Les autres conditions de passation d'expérience, les consignes, les types de réponses étaient les mêmes que dans les expériences précédentes (§5.4.2).

10.2.4 Sujets

20 sujets français ont participé à l'expérience, dotés d'une vision et audition normale ou corrigée (5 femmes et 15 hommes, entre 20 et 28 ans avec 22,3 ans en moyenne, 19 droitiers et 1gaucher). Tous les sujets ont donné un consentement éclairé à participer à l'expérience et n'étaient pas au courant du but de l'étude.

10.3 Résultats

10.3.1 Scores bruts

Nous présentons les réponses des sujets dans la matrice de confusion (Tableau 23). Le taux d'erreur (réponses absentes) est **5,36** %. La proportion d'erreur est similaire à l'expérience précédente. La distribution des erreurs est également comparable (l'absence de réponse est plus rare que le cas de réponses multiples). Les taux d'erreurs sont identiques à travers les différentes conditions. Les taux de réponses multiples pour les cibles « Ba » et « McGurk »sont similaires, et l'absence de réponse plus rare pour les cibles « Ba ».

La baisse du taux d'erreurs par rapport aux Expériences 1 et 2 (12,5% dans l'Expérience 1 et 10,45% dans l'Expérience 2) est sans doute due ici, comme dans l'Expérience 4, à l'absence des cibles enregistrées en contexte incohérent dans la présente expérience. Ceci diminue la variété physique de l'ensemble des stimuli cibles présentés, en deux sous-groupes : cibles « Ba » et « McGurk » cohérentes. L'absence des cibles « McGurk » incohérentes, plus dispersées en termes des paramètres physiques (Figure 48), conduit probablement à une meilleure détection des cibles.

Tableau 23 – Matrice de confusion

Stimuli		**Stimuli présentés**	**Réponse « ba »**		**Réponse « da »**		**Plusieurs réponses « ba »**		**Plusieurs réponses « da »**		**Absence de réponses**		**Absence plusieurs réponses**	
Coh	**Ba**	200	182	(91%)	2	(1%)	7	(4%)	0	(0%)	0	(0%)	9	(5%)
	McG	640	395	(62%)	174	(27%)	27	(4%)	13	(2%)	8	(1%)	23	(4%)
P	**Ba**	200	180	(90%)	4	(2%)	10	(5%)	1	(1%)	0	(0%)	5	(3%)
	McG	640	429	(67%)	111	(17%)	53	(8%)	4	(1%)	18	(3%)	25	(4%)
T	**Ba**	200	183	(92%)	5	(3%)	7	(4%)	0	(0%)	1	(1%)	4	(2%)
	McG	640	397	(62%)	162	(25%)	27	(4%)	4	(1%)	18	(3%)	32	(5%)
PT	**Ba**	200	182	(91%)	2	(1%)	7	(4%)	0	(0%)	3	(2%)	6	(3%)
	McG	640	490	(77%)	89	(14%)	28	(4%)	5	(1%)	8	(1%)	20	(3%)

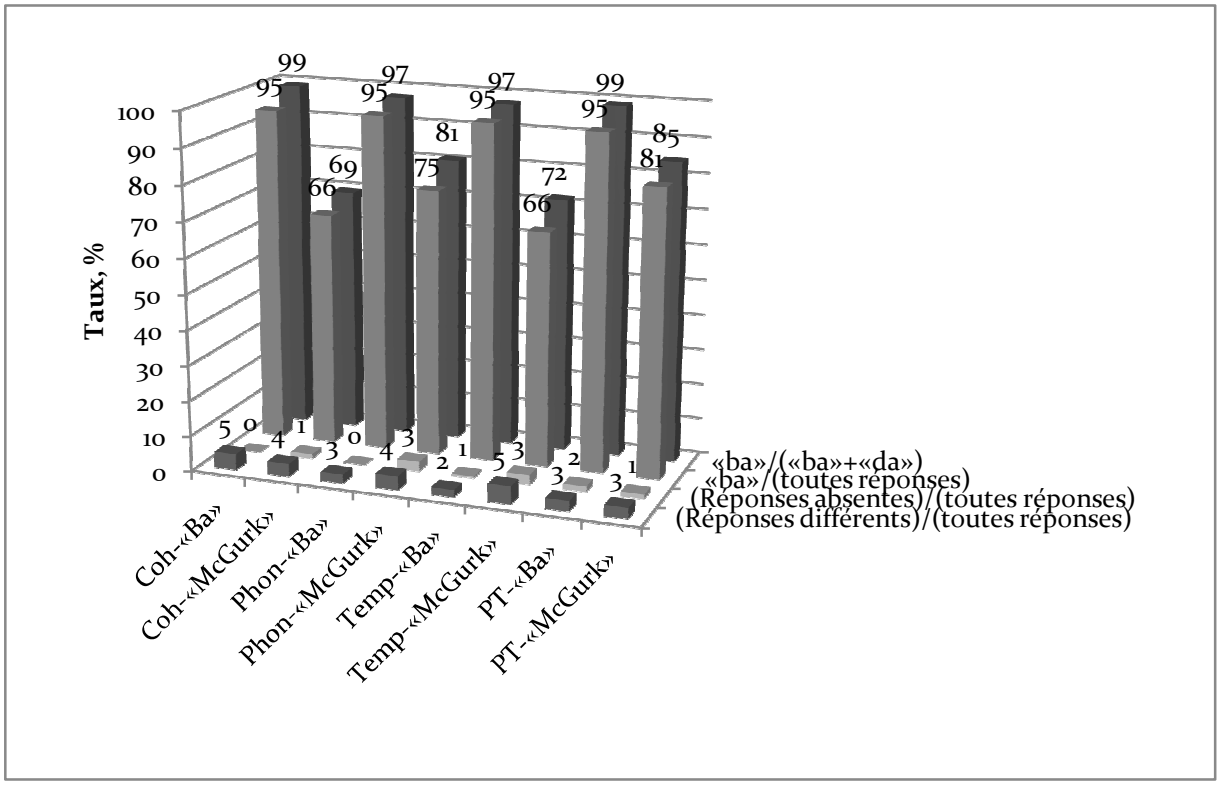


Figure 81 – Expérience 5, données brutes, %.

L'analyse des réponses des participants est présentée sur la Figure 82 en triant les participants selon leur score de réponses aux cibles en contexte cohérent. Il apparaît dans cette expérience que le nombre de participants avec un taux d'effet McGurk faible est plus élevé (presque la moitié). Comme dans les expériences précédentes, ces taux varient beaucoup selon les sujets. Il apparaît un effet systématique des contextes incohérents phonétique et phonetico-temporel qui produisent une augmentation du nombre de réponses « ba » (donc une baisse d'effet McGurk) pour presque tous les sujets. La différence entre contexte cohérent et contexte incohérent temporel est moins claire.

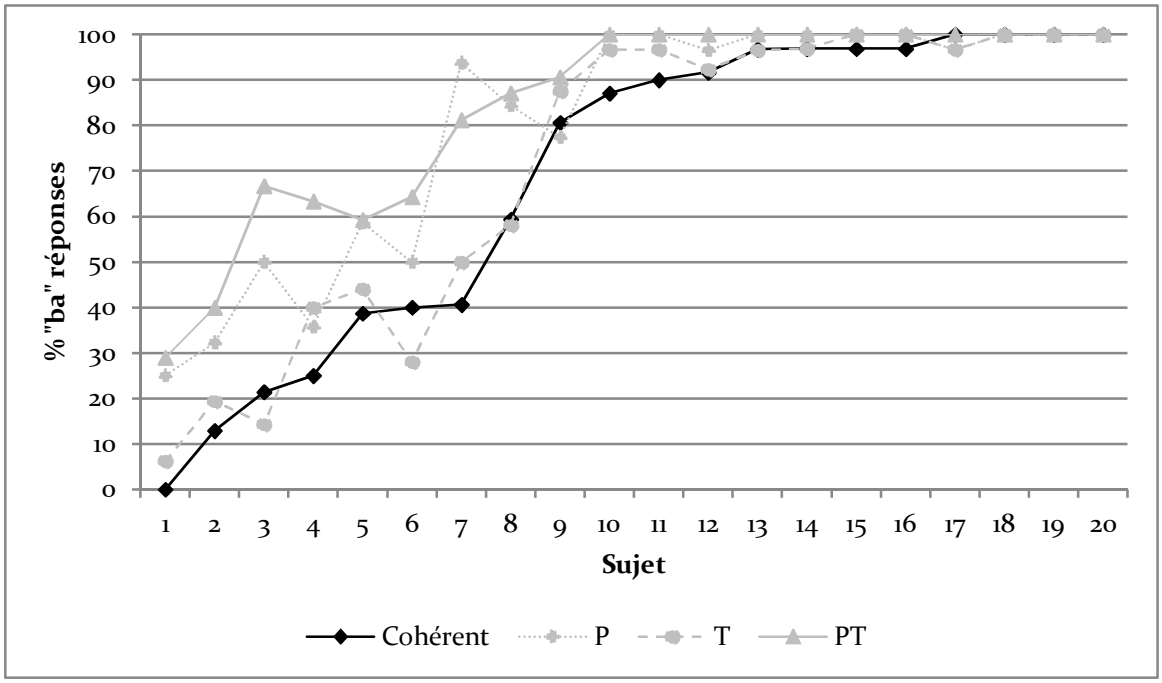


Figure 82 - Perception d'effet McGurk par sujet (« ba »/(« ba » + « da »)).

10.3.2 Analyses statistiques des pourcentages de réponse

10.3.2.1 Effet de la cible et du contexte

Regardons d'abord les pourcentages globaux de réponse aux cibles « Ba » et « McGurk » en fonction du contexte (Figure 83). Les cibles « Ba » sont perçues correctement dans tous les contextes, avec 100% de réponses « ba ». Les taux d'effet McGurk sont globalement plus faibles par rapport aux expériences précédentes. Ainsi, dans le cas le plus favorable avec le contexte cohérent, nous n'obtenons que 25% de réponses « da ». Comme nous l'avons vu, près de la moitié des sujets présentent un effet McGurk faible (Figure 82), contre environ un quart dans les expériences précédentes. Le contexte incohérent temporel (T) produit une baisse d'effet McGurk jusqu'à 21%, le contexte incohérent phonétique (P) jusqu'à 11% et le contexte phonético-temporel (PT) jusqu'à 8%.

Une ANOVA à mesures répétées sur les facteurs contexte et cible (Tableau 24) montre un effet significatif des deux facteurs (contexte [$F(3,57)=10.69$, $P<0.001$], cible [$F(1,19)=16.86$, $P<0.005$]), ainsi que de leur interaction [$F(3,57)=11.13$, $P<0.001$]. Une analyse post hoc montre que l'augmentation du taux de réponses « ba » pour les cibles « McGurk » est significative entre les contextes C et T d'une part, et P et PT d'autre part ($P<0.05$). La réduction d'effet McGurk d'environ 50% produite par les contextes phonétique et phonético-temporel est de même ordre de grandeur que celle produite par le contexte incohérent de l'Expérience 4.

L'effet sujet est une fois encore significatif [$F(1,19)=701.48$, $P<0.001$] (Figure 82).

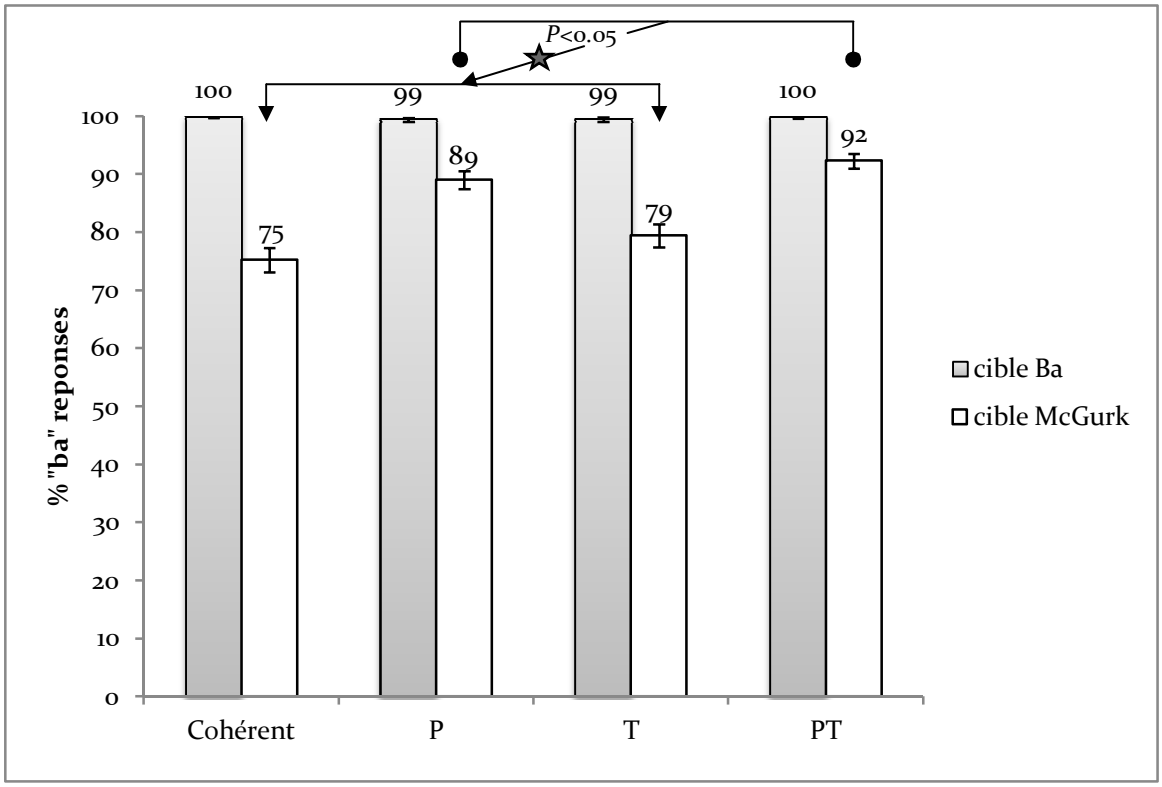


Figure 83 - Taux des réponses (« ba »/(« ba » + « da ») dans l'Expérience 5

Tableau 24– ANOVA à mesures répétées: cible, contexte.

Source		Somme des carrés de type III	ddl	Moyenne des carrés	D	Sig.
Contexte : coh, incoh	Sphéricité supposée	,392	3	,131	10,685	,000
Erreur (contexte)	Sphéricité supposée	,697	57	,012		
Cible : Ba, McGurk	Sphéricité supposée	4,899	1	4,899	16,855	,001
Erreur (cible)	Sphéricité supposée	5,523	19	,291		
*Contexte * cible*	Sphéricité supposée	,394	3	,131	11,128	,000
*Erreur (contexte*cible)*	Sphéricité supposée	,673	57	,012		

10.3.2.2 *Incohérence temporelle vs. incohérence phonétique*

L'analyse globale présentée ci-dessus ne fait pas apparaître d'effet significatif de l'incohérence temporelle. Néanmoins, on observe une différence d'environ 3% entre les contextes T et C d'une part, P et PT d'autre part. Nous avons donc choisi dans un second temps d'effectuer une analyse sur les seules cibles « McGurk » en décomposant les conditions contextuelles (C, P, T, PT) selon deux facteurs P et T[3]. Cette analyse montre que les deux facteurs produisent un effet significatif (T [$F(1,19)=12.83$, $P<0.005$], P [$F(1,19)=31.01$, $P<0.001$]), mais pas leur interaction [$F(1,19)=0.02$, $P=0.88$] (Tableau 25).

Cette analyse montre ainsi une influence, faible mais significative, d'une incohérence temporelle sur le processus de déliage, bien que les faible fluctuations imposées restent chacune compatibles avec la fenêtre d'intégration phonétique audiovisuelle (van Wassenhove et al, 2007).

Tableau 25– ANOVA à mesures répétées selon les facteurs T, P pour les cibles « McGurk »

Source		Somme des carrés de type III	ddl	Moyenne des carrés	D	Sig.
T	Sphéricité supposée	,057	1	,057	12,832	,002
Erreur (T)	Sphéricité supposée	,084	19	,004		
P	Sphéricité supposée	,699	1	,699	31,011	,000
Erreur (P)	Sphéricité supposée	,428	19	,023		
*T * P*	Sphéricité supposée	,000	1	,000	,022	,883
*Erreur (T*P)*	Sphéricité supposée	,145	19	,008		

10.4 Discussion

Dans cette expérience nous avons décomposé l'incohérence maximale imposée dans les expériences précédentes selon deux dimensions, phonétique et temporelle, qui produisent chacune un effet significatif et semblent ainsi participer l'une et l'autre au processus de déliage.

Un premier point notable est que l'incohérence phonétique produit un effet de déliage qui semble globalement comparable avec celui d'une incohérence « totale », selon les résultats de l'Expérience 4. Ce point, un peu surprenant, peut être dû au fait qu'il n'y a pas dans l'Expérience 5 de discontinuité visuelle entre contexte et cible, alors que nous avons dû en

[3] Cette décomposition est effectuée en recodant les 4 contextes C : 1, P : 2, T : 3, PT : 4, sur deux dimensions
T : 0/1, P : 0/1, soit un codage global C : 00, P : 01, T : 10, PT : 11.

imposer une dans l'Expérience 4. Comme nous ne disposons pas d'une condition d'incohérence complète dans cette expérience, impossible à produire pour des raisons de construction même des stimuli, nous ne disposons bien évidemment pas de comparaison directe de ces deux types de contexte, qui nous permettrait de savoir si cette discontinuité joue un rôle, ou si l'incohérence phonétique est effectivement le responsable majeur de l'effet de déliage dans l'Expérience 4.

Le fait en soit d'obtenir un effet de déliage fort avec une incohérence phonétique imposée sur une cohérence temporelle aussi parfaite que possible est aussi une relative surprise. Nous attendions que les mécanismes de corrélation audiovisuelle soient au cœur du processus de liage, il apparaît que ces mécanismes de corrélation ne portent pas sur les seules fluctuations d'intensité globale, mais comprennent également des composantes phonétiques fines (donc impliquent effectivement des analyses spectrales détaillées, et sans doute des bandes de fréquence spécifiques) qui jouent un rôle majeur dans les calculs de cohérence audiovisuelle.

En ce qui concerne les effets d'incohérence temporelle, il apparaît que l'impact d'une incohérence temporelle est faible mais significatif, malgré la petitesse des fluctuations temporelles imposées, qui toutes étaient comprises dans la fenêtre d'intégration audiovisuelle supposée. Bien évidemment, on peut s'attendre à ce que des variations temporelles plus amples, en dehors de la fenêtre d'intégration audiovisuelle, auraient produit un effet beaucoup plus fort. Ce point, non étudié dans le présent travail, fera partie de nos perspectives.

Les expériences qui suivent vont s'attacher maintenant à mieux comprendre la dynamique temporelle des effets de liage/déliage.

Chapitre 11. Expérience 6. Caractérisation de la dynamique temporelle.

11.1 Objectifs et hypothèses

Dans l'Expérience 4, la seule expérience où nous avons pu faire une analyse du rôle de la durée du contexte dans le processus de déliage, nous n'avons pas observé de différence significative d'effet McGurk en contexte incohérent entre des durées de contexte de 5 et 20 syllabes. Comme on l'observe sur la Figure 77, pour la durée du contexte la plus courte de 5 syllabes, équivalent à 3 secondes environ, nous observons déjà un déliage maximal. L'objectif de la nouvelle expérience présentée dans ce chapitre est d'étudier le processus du liage sur des durées de contexte plus courtes, sachant que tous les changements possibles peuvent apparaître avant 5 syllabes (~3 s). C'est donc sur cette plage de 0 à 3 s de contexte que nous allons nous focaliser maintenant.

11.2 Méthodologie

11.2.1 Principe

Dans cette expérience nous avons repris le même schéma expérimental général que dans les expériences précédentes, qui consiste à présenter des stimuli comprenant un contexte et une cible. Nous avons repris les contextes de base cohérents et maximalement incohérents de l'Expérience 4 (Chapitre 9), et comme nous avons vu dans l'Expérience 5 au Chapitre 10 que le contexte « phonétique » (avec des incohérences de contenu phonétique sans incohérence temporelle) semble produire également de forts effets de déliage des sources auditives et visuelles nous avons décidé de l'intégrer également dans cette expérience. Nous pourrons ainsi comparer directement l'influence de ces deux contextes sur l'effet McGurk.

Le but de cette expérience est d'observer la dynamique du processus de déliage sur des durées courtes, nous avons donc choisi des durées de contexte entre 1 et 5 syllabes. Nous avons également introduit des cibles sans contexte pour fournir une ligne de base au processus de déliage. Bien évidemment, nous utilisons toujours deux types de cibles : « Ba » et « McGurk ».

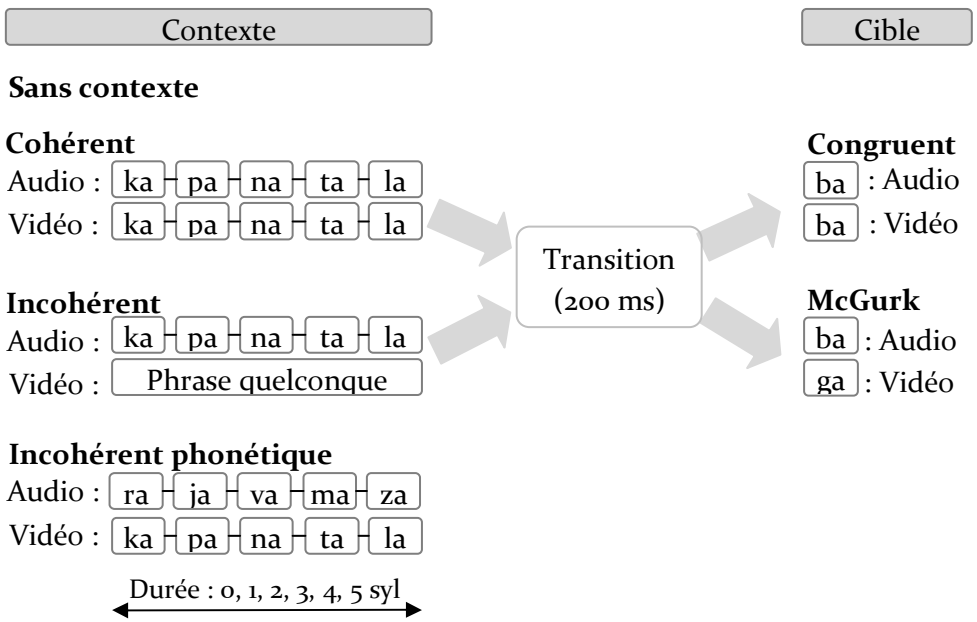


Figure 84 - Principe expérimental

11.2.2 Stimuli

Pour réaliser cette expérience nous avons repris les mêmes séquences que celles de l'Expérience 4 (voir description § 9.2.2) pour les contextes cohérent et incohérent. La création du contexte phonétique est décrite dans l'Expérience 5 (§ 10.2). La seule modification que nous avons appliquée à ces stimuli est une coupure systématique de la partie initiale des contextes pour créer des durées de contexte plus courtes. Pour ce faire nous avons repris les contextes de 5 syllabes tels quels, réduit les contextes de base de 10 syllabes pour créer par coupure initiale les contextes de 4 syllabes, et de même, les séquences de contexte de 15 syllabes ont été réduites à 3 syllabes et les séquences de 20 syllabes réduites à deux syllabes. Pour préparer un contexte d'une seule syllabe nous avons systématiquement pris la 10ème syllabe dans les contextes de 20 syllabes. La coupure de début du contexte est effectuée systématiquement 80 ms avant le début acoustique de la première syllabe du contexte. En ce qui concerne les séquences contextuelles à une seule syllabe, elles se terminent, comme tous les autres contextes, 240 ms avant le burst de la syllabe suivante (11ème syllabe). Pour mémoire, nous avons démarré systématiquement toutes les cibles 240 ms avant le burst. En procédant de cette manière nous avons d'une part assuré une variété de contextes et d'autre part nous avons assuré une cohérence maximale de cette expérience par rapport aux autres.

11.2.3 Plan d'expérience

Nous avons maintenu les proportions ¾ de cibles McGurk et ¼ de cibles Ba. Nous avons ainsi préparé 12 cibles « McGurk » et 4 cibles « Ba » qui étaient associées à 3 types de contextes (cohérent, incohérent, incohérent phonétique), de 5 durées possibles (1, 2, 3, 4 et 5 syllabes) avec 4 variantes pour chaque durée. La répartition des cibles permet d'assurer que l'ensemble des 12 cibles McGurk sont associées à chaque contexte et chaque durée de contexte, avec une distribution aléatoire des cibles entre les 4 variantes possibles de chaque durée de contexte. Ceci reprend globalement les principes de l'Expérience 4, décrite en détail dans le paragraphe 9.2.2.4. Nous avons également introduit un jeu de cibles isolées (12 cibles « McGurk » et 4 cibles « Ba »).

Au total nous avons donc présenté 256 stimuli (3*5*(12+4) + (12+4)). La répartition des stimuli de chaque type est indiquée dans le Tableau 26.

Tableau 26 - Nombre de stimuli présentés dans l'Expérience 6

Sans contexte		**Cohérent**		**Incohérent**		**Incohérent phonétique**	
« Ba »	**« McGurk »**	**« Ba »**	**« McGurk »**	**« Ba »**	**« McGurk »**	**« Ba »**	**« McGurk »**
4 (1/4)	12 (3/4)	20 (1/4)	60 (3/4)	20 (1/4)	60 (3/4)	20 (1/4)	60 (3/4)

Tous les stimuli ont étaient mélangés aléatoirement et présentés dans un bloc continu de durée environ de 15 minutes. Nous avons préparé 5 blocs différents, dont chacun était présentés aux 4 sujets en contrôlant entre les sujets la latéralisation des réponses.

En résumé, l'expérience consiste en :

- 256 stimuli au total répartis aléatoirement dans un bloc entier
- 5 blocs possibles (pour 5 ordres différents des stimuli), contrebalancés entre les sujets

- Facteur durée : 1, 2, 3, 4, 5 syllabes CV et une condition sans contexte
- 2 types de réponse possible « ba » ou « da » en ligne, choix forcé

Nous avons utilisé ici le logiciel Presentation® (Version 0.70, www.neurobs.com).Les autres conditions de passation d'expérience, les consignes, les types de réponses étaient les mêmes que dans l'Expérience 1 (Paragraphe 5.4.2).

11.2.4 Sujets

20 sujets français ont participé à l'expérience avec une vision et audition normale ou corrigée (4 femmes et 16 hommes, entre 23 et 54 ans avec 26,6 ans en moyenne, 19 droitiers et 1 gaucher). Tous les sujets ont donné leur consentement éclairé à participer à l'expérience et n'étaient pas au courant du but de l'étude.

11.3 Résultats

11.3.1 Scores bruts

Nous avons appliqué la même méthode de traitement des réponses et d'analyse statistique que dans toutes les expériences précédentes (Paragraphe 5.5). Néanmoins, étant donné que nous avons des durées de contexte plus courtes et même dans certains cas une absence de contexte, nous avons réduit la période d'acceptation des réponses de 3500 ms à 1200 ms. Nous avons vérifié que cette fenêtre réduite permettait néanmoins de valider la grande majorité des réponses.

Nous présentons le détail des réponses des sujets dans la matrice de confusion (Tableau 27). Le taux d'erreurs (réponses absentes) est de 5,8%. Le taux d'erreurs moyen est comparable à celui obtenu dans les Expériences 4 et 5. Même dans la condition sans contexte le nombre d'erreurs est comparable, ainsi le fait de raccourcir le contexte n'engendre pas un taux d'erreur plus grand.

Tableau 27 – Matrice de confusion (C pour « contexte cohérent », I pour « contexte incohérent », P pour « contexte phonétiquement incohérent »

Stimuli		Stimuli présentés	Réponse « ba »		Réponse « da »		Plusieurs réponses « ba »		Plusieurs réponses « da »		Absence de réponses		Absence plusieurs réponses	
Sans contexte	Ba	80	76	(95%)	0	(0%)	1	(1%)	0	(0%)	0	(0%)	3	(4%)
	McG	240	116	(48%)	102	(43%)	6	(3%)	5	(2%)	2	(1%)	9	(4%)
C	Ba	400	362	(91%)	6	(2%)	6	(2%)	0	(0%)	6	(2%)	20	(5%)
	McG	1200	511	(43%)	576	(48%)	19	(2%)	24	(2%)	22	(2%)	48	(4%)
I	Ba	400	373	(93%)	7	(2%)	4	(1%)	0	(0%)	2	(1%)	14	(4%)
	McG	1200	766	(64%)	333	(28%)	14	(1%)	15	(1%)	15	(1%)	57	(5%)
P	Ba	400	369	(92%)	5	(1%)	9	(2%)	0	(0%)	1	(0%)	16	(4%)
	McG	1200	585	(49%)	490	(41%)	26	(2%)	19	(2%)	24	(2%)	56	(5%)

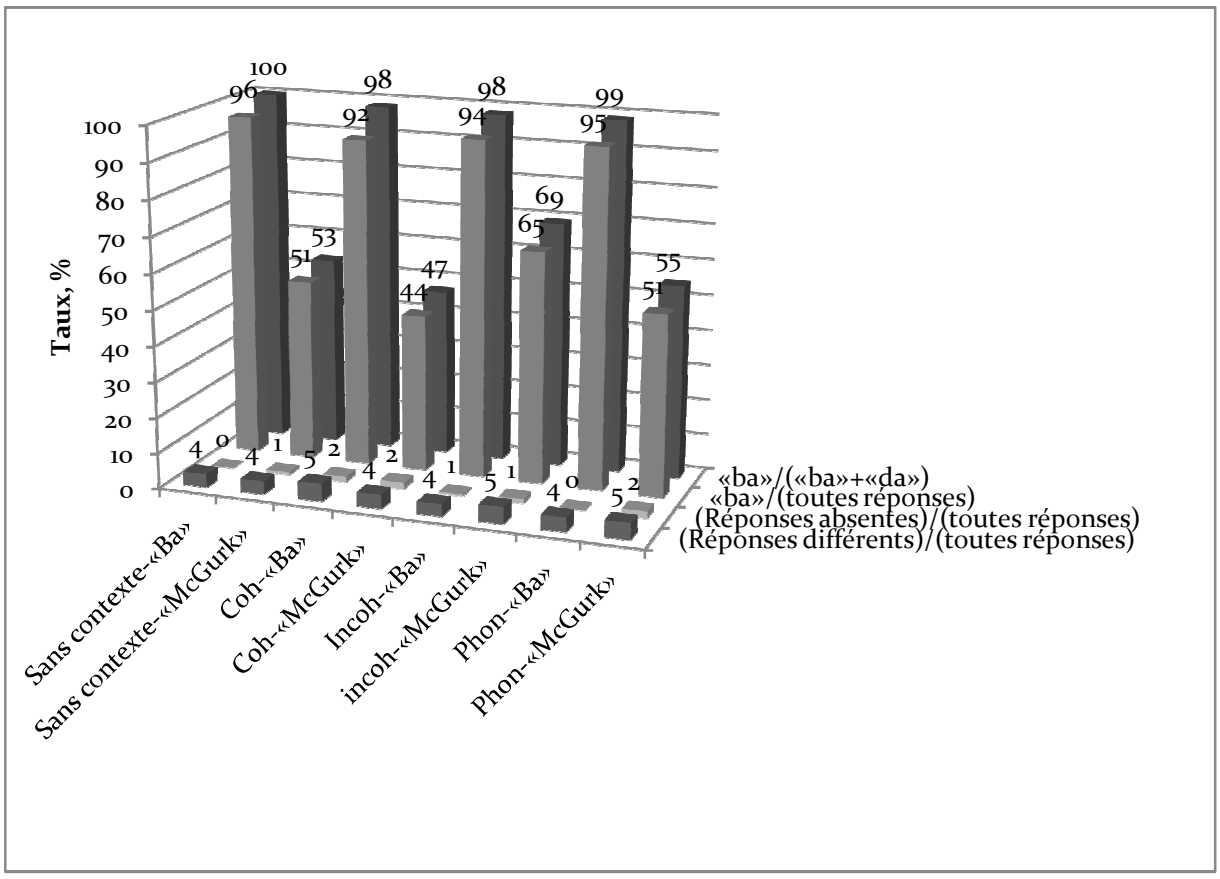


Figure 85 – Expérience 6, données brutes, %.

L'analyse du comportement des participants pour les cibles « McGurk » est présentée sur la Figure 86 en ordonnant les sujets en fonction de leurs réponses en contexte cohérent. On retrouve de fortes variations de l'effet McGurk selon les sujets. Nous observons la différence entre les contextes cohérent et incohérent, pour tous les sujets, dont les courbes ne se croisent jamais. Le contexte phonétique semble fournir plus de réponses « ba » pour la plupart des sujets, mais cette différence est moins forte que pour le contexte incohérent. La perception d'une cible seule sans contexte est plus irrégulière selon les sujets.

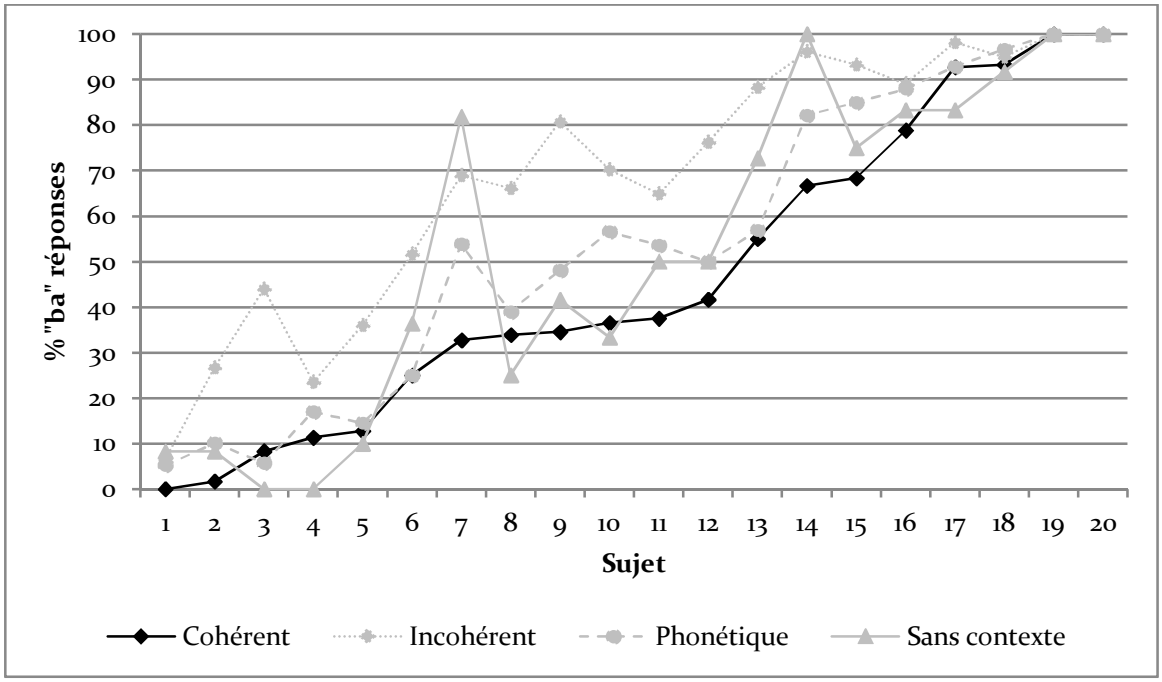


Figure 86 - Perception d'effet McGurk par sujet (« ba »/(« ba » + « da »)).

11.3.2 Analyses statistiques des pourcentages de réponse

11.3.2.1 Effet de la cible et du contexte

Commençons notre analyse en regardant les résultats globaux des cibles « Ba » et « McGurk » en fonction du contexte (Figure 87). Sur cette figure nous présentons les conditions « sans contexte » à titre indicatif, afin d'avoir un repère permettant d'alimenter nos réflexions. Cependant cette condition n'est jamais prise en compte dans les analyses de la variance, étant donné que le nombre de stimuli n'est pas équivalent à celui des autres conditions de contexte.

Les cibles « Ba » sont perçues correctement dans tous les contextes, avec 100% de réponses « ba ». L'effet McGurk pour les cibles isolées est à 46% (54% de réponses « ba »), ce qui est plutôt élevé (Cathiard et al, 2001).

L'effet McGurk en contexte cohérent est plus élevé que dans le cas sans contexte, avec 53% de réponses « da ». Ce score est également plus élevé que dans les Expériences 1, 2 et 4, qui présentaient environ 40% de perception d'effet McGurk. En contexte incohérent le taux d'effet McGurk diminue à 27% et en contexte incohérent phonétique à 43%.

Une ANOVA à mesures répétées sur les facteurs contexte et cible (Tableau 28) montre un effet significatif des deux facteurs [contexte : $F(2.43,46.18)=10.25$, $P<0.001$], [cible : $F(1,19)=55.1$, $P<0.001$], et de leur interaction [$F(3,57)=14.67$, $P<0.001$]. Une analyse Post hoc montre que la perception d'effet McGurk est significativement différente entre les trois contextes ($P<0.05$). Cette expérience confirme que le contexte phonétique est suffisant pour délier les flux auditifs et visuels, mais contrairement à nos premières conclusions – provenant d'analyses indirectes – dans l'Expérience 5, son effet n'est pas aussi grand que celui du contexte incohérent.

L'effet sujet est également significatif une nouvelle fois [$F(1,19)=536.27$, $P<0.001$]).

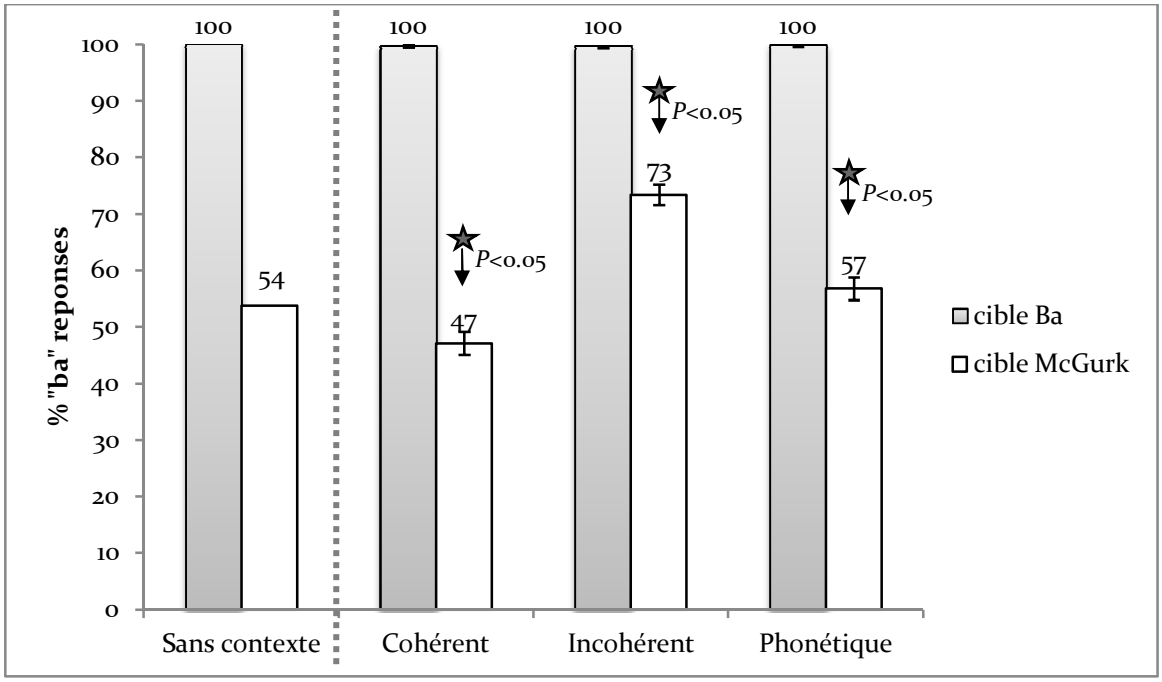


Figure 87 - Taux des réponses (« ba »/(« ba » + « da »)) dans l'Expérience 6.

Tableau 28– ANOVA à mesures répétées : cible, contexte.

Source		Somme des carrés de type III	ddl	Moyenne des carrés	D	Sig.
Contexte : coh, incoh, phon	Huynh-Feldt	,368	2,431	,151	10,251	,000
Erreur (contexte)	Huynh-Feldt	,682	46,181	,015		
Cible : Ba, McGurk	Sphéricité supposée	17,707	1	17,707	55,106	,000
Erreur (cible)	Sphéricité supposée	6,105	19	,321		
*Contexte * cible*	Sphéricité supposée	,489	3	,163	14,670	,000
*Erreur (contexte*cible)*	Sphéricité supposée	,634	57	,011		

11.3.2.2 Effet de la durée du contexte

Passons à la question principale de cette expérience, l'analyse de la durée du contexte. Une ANOVA à mesures répétées sur les facteurs durée et contexte centrée sur les cibles McGurk montre un effet significatif des deux facteurs (durée [F(4,76)=7.21, P<0.001], contexte [F(2,38)=46.7, *P*<0.001]), ainsi que de leur interaction [*F*(8,152)=4.72, *P*<0.001] (Tableau 29). Une analyse posthoc (p<0.05) centrée sur l'interaction contexte*durée montre que ce n'est que dans le contexte incohérent que nous observons un effet de la durée, avec une différence significative de réponse entre les durées 1-2 syl et 4 syl. Nous reviendrons sur les interprétations possibles de cette observation au vu des analyses qui suivront.

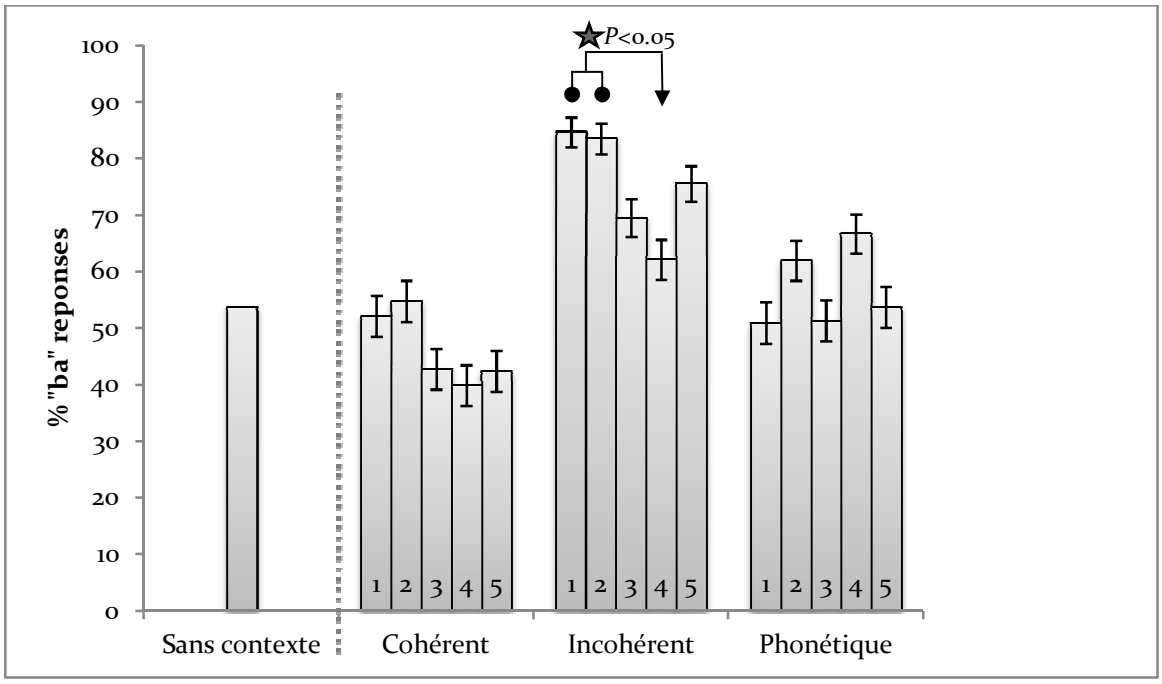


Figure 88 - Taux des réponses (« ba »/ « ba » + « da ») en fonction du contexte et de la durée

Tableau 29– ANOVA à mesures répétées: effet du contexte et de la durée.

Source		Somme des carrés de type III	ddl	Moyenne des carrés	D	Sig.
Contexte : coh, incoh, phon	Sphéricité supposée	4,809	2	2,405	46,693	,000
Erreur (contexte)	Sphéricité supposée	1,957	38	,051		
Durée : 1, 2, 3, 4, 5 syls	Sphéricité supposée	,744	4	,186	7,206	,000
Erreur (durée)	Sphéricité supposée	1,962	76	,026		
*Contexte * durée*	Sphéricité supposée	1,012	8	,127	4,716	,000
*Erreur (contexte*durée)*	Sphéricité supposée	4,078	152	,027		

Revenons à l'effet de la durée tous contextes confondus (Figure 89). Il apparaît un taux de réponses « ba » plus élevé pour des durées 1 et 2 syllabes par rapport aux durées 3, 4, 5 syllabes. Une analyse posthoc montre un effet significatif (p<0.5) entre les durées 1 et 3 syllabes ainsi qu'entre les durées 2 vs. 3 et 4 syllabes. Nous reviendrons plus loin sur ce résultat un peu surprenant.

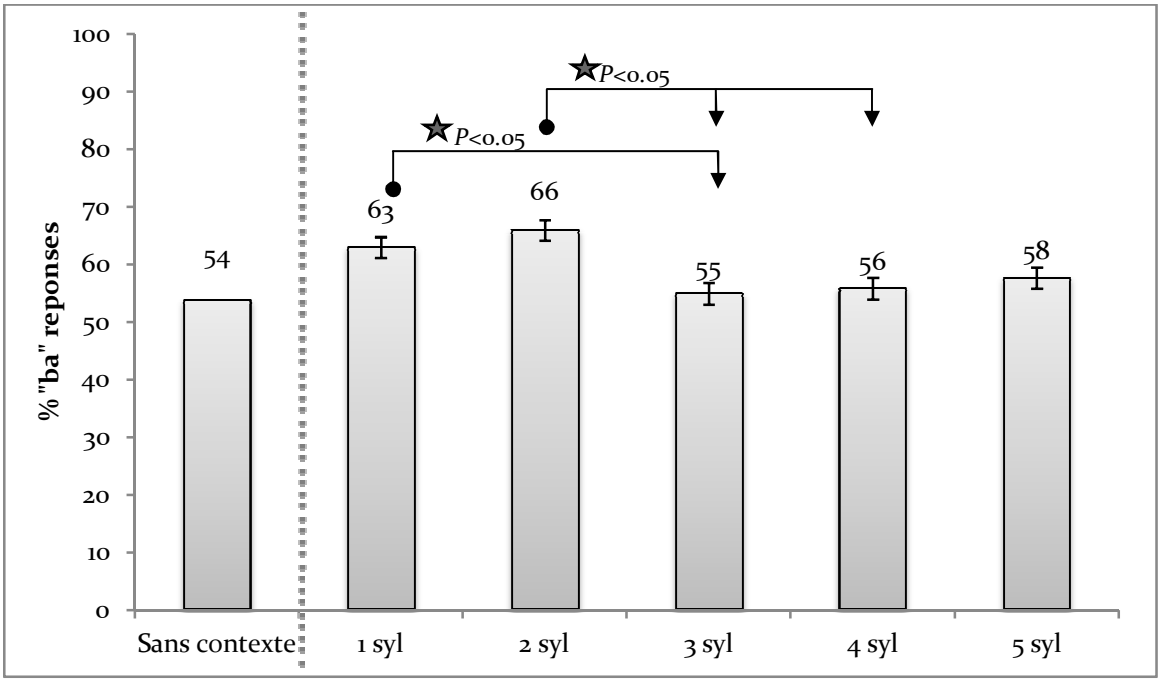


Figure 89 - Taux des réponses (« ba »/ « ba » + « da ») en fonction de la durée du contexte, tous contextes confondus

Notons enfin un résultat important : le déliage semble très rapide, puisque la baisse de l'effet McGurk dans les deux contextes incohérents est aussi forte avec une seule syllabe qu'avec plusieurs, voire même plus forte en contexte incohérent non phonétique.

11.3.3 Temps de réponses

Passons maintenant à l'analyse des temps de réponse. Sur la Figure 90 nous présentons le temps de réponse en fonction de la cible et du contexte. La condition sans contexte est présentée à titre indicatif et n'entre pas dans les analyses. Globalement, les cibles « Ba » conduisent à un temps de traitement autour de 600 ms dans tous les contextes et les cibles McGurk demandent un temps de l'ordre de 660-690 ms. Ainsi, les cibles « McGurk » sont traitées environ 70 ms plus lentement que les cibles « Ba » et ce pour tous les contextes. Une ANOVA à mesures répétées (Tableau 30) montre un effet significatif pour le facteur cible [$F(1,19)=37.86$, $P<0.001$] et non significatif pour le facteur contexte [$F(2,38)=0.72$, $P=0.5$], ainsi que pour leur interaction [$F(2,38)=2.5$, $P=0.1$]. Ainsi l'ANOVA confirme que les cibles « McGurk » demandent plus de temps de traitement que les cibles « Ba ».

Nous avions déjà observé un pattern similaire dans l'Expérience 4 (quoique plus faible, et n'apparaissant, rappelons-le, que dans une analyse avec le facteur sujet comme effet fixe). Cette fois nous obtenons des différences de temps de réponses plus grandes (autour de 70 ms ici contre 25 ms dans l'Expérience 4) avec des temps moyens globalement plus élevés (640 ms ici contre 565 ms dans l'Expérience 4).Cette augmentation du temps de réponse pourrait être due à une durée de contexte plus courte dans cette expérience. Avec des durées de contexte plus longues, comme c'est le cas dans l'Expérience 4, le sujet peut anticiper l'arrivée d'une syllabe cible par la rythmicité du contexte, tandis que dans le cas de durées plus courtes l'accrochage sur le rythme est plus difficile.

Le fait que la différence de temps de réponses entre cibles soit maintenue du contexte cohérent aux contextes incohérents est particulièrement intéressant dans le contexte d'une critique possible de nos données qui serait que les sujets prêtent de moins en moins d'attention aux stimuli visuels en contexte incohérent. Si c'était le cas, les temps de réponses baisseraient dans le cas de cibles McGurk, ce qui n'est pas le cas. Ainsi, le maintien d'une différence significative de temps de réponse entre les deux cibles quel que soit le contexte confirme que les sujets regardent bien le locuteur dans tous les contextes, et que c'est bien dans le processus de traitement (liage) et de décision qu'il faut chercher l'explication de la baisse d'effet McGurk en contexte incohérent.

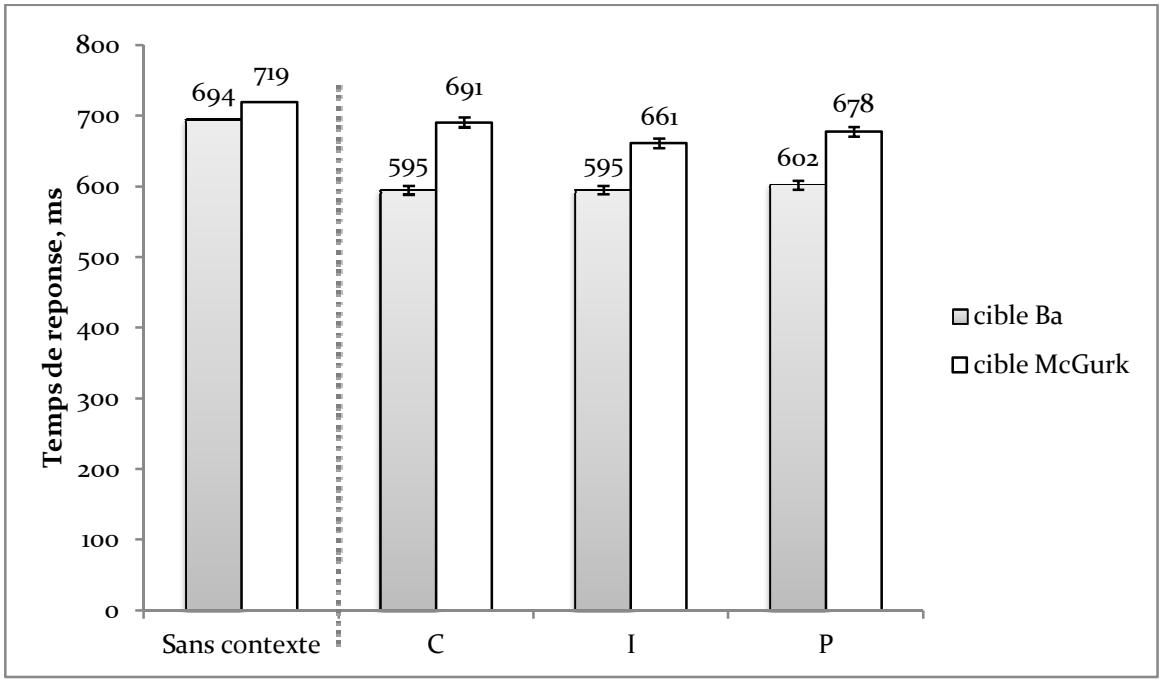


Figure 90 - Temps de réponse, ms

Tableau 30 - ANOVA à mesures répétées des temps des réponses selon les facteurs cible, contexte et sujet

Source		Somme des carrés de type III	ddl	Moyenne des carrés	D	Sig.
Contexte : coh, incoh, phon	Sphéricité supposée	,011	2	,005	,715	,496
Erreur (contexte)	Sphéricité supposée	,286	38	,008		
Cible : Ba, McGurk	Sphéricité supposée	,465	1	,465	37,861	,000
Erreur (cible)	Sphéricité supposée	,234	19	,012		
*Contexte * cible*	Sphéricité supposée	,010	2	,005	2,504	,095
*Erreur (contexte*cible)*	Sphéricité supposée	,079	38	,002		

Revenons à l'observation que nous avions faite d'une augmentation du nombre de réponses « ba » pour des durées de contexte de 1 et 2 syllabes, tous contextes confondus. Sur la Figure 91 nous présentons les variations de temps de réponse en fonction des facteurs durée et contexte pour des cibles McGurk. Une ANOVA à mesures répétées centré pour ces cibles McGurk (Tableau 31) montre un effet significatif du facteur durée [$F(4,76)=6.68$, $P<0.001$], mais pas d'effet significatif du facteur contexte [$F(2,38)=2.54$, $P=0.92$] ni de leur interaction [$F(5.37,102.1)=0.97$, $P=0.44$].

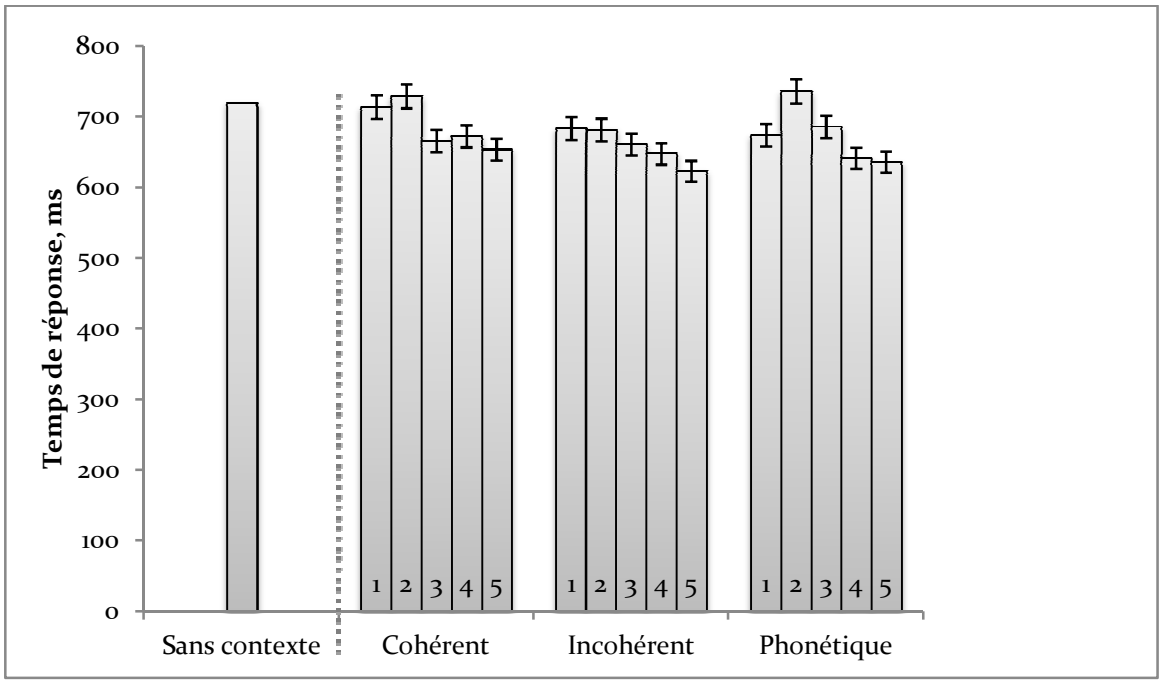


Figure 91– Temps de réponse (ms) pour les cibles « McGurk » en fonction du contexte et de la durée

Tableau 31– ANOVA à mesures répétées sur les temps de réponses pour les cibles « McGurk », en fonction du contexte et de la durée.

Source		Somme des carrés de type III	ddl	Moyenne des carrés	D	Sig.
Contexte : coh, incoh, phon	Sphéricité supposée	,084	2	,042	2,537	,092
Erreur (contexte)	Sphéricité supposée	,631	38	,017		
Durée : 1, 2, 3, 4, 5 syls	Sphéricité supposée	,487	4	,122	6,677	,000
Erreur (durée)	Sphéricité supposée	1,385	76	,018		
*Contexte * durée*	Huynh-Feldt	,086	5,373	,016	,970	,443
*Erreur (contexte*durée)*	Huynh-Feldt	1,679	102,095	,016		

Sur la Figure 92 on observe une tendance à une augmentation de temps de réponse pour les durées courtes. Un test posthoc confirme l'existence d'une différence significative de temps de réponse entre les durées 1 syllabe vs. 5 syllabes et 2 syllabes vs. 4 et 5 syllabes. On pourrait alors supposer qu'il existe un lien entre l'augmentation du temps de réponse et la diminution de perception d'effet McGurk pour les durées courtes, d'une et deux syllabes (Figure 89, Figure 92). Ce lien pourrait provenir d'une baisse attentionnelle des sujets pour les contextes de durée courte, liée à la vitesse d'apparition et de disparition de stimulus. Nous savons en effet qu'une demande attentionnelle forte suite à l'ajout d'une tâche perturbatrice provoque une diminution d'effet McGurk, interprétée par les auteurs comme due au fait que l'attention nécessaire à l'intégration multisensorielle est perturbée par la tâche seconde (Alsius et al, 2005). Ainsi, on pourrait de demander si la rapidité de la tâche pour des contextes courts ne pourrait pas conduire à la fois à une baisse d'effet McGurk due à ce facteur attentionnel, et à une augmentation de temps de réponse. Cependant, en regardant plus précisément, si cette hypothèse expliquait intégralement les résultats alors elle devrait s'appliquer au cas d'une cible seule sans contexte. Or une telle cible est associée à la fois à un temps de réponse élevé, et à un effet McGurk élevé également. Ainsi, l'hypothèse de la charge attentionnelle ne suffit pas à interpréter les résultats, et c'est bien sur un mécanisme lié au processus de fusion lui-même – en l'occurrence, selon notre interprétation, un mécanismes de liage-déliage – qu'il faut chercher l'explication des grandes tendances de nos résultats.

Nous avons alors imaginé que des mécanismes neuronaux dynamiques de type « adaptation neuronale » appliqués au déliage proprement dit pourraient être en jeu : une courte durée d'incohérence produirait un effet de déliage maximal, puis le système se stabiliserait, produisant un niveau de déliage moindre à partir des trois syllabes. Cette interprétation reste très fragile faute d'éléments supplémentaires et de confirmation expérimentale. Enfin, il n'est pas impossible qu'un effet de surprise associant augmentation du temps de réponse et diminution de l'effet McGurk vienne moduler le phénomène de liage-déliage et produire les effets complexes observés.

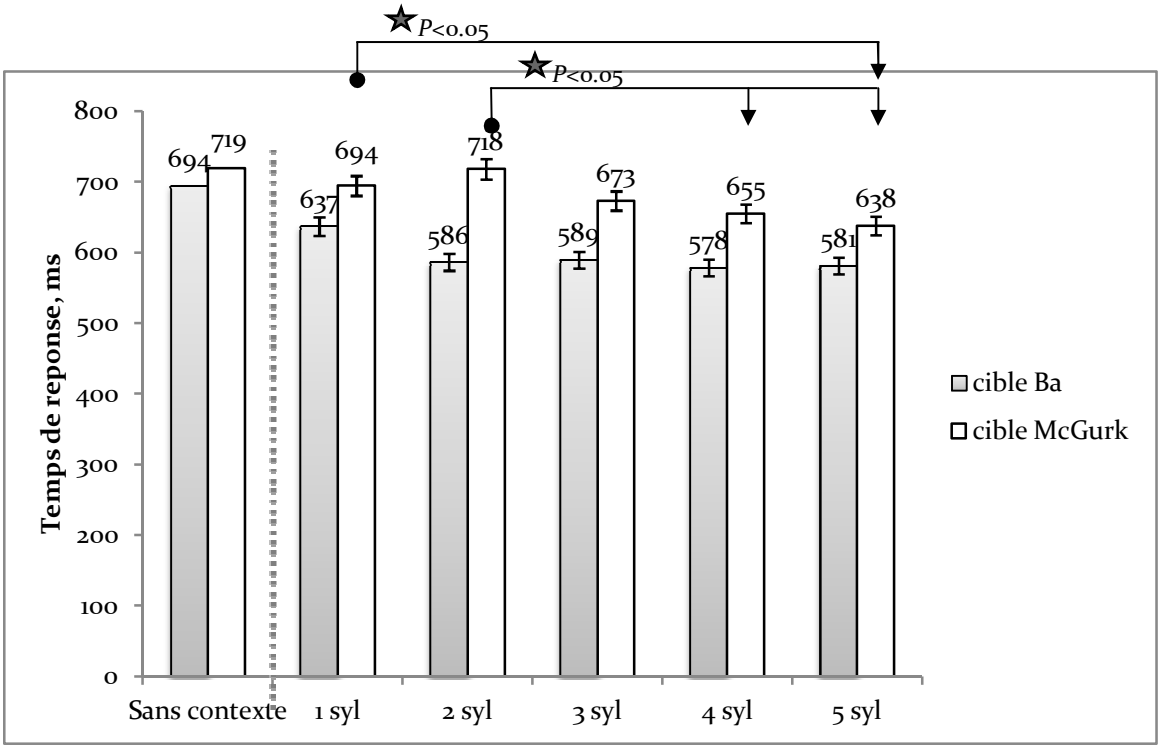


Figure 92 – Temps de réponse en fonction de la durée, ms

11.4 Discussion

Les résultats de cette expérience confirment les résultats obtenus précédemment : l'incohérence du contexte provoque un déliage de la parole audiovisuelle qui modifie les résultats de la fusion, et leur indicateur dans ce travail, le niveau d'effet McGurk. Les deux contextes « phonétique » et « incohérent » permettent l'un et l'autre de produire un effet significatif de déliage, mais le contexte incohérent fournit un effet plus important que le contexte phonétique, qui est, de fait, « moins incohérent ».

Le déliage est rapide : une durée d'incohérence de l'ordre de la syllabe suffit pour obtenir un déliage maximal. Ceci n'est pas si clair dans le cas du contexte phonétique, qui implique vraisemblablement d'autres mécanismes avec une interaction entre les bas niveaux des primitives spectro-temporelles et le niveau phonétique plus élevé. En tout état de cause, les résultats de cette expérience, confirmant l'Expérience 5, montrent que le processus de liage peut être influencé par les niveaux supérieurs de décodage phonétique. Enfin, l'analyse des temps de réponse montre que les cibles « McGurk » sont traitées plus lentement que les cibles « Ba » dans tous les contextes. Ceci confirme a minima que les sujets traitent l'information visuelle dans tous les cas.

Pendant les 2 premières syllabes du contexte incohérent on observe une diminution de la perception d'effet McGurk et une augmentation du temps de réponse. Nous interprétons ces résultats – qui demandent à être confirmés –comme l'existence d'un possible effet d'adaptation déliage, avec amplification de l'effet à durées courtes, puis saturation et décroissance à durées plus longues. Cet effet est probablement renforcé par un mécanisme de charge attentionnelle lié à l'effet de surprise des sujets à l'arrivée d'un signal court et inattendu.

La dernière étape de notre chemin d'expérimentation consiste à étudier comment les processus de déliage peuvent être contrebalancés par des processus de reliage : autrement dit, comment on retourne dans un état « lié » après une phase de déliage. Ce sera l'objet du chapitre suivant.

Chapitre 12. Expérience 7. Mise en évidence d'un processus de reliage

12.1 Objectifs et hypothèses

Dans les expériences précédentes nous avons mis en évidence un phénomène de modulation contextuelle de l'effet McGurk, que nous avons interprété en relation avec l'hypothèse d'une étape de liage audiovisuel et l'existence d'un processus de déliage dont nous avons étudié certaines caractéristiques, notamment l'influence de l'incohérence phonétique et temporelle, ainsi que la durée minimale pour provoquer un déliage et une baisse significative de l'effet McGurk. Nous avons notamment montré que le mécanisme de déliage peut être très rapide : une syllabe incohérente est suffisante pour délier les sources auditif et visuel. La question que nous nous posons dans ce chapitre concerne la possibilité de rebasculer d'un état délié à un état lié. Il s'agit donc de tenter de mettre en évidence un mécanisme de « reliage ». Cette question peut se décomposer en deux points. D'une part, comment peut-on « re-lier » les sources auditifs et visuelles, par quel type de matériau audiovisuel ? Ensuite, quelle est la durée minimale nécessaire pour un reliage réussi, ce qui implique donc la caractérisation dynamique du processus de reliage?

12.2 Méthodologie

12.2.1 Principe

Pour mettre en place cette expérience qui vise à étudier le reliage audiovisuel de la parole, il nous faut commencer par expliciter certaines hypothèses. D'abord, rappelons que nous considérons le processus de liage comme un système dynamique qui change son état au cours de temps en fonction des entrées et produit une action de modulation de la fusion audiovisuelle. Ensuite, nous supposons que l'état par défaut est lié. Cette seconde hypothèse repose sur trois séries d'éléments : (i) nous savons que le signal visuel améliore la détection de parole (Schwartz et al, 2004), (Kim & Davis, 2004), (Grant & Seitz, 2000), ce qui est compatible avec une hypothèse de liage a priori ; (ii) le fait même de l'existence de l'effet McGurk chez la plupart des sujets, ainsi que l'observation des résultats de l'expérience précédente qui montrent que le taux d'effet McGurk en contexte cohérent est équivalent à celui d'un effet McGurk sans contexte (Figure 88) ; (iii) enfin, plus généralement, la théorie de la Gestalt (§3.2.1) et les hypothèses de Bregman dans le cadre de l'analyse de scènes auditives (§3.2.2), reposent sur l'hypothèse d'un état par défaut de groupement spontané des primitives, ce qui est d'ailleurs confirmé par Hupé et al. (Hupé & Pressnitzer, 2012) aussi bien dans les mécanismes de multistabilité auditive (streaming auditif) que visuelle (« plaids »).

Sur cette base, nous proposons un paradigme dans lequel nous cherchons d'abord à délier les deux sources par un « contexte » incohérent, puis nous essayons de les relier par un stimulus que nous baptisons « reset » (Figure 93). Nous avons étudié deux types de reset : un reset cohérent, qui consiste en une séquence de syllabes audiovisuelles cohérentes et un reset fixe, qui consiste en un image fixe du locuteur associée à un silence. Ainsi, le stimulus complet consiste en une séquence (1) d'un contexte incohérent (2) suivi d'un reset (3) puis d'une cible. Nous gardons les cibles« Ba » ou « McGurk » des expériences précédentes.

Dans cette expérience, nous intégrons aussi un stimulus contrôle, sans reset, qui consiste en un contexte cohérent et une cible. Ce stimulus contrôle nous permet de comparer les résultats de cette expérience par rapport à ceux des expériences précédentes.

Dans l'expérience précédente nous avons mis en évidence de faibles différences significatives entre les durées de contexte incohérent les plus courtes (1, 2 syllabes, avec un maximum de réponses « ba ») et les durées plus longues (3, 4 syllabes) (Figure 88). Aussi, dans cette expérience, nous avons choisi d'utiliser deux durées de contexte incohérent, produisant des effets potentiellement différents, 2 et 4 syllabes, afin d'assurer une variété des conditions expérimentales et de vérifier si la différence se reproduit.

Dans l'expérience précédente nous avons vu que l'évolution temporelle du mécanisme de déliage est un processus rapide. Dans cette expérience nous nous attendons à observer que le re-liage fonctionne également sur des dynamiques rapides, aussi nous nous proposons de tester des durées de reset de 0 (pas de reset), 1, 2 et 3 syllabes, correspondant approximativement à des silences de 0 (pas de reset), 480 ms, 1000 ms et 1480 ms dans le reset fixe.

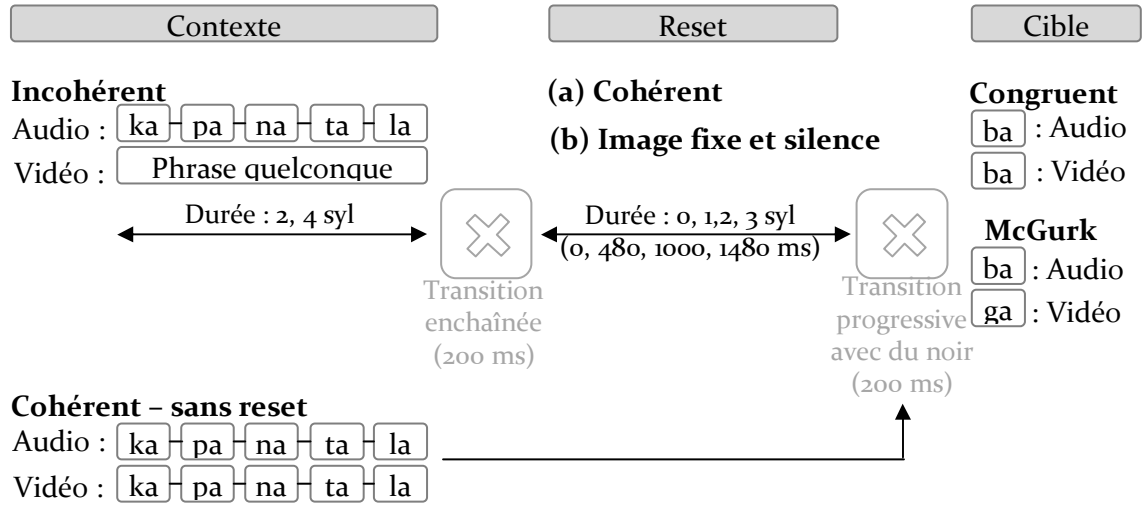


Figure 93 - Principe expérimental

12.2.2 Stimuli

12.2.2.1 Contexte

Les stimuli de contexte cohérent et incohérent sont les mêmes que dans l'expérience précédente, en nous limitant aux durées de 2 et 4 syllabes. Rappelons que le contexte se termine par la coupure d'une cible 240ms avant son burst. Nous avons repris les 4 variations de chaque type et de chaque durée du contexte.

12.2.2.2 Reset

La préparation du reset cohérent est effectuée de manière similaire à celle du contexte cohérent dans le corpus de base. Les contextes de 15 syllabes ont servi pour créer des resets de 3 syllabes, les contextes de 20 syllabes pour créer des resets de 2 syllabes, et la dixième syllabe

des contextes de 20 syllabes pour créer des resets d'une syllabe. La coupure est effectuée de la même manière que pour les contextes, selon la même procédure que dans l'expérience précédente. Nous avons 4 exemplaires de chaque condition.

Le reset fixe est préparé à l'aide d'une image du locuteur avec une expression neutre (Figure 94) associée à un silence acoustique. Nous avons choisi des durées similaires à celles des resets cohérents soit 480 ms pour une syllabe, 1000 ms pour 2 syllabes et 1480 ms pour 3 syllabes.

Figure 94 - image neutre utilisé pour créer un reset fixe

En faisant la coupure de cette manière nous pouvons enchaîner le contexte, le reset et la cible de façon homogène, sans rupture perceptive du rythme pour tout un stimulus. De plus nous avons assuré une variété des contextes et des resets et nous avons gardé la cohérence de cette expérience par rapport aux expériences précédentes.

12.2.2.3 Assemblage d'un stimulus complet

Afin de créer un stimulus complet (contexte plus reset plus cible) il nous faut joindre les parties issues de séquences différentes. Pour éviter une rupture brutale perceptive en modalité visuelle et obtenir une transition plus lisse nous avons appliqué systématiquement une transition enchaînée entre contexte et reset et une transition progressive (passant par une image noire) entre reset et cible. La durée des deux transitions est de 200 ms ou 5 images.

Dans la transition enchaînée la première image du reset est fusionnée avec les trois dernières images du contexte avec des proportions croissant de 25% à 33% puis 50%., puis la dernière image du contexte est fusionnée avec les deux premières images du reset dans les proportions 33% et 25% (Figure 95). La transition progressive par une image noire est réalisée sur le même principe, fusionnant les trois dernières images du reset et les deux premières images de la cible, à travers un masque avec le niveau du noir 1/3-2/3-1-2/3-1/3. Le principe est détaillé dans la §9.2.2.4.

Nous avons appliqué les transitions différentes qui nous permettent d'une part de maintenir la cohérence entre l'information contextuelle et la cible testée dans nos expériences précédentes et d'autre part d'avoir un passage lisse, très peu visible, entre le contexte et le reset.

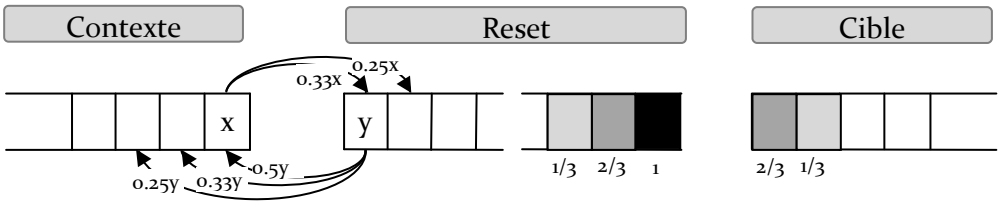


Figure 95 – Transition enchaînée et transition progressive avec image noire sur 5 images (200ms)

12.2.3 Plan d'expérience

Nous avons séparé cette expérience en deux blocs selon le type de reset. Nous avons préparé 5 variantes de chaque bloc. Ces variantes et l'ordre de passage des blocs sont distribués aléatoirement entre les sujets, en contrôlant également la latéralisation des réponses.

Au total nous avons préparé une groupe de 12 cibles McGurk et 4 cibles Ba enchaînées dans une première condition avec le contexte cohérent pour 2 durées (2 et 4 syllabes) avec 4 variations pour chaque durée, et dans une deuxième condition avec le contexte incohérent de durée 2 ou 4 syllabes et avec un reset soit cohérent, soit fixe de durée 0 (absence de reset), 1,2, 3 syllabes avec quatre occurrences différentes pour chaque durée du reset. La répartition des cibles est faite de façon à présenter les mêmes 12 cibles McGurk pour chaque durée du contexte, mais la répartition entre les variantes des durées de reset est aléatoire. Comme dans toutes nos expériences nous conservons les proportions ¾ de cibles McGurk et ¼ de cibles Ba. Au total dans un bloc (selon le reset, cohérent vs. fixe) nous avons présenté 160 stimuli (2*(12+4) + 2*4*(12+4)). La répartition des stimuli de chaque type est indiquée dans le Tableau 32.

Tableau 32 - Nombre des stimuli présentés dans un bloc donné (reset cohérent ou reset fixe)

Contexte cohérent		**Contexte incohérent**			
Sans reset		**Sans reset**		**Reset (cohérent ou fixe)**	
« ba »	**« McGurk »**	**« ba »**	**« McGurk »**	**« ba »**	**« McGurk »**
8 (1/4)	24 (3/4)	8 (1/4)	24 (3/4)	24 (1/4)	72 (3/4)

En résumé, l'expérience consiste en :

- 320 stimuli au total, présentés en deux blocs en fonction de type du reset, répartis aléatoirement à l'intérieur du bloc. La durée d'un bloc est environ 10 minutes.
- 5 variantes de chaque type de bloc, chacune présentée à 4 sujets
- Durée du contexte : 2 ou 4 syllabes CV
- Durée du reset 0 (sans reset), 1, 2, 3 syllabes CV
- 2 types de réponse possible « ba » ou « da » en ligne, choix forcé

Dans cette expérience nous avons utilisé le logiciel Presentation® software (Version 0.70, www.neurobs.com), ce qui permet de mesurer les temps de réponse. Les autres

conditions de passation d'expérience, les consignes, les types de réponses étaient les mêmes que dans l'Expérience 1 (Paragraphe 5.4.2). Une moitié des sujets a commencé par un bloc « reset fixe », l'autre moitié par un bloc « reset cohérent ».

12.2.4 Sujets

20 sujets français ont participé à l'expérience avec une vision et une audition normale ou corrigée (9 femmes et 11 hommes, entre 18 et 60 ans avec 25,7 ans en moyenne, 19 droitiers et 1 gaucher). Tous les sujets ont donné un consentement éclairé à participer à l'expérience et n'étaient pas au courant du but de l'étude.

12.3 Résultats

12.3.1 Scores bruts

Nous avons appliqué la même méthode de traitement des réponses et d'analyse statistique que dans toutes nos expériences (Paragraphe 5.5). Mais, comme au chapitre 11, comme nous utilisons des contextes de durée courte, dans cette expérience nous avons de nouveau réduit le période d'acceptation des réponses de 3500 ms à 1200 ms. Nous avons vérifié que cela nous a permis d'acquérir la grande majorité des réponses.

Nous présentons le détail des réponses fournies par les sujets dans la matrice de confusion (Tableau 33). Le taux d'erreur (réponses absentes) est de 10,2%. Nous constatons une remontée sensible du taux d'erreur, mais avec une distribution comparable aux expériences précédentes. Cette augmentation peut être attribuable à la complexité apparente des séquences.

Tableau 33 – Matrice de confusion

	Stimuli		Stimuli présentés	Réponse « ba »		Réponse « da »		Plusieurs réponses « ba »		Plusieurs réponses « da »		Absence de réponses		Absence plusieurs réponses	
Bloc avec reset cohérent	coh sans reset	Ba	160	134	(84%)	2	(1%)	4	(3%)	1	(1%)	1	(1%)	18	(11%)
		McG	480	161	(34%)	217	(45%)	14	(3%)	11	(2%)	23	(5%)	54	(11%)
	Incoh	Ba	640	530	(83%)	8	(1%)	26	(4%)	1	(0%)	4	(1%)	71	(11%)
		McG	1920	862	(45%)	723	(38%)	83	(4%)	27	(1%)	58	(3%)	167	(9%)
Block avec reset fix	coh sans reset	Ba	160	137	(86%)	3	(2%)	5	(3%)	0	(0%)	1	(1%)	14	(9%)
		McG	480	179	(37%)	242	(50%)	9	(2%)	5	(1%)	23	(5%)	22	(5%)
	Incoh	Ba	640	566	(88%)	9	(1%)	14	(2%)	0	(0%)	0	(0%)	51	(8%)
		McG	1920	1169	(61%)	552	(29%)	40	(2%)	13	(1%)	37	(2%)	109	(6%)

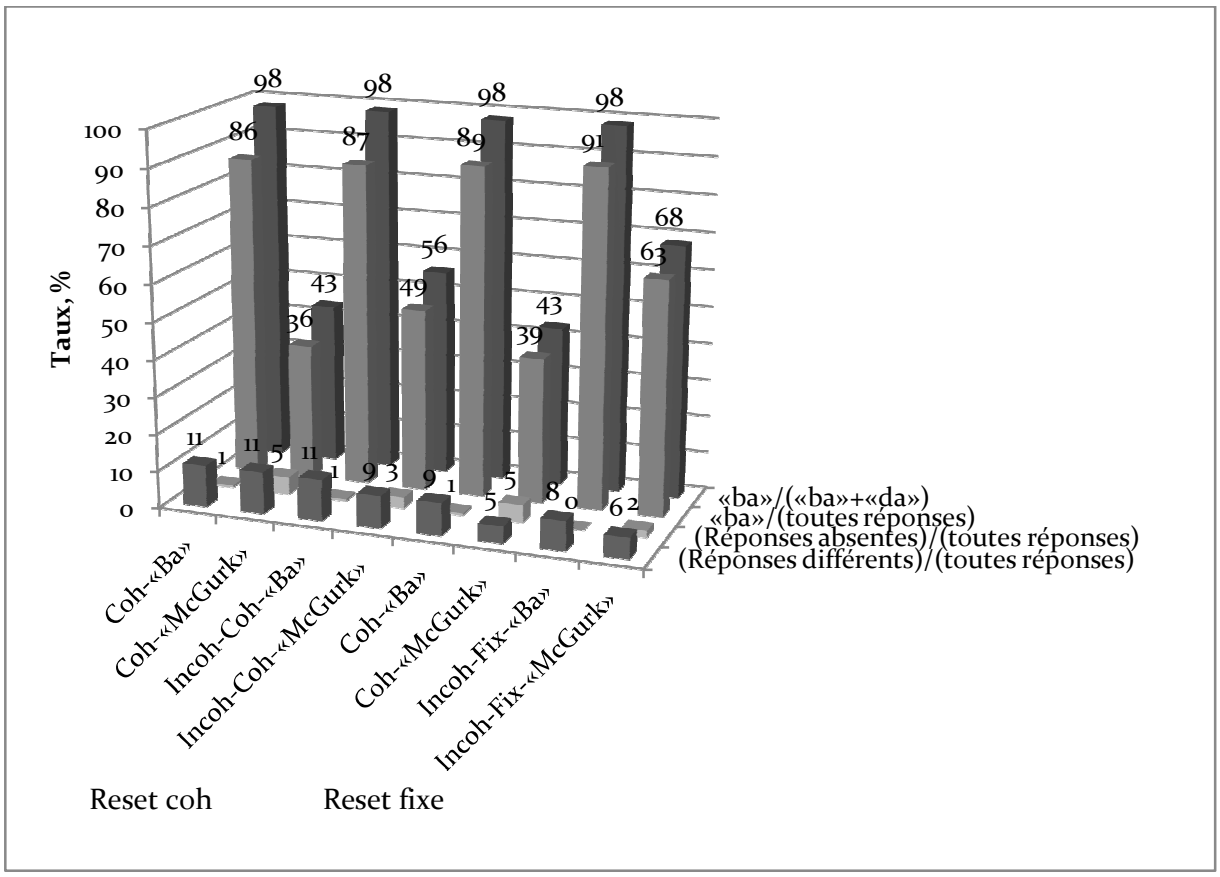


Figure 96 – Expérience 7, données brutes, %.

L'analyse du comportement des participants pour les cibles « McGurk » est présentée sur la Figure 97. Dans les expériences précédentes nous avions vu que le comportement des sujets était bien dépendant du contexte. Dans cette expérience nous présentons l'analyse en fonction du reset en ordonnant les sujets en fonction de leurs réponses en reset cohérent. La plupart des sujets de cette expérience présentent une bonne perception d'effet McGurk. Nous observons une tendance à avoir des scores en reset cohérent inférieurs à ceux de la condition en reset fixe, avec des variations parfois nettes entre les sujets.

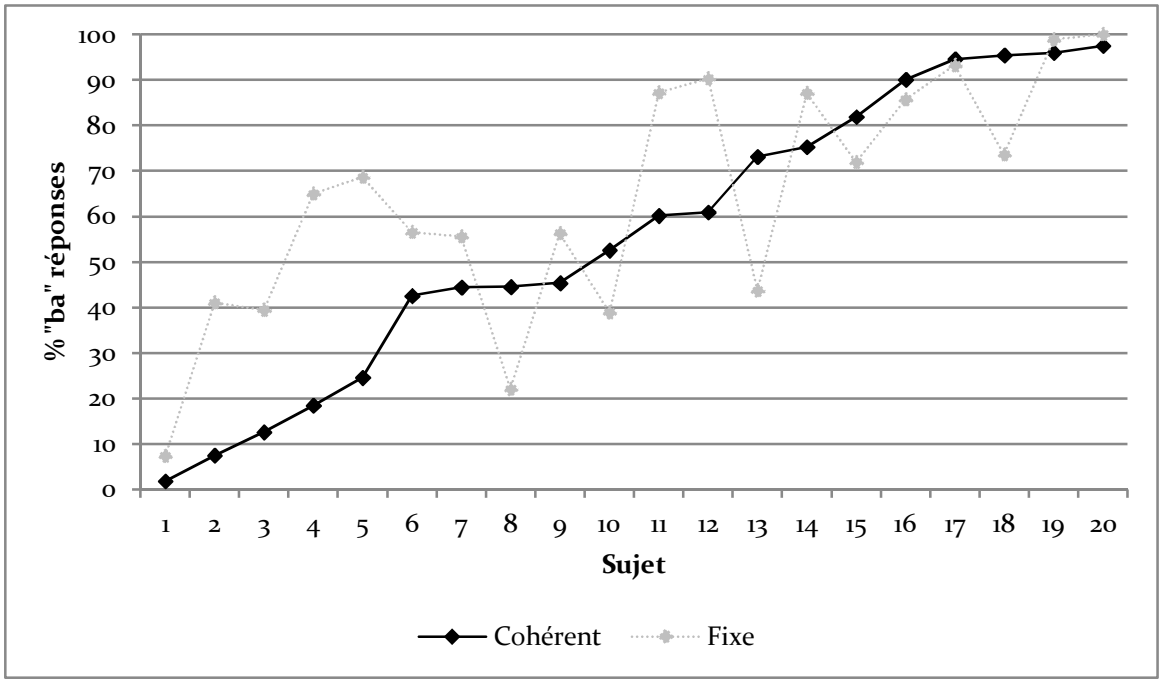


Figure 97 - Perception d'effet McGurk par sujet selon le reset (« ba »/(« ba » + « da »)).

12.3.2 Analyses statistiques des pourcentages de réponse

12.3.2.1 Effet du reset et de la durée du reset

Comparons d'abord les résultats de cette expérience avec nos études précédentes. Sur la Figure 98 nous voyons que la perception d'effet McGurk dans le cas du contexte cohérent est de 55% (45% de réponses « ba ») contre 20-25% dans le cas du contexte incohérent sans reset (75 à 80% de réponses « ba »). Ces données confirment nos résultats précédents.

Les variations de scores dans la Figure 98 en fonction du type de reset et de la durée nous suggèrent des réponses à la question principale de cette expérience et nous informent sur le processus de re-liage. En effet, les résultats semblent clairement différents entre le reset cohérent et le reset fixe. Si l'on part de la condition sans reset dans les deux blocs, qui donne un score maximal de réponses « ba » autour de 75 à 80% (20-25% d'effet McGurk), on observe que le score évolue peu pour le reset fixe : pour un reset de l'équivalent d'une syllabe, soit une durée de 480 ms, le score baisse un peu à 69% (remontée légère de l'effet McGurk) pour rester stable jusqu'à une durée de reset de 1480 ms : il semble ainsi qu'il y n'ait que peu de « re-liage » dans cette condition. Dans le cas du reset cohérent au contraire, nous observons une décroissance régulière du score de réponses « ba », correspondant à une augmentation de l'effet McGurk que nous interprétons comme un reliage progressif jusqu'à retour au niveau de liage de base, équivalent à celui produit par un contexte cohérent, pour une durée de reset cohérent de trois syllabes.

Une ANOVA à mesures répétées centrée sur les stimuli McGurk et le contexte incohérent (Tableau 34) montre un effet significatif des facteurs reset [$F(1,19)=5.23$, $P=0.034$] et durée du reset [$F(3,57)=14.35$, $P<0.001$], ainsi que de leur interaction [$F(3,57)=8.46$, $P<0.001$]. Un test post hoc montre une absence de différence significative entre toutes les conditions de

durées pour le reset fixe. Dans le cas du reset cohérent il existe au contraire une différence entre les durées 0 et 2 syllabes d'une part, et entre (0, 1, 2) et 3 syllabes d'autre part ($P<0.05$). Ainsi l'effet significatif du facteur « durée du reset » vient du reset cohérent. Cette analyse confirme parfaitement la lecture qualitative faite précédemment.

Les résultats de cette expérience montrent donc une différence claire du rôle des resets cohérent et fixe sur l'état de liage audiovisuel. Une simple pause de durée allant jusqu'à une seconde et demie ne suffit pas à relier les deux sources. Au contraire, il apparaît que tout au long de cette pause l'état cognitif du sujet est en quelque sorte « gelé » dans un état « délié ». Pour que le reliage puisse avoir lieu il faut introduire un signal cohérent. Au contraire du déliage, qui est très rapide, le reliage a un caractère progressif et demande plus de temps. Au minimum trois syllabes semblent nécessaires (~1480 ms) pour relier complètement les deux sources, ainsi que l'atteste la différence significative entre les conditions de reset cohérent à 2 vs. 3 syllabes dans le test post-hoc mentionné ci-dessus.

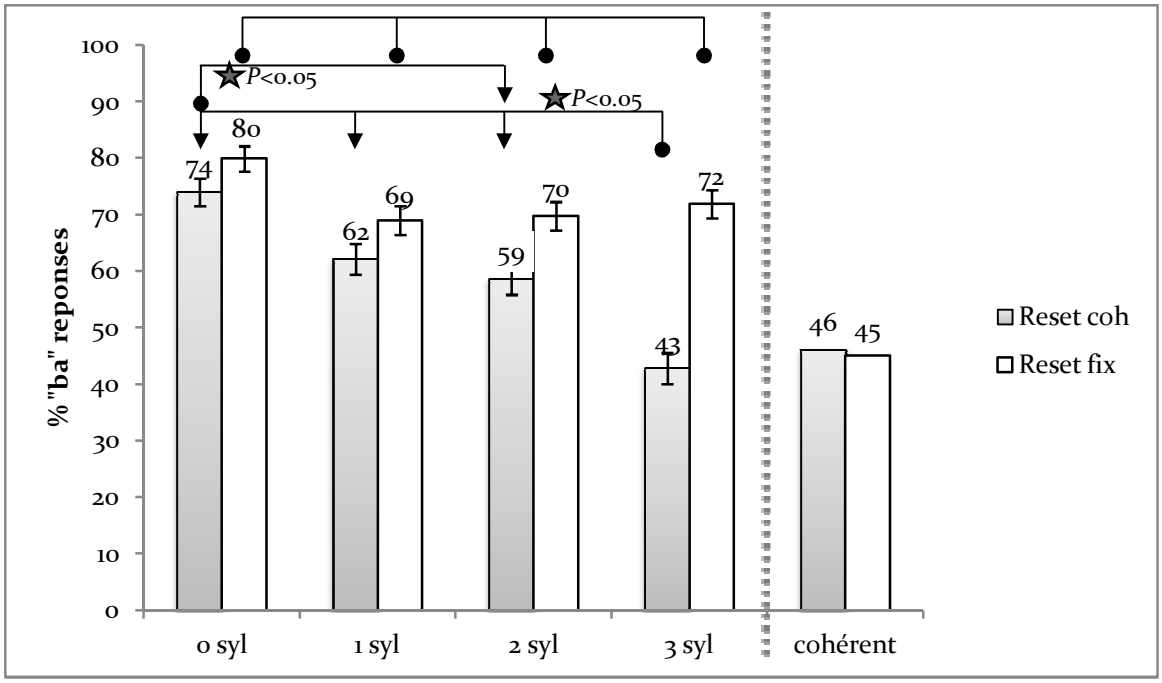


Figure 98 - Taux de réponses (« ba »/(« ba » + « da »)) dans l'Expérience 7.

Tableau 34– ANOVA à mesures répétées: reset, durée du reset, durée du contexte.

Source		Somme des carrés de type III	ddl	Moyenne des carrés	D	Sig.
Reset : coh. vs.fixe	Sphéricité supposée	1,643	1	1,643	5,229	,034
Erreur (reset)	Sphéricité supposée	5,969	19	,314		
Durée du reset : 0, 1, 2, 3 syls	Sphéricité supposée	1,941	3	,647	14,345	,000
Erreur (durée du reset)	Sphéricité supposée	2,571	57	,045		
Durée du contexte : 2, 4 syl	Sphéricité supposée	,482	1	,482	25,046	,000
Erreur (durée du contexte)	Sphéricité supposée	,365	19	,019		
*Reset * Durée du reset*	Sphéricité supposée	,799	3	,266	8,463	,000
*Erreur (Reset * Durée du reset)*	Sphéricité supposée	1,794	57	,031		
*Reset * Durée du contexte*	Sphéricité supposée	,011	1	,011	,531	,475
*Erreur (Reset * Durée du contexte)*	Sphéricité supposée	,376	19	,020		
*Durée du reset * Durée du contexte*	Sphéricité supposée	,078	3	,026	1,420	,246
*Erreur (Durée du reset * Durée du contexte)*	Sphéricité supposée	1,044	57	,018		
*Reset * Durée du reset * Durée du contexte*	Sphéricité supposée	,082	3	,027	1,688	,180
*Erreur (Reset * Durée du reset * Durée du contexte)*	Sphéricité supposée	,926	57	,016		

12.3.2.2 Effet de la durée du contexte

Dans l'Expérience 6 nous avions observé une différence entre les durées de contexte incohérent courtes (1,2 syllabes) et moins courtes (3, 4 syllabes). Dans l'expérience présente nous avons donc testé d'éventuelles différences de score entre durées de contexte de 2 et 4 syllabes. Sur la Figure 99 nous observons une différence de perception d'effet McGurk selon la durée du contexte. La durée la plus courte (deux syllabes incohérentes) fournit des scores plus élevés donc conduit à moins d'effet McGurk que la durée plus longue de 4 syllabes et ce dans les deux resets. L'ANOVA présentée ci-dessus confirme l'effet significatif de la durée du contexte (Tableau 34). Ainsi cette expérience reproduit le résultat selon lequel les durées de contexte 2 et 4 syllabes sont perçues différemment. Cette différence persiste dans les deux

resets. Elle suggère qu'une durée courte produit plus de déliage qu'une durée plus longue, ce que nous avons interprété comme la conséquence possible d'un mécanisme d'adaptation au déliage.

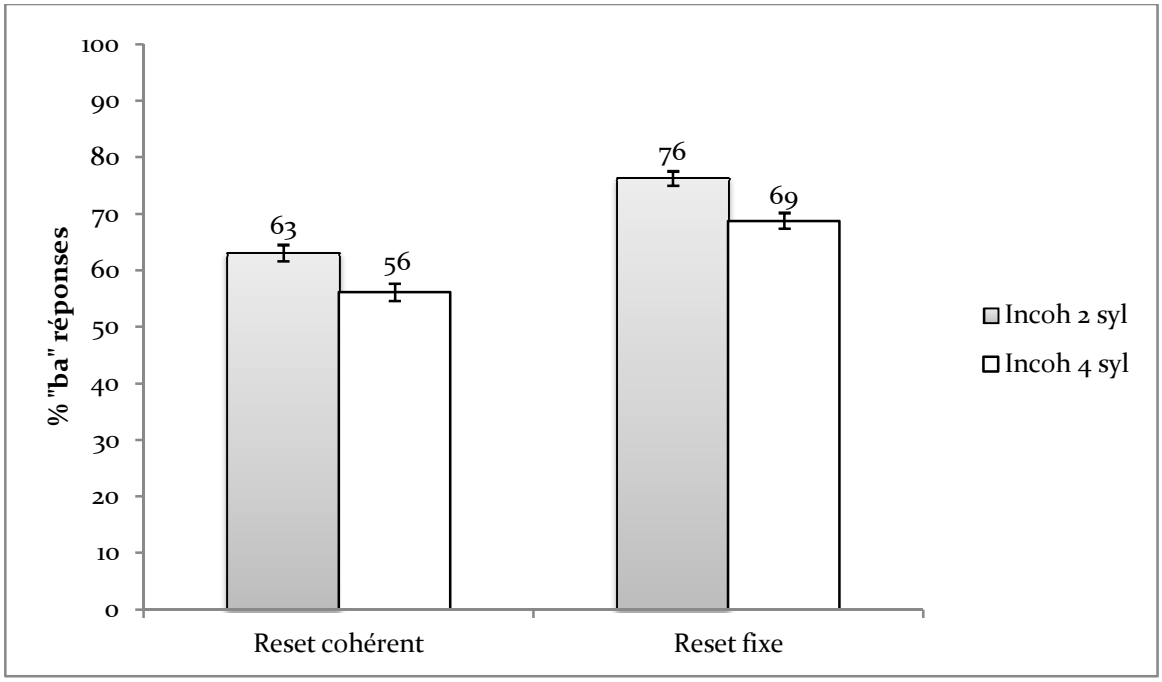


Figure 99–Effet de la durée de contexte sur le taux de réponses (« ba »/(« ba » + « da »)) dans l'Expérience 7

12.3.3 Temps de réponse

Nous allons maintenant analyser les temps de réponse. Sur la Figure 100 nous présentons le temps de réponse pour les cibles Ba et McGurk en contexte incohérent, toutes durées de contexte et de reset confondues, mais en fonction du reset (cohérent vs. fixe). Les ordres de grandeur des temps de réponse sont cohérents avec ceux de l'expérience précédente (§11.3.3). Comme dans les expériences précédentes les temps de réponse pour les cibles McGurk sont plus longs que pour les cibles Ba. Une ANOVA à mesures répétées sur trois facteurs (Tableau 35) montre un effet significatif du facteur cible [$F(1,19)=18.36$, $P<0.001$]. Les facteurs reset [$F(1,19)=1.59$, $P=0.22$] et durée du reset [$F(2.37,45.03)=0.59$, $P=0.59$] ne sont pas significatifs, de même que chacune des interactions entre facteurs. Cette analyse confirme que le temps de traitement dépend uniquement de la cible et pas du contexte, et que donc, si les effets de contexte (déliage, reliage) peuvent moduler l'effet McGurk positivement ou négativement, le temps de traitement lui n'est pas modulé. Ce temps semble ainsi dépendre uniquement de la mesure faite par le sujet de l'incohérence locale entre composante auditive et visuelle dans la cible, congruente ou non, les cibles non congruentes (McGurk) demandant plus de temps de traitement que les cibles congruentes Ba. Nous avons déjà observé et discuté ce phénomène dans les expériences précédentes (§11.3.3).

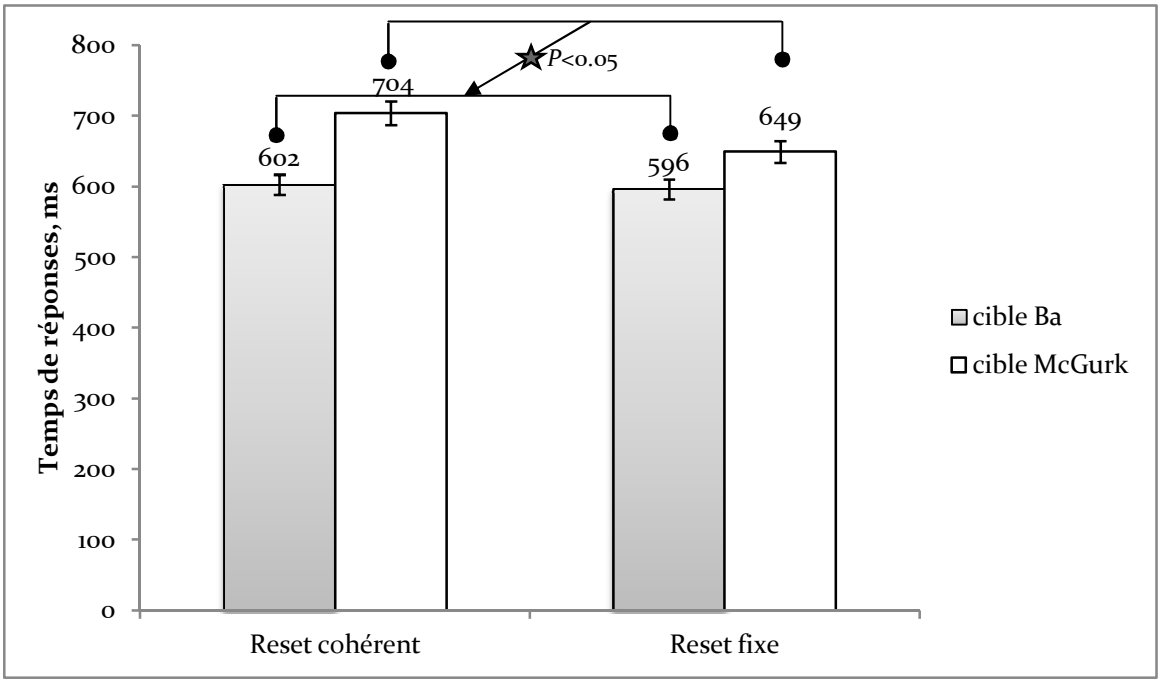


Figure 100 - Temps de réponse, ms, dans l'Expérience 7

Tableau 35– ANOVA à mesures répétées des temps de réponses dans l'Expérience 7 : reset, durée du reset, cible

Source		Somme des carrés de type III	ddl	Moyenne des carrés	D	Sig.
Reset : coh. vs. fixe	Sphéricité supposée	,088	1	,088	1,586	,223
Erreur (reset)	Sphéricité supposée	1,056	19	,056		
Durée du reset : 0, 1, 2, 3 syls	Huynh-Feldt	,047	2,370	,020	,591	,586
Erreur (durée du reset)	Huynh-Feldt	1,509	45,032	,034		
Cible : Ba, McGurk	Sphéricité supposée	1,004	1	1,004	18,364	,000
Erreur (cible)	Sphéricité supposée	1,038	19	,055		
*Reset * Durée du reset*	Huynh-Feldt	,023	1,595	,015	,302	,692
*Erreur (Reset * Durée du reset)*	Huynh-Feldt	1,466	30,300	,048		
*Reset * Cible*	Sphéricité supposée	,047	1	,047	2,404	,137
*Erreur (Reset * Cible)*	Sphéricité supposée	,370	19	,019		
*Durée du reset * Cible*	Huynh-Feldt	,031	1,864	,017	,463	,620

*Erreur (Durée du reset * Cible)*	Huynh-Feldt	1,276	35,420	,036		
*Reset * Durée du reset * Cible*	Huynh-Feldt	,015	2,120	,007	,288	,764
*Erreur (Reset * Durée du reset * Cible)*	Huynh-Feldt	1,005	40,275	,025		

12.4 Discussion

Cette dernière expérience a permis de répondre à la question que nous nous posions au départ. Dans le cadre théorique que nous avons choisi, nous avons d'abord mis en évidence par toute une série d'expériences des effets de modulation négative de l'effet McGurk par divers types de stimuli contextuels audiovisuels incohérents. Nous avons interprété ces effets en supposant l'existence de mécanismes de déliage, intégrés au sein d'un processus général de liage audiovisuel. Les premières expériences de la Partie III ont permis de caractériser ces processus de déliage.

Nous savons maintenant que l'on peut, suite à un processus de déliage, produire des effets de « reliage ». Mais le premier résultat important de ce chapitre est que le reliage ne peut se faire qu'en présentant des matériaux audiovisuels cohérents, c'est-à-dire en donnant des éléments de cohérence qui permettent d'effacer progressivement les éléments incohérents précédents. La dynamique de ce processus de reliage par cohérence est rapide mais néanmoins moins rapide que le processus de déliage lui-même : une ou deux syllabes produisent un déliage maximal, là où il en faut 3 pour effacer totalement ce déliage et revenir au niveau d'effet McGurk initial.

Le second résultat est que la présentation d'un stimulus stationnaire (image fixe et silence) ne permet par contre pas de relier les composantes auditives et visuelles. On peut ainsi en quelque sorte geler l'état cognitif du sujet dans un état délié, ce qui est un résultat assez inattendu et extrêmement intéressant, qui permettra certainement de mettre en place de nouveaux paradigmes dans l'avenir, nous y reviendrons dans notre discussion générale.

Il est intéressant à ce point de rappeler que nous avions déjà tenté de forcer le reliage des deux modalités sensorielles par une alerte audiovisuelle attentionnelle dans l'Expérience 2, et ce sans succès : une simple alerte avant la cible n'avait pas permis de restaurer l'état lié et de remonter l'effet McGurk à son niveau initial. Ce résultat négatif apparaît en bonne cohérence avec l'absence de reliage fournie par un stimulus stationnaire comme le reset fixe utilisé ici.

Notons enfin que cette expérience a permis également de confirmer que les contextes incohérents de durées courtes (2 syllabes) produisent un effet de déliage légèrement mais significativement plus important que les durées plus longues (4 syllabes), ce que nous proposons d'interpréter par un effet d'adaptation au déliage.

Enfin, nous confirmons une nouvelle fois dans cette expérience que les cibles McGurk sont traitées plus lentement que les cibles Ba, et ce indépendamment des effets de déliage et de reliage sur le percept lui même, ce qui confirme que les sujets traitent l'information visuelle dans tous les cas, et qui fournit une donnée expérimentale forte qu'il conviendra d'interpréter dans le cadre d'un modèle global du processus de liage audiovisuel.

C'est précisément à ce portrait général que nous allons maintenant nous attacher dans la discussion générale qui va suivre.

Partie IV
Synthèse

Nous disposons maintenant, en fin de ce travail, d'un corpus important de résultats expérimentaux qui nous semblent originaux et riches. Cette dernière partie vise à en proposer une synthèse, en partant d'une discussion générale qui permettra de faire ressortir nos principales hypothèses et de les mettre en regard de propositions computationnelles et neuronatomiques. Puis nous proposerons une série de perspectives articulées autour de propositions expérimentales diverses.

Chapitre 13. Discussion

13.1 Résumé des principaux résultats

Dans ce manuscrit nous avons présenté une série de sept expériences, dont nous allons proposer ici un résumé.

Nos études ont toutes porté sur la question de la fusion audiovisuelle, que nous avons évaluée par l'étude de la perception d'effet McGurk. Dans nos premières expériences présentées au sein de la Partie II, nous avons réussi à moduler la perception d'effet McGurk en variant le contexte préalable, soit complètement cohérent soit complètement incohérent (Expériences 1 à 4). Le contexte cohérent consiste en une succession de syllabes audiovisuelles, et le contexte incohérent est composé de la même succession de syllabes auditives associées cette fois à un flux de phrases libres présentées dans la modalité visuelle. La modulation a atteint 50% de réduction de l'effet McGurk (passant de 40 à 20 %), dans l'Expérience 4, avec même une disparition quasi complète de l'effet McGurk dans les Expériences 1 et 2, mais sur des stimuli cible sans doute insuffisamment contrôlés, selon les données de l'Expérience 3.

Il est important de rappeler que les sujets ne savaient pas quand les cibles apparaissaient dans le film avec un mélange aléatoire de stimuli avec contexte cohérent et incohérent. Les sujets n'ont donc très vraisemblablement pas modifié leur attention visuelle consciente d'un stimulus au suivant. Notre interprétation globale est celle de l'implication d'un mécanisme de modulation de la fusion audiovisuelle, au sein duquel l'incohérence des flux auditifs et visuels amène les sujets à diminuer le rôle de l'input visuel dans le processus de fusion.

Nous avons alors entrepris dans la Partie III de mieux caractériser ce mécanisme de modulation. Nous avons pu montrer que cet effet apparaît très rapidement. Un contexte composé d'une seule syllabe incohérente est suffisant pour obtenir un effet de réduction de l'effet McGurk maximal (voir résultats de l'Expérience 6, rappelés dans la Figure 101). La modulation décroît ensuite à partir d'une durée de trois syllabes puis reste stable, sans augmentation de l'effet de modulation négative de l'effet McGurk même pour des durées longues allant jusqu'à 20 syllabes, selon l'Expérience 4. Ainsi avec des contextes de forte

incohérence la rupture de fusion est quasi instantanée et réduit immédiatement et de manière stable le poids de l'information visuelle dans le processus de fusion.

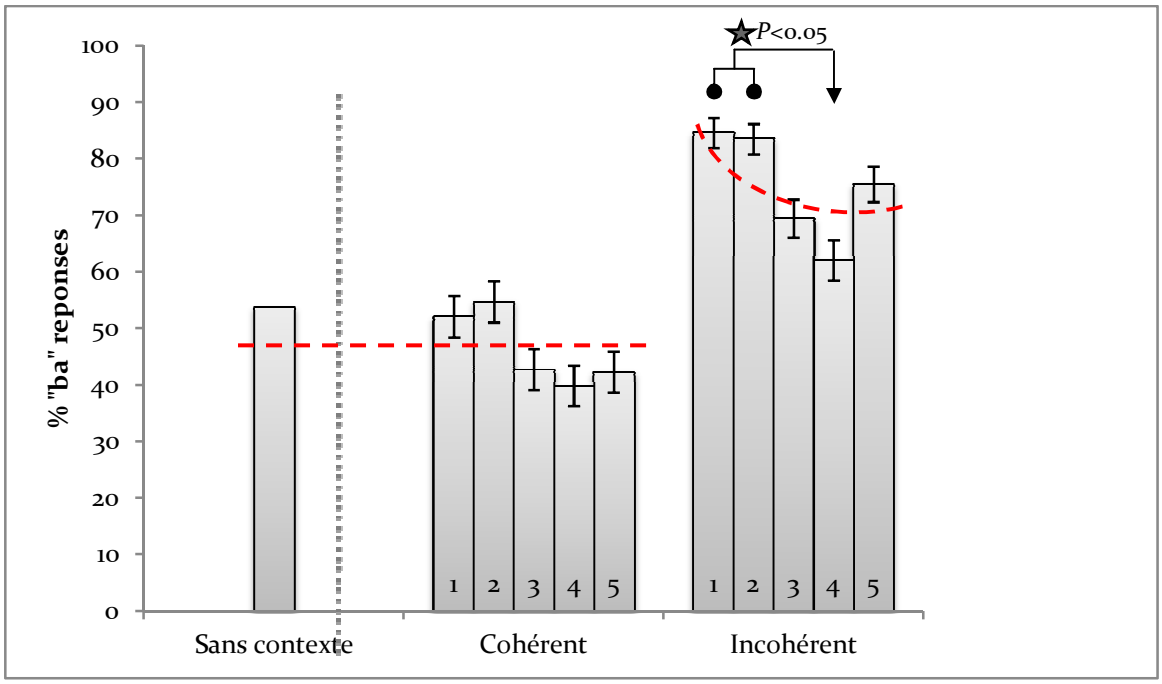


Figure 101 - Identification de cibles « McGurk » (pourcentage de réponses « ba »/ « ba » + « da ») en fonction de la durée du contexte cohérent vs. incohérent, d'après l'Expérience 6

Comme indiqué ci-dessus, dans un premier temps nous avons cherché à démontrer le bien-fondé de notre hypothèse de modulation en utilisant des stimuli contextuels caractérisés par une incohérence maximale. Nous nous sommes alors posés la question de décomposer l'incohérence en composantes élémentaires. Pour ce faire nous avons décomposé l'incohérence maximale selon deux dimensions, incohérence temporelle et incohérence phonétique. Pour l'incohérence phonétique nous avons réalisé un montage dans lequel des syllabes auditives et visuelles incohérentes étaient associées en maintenant la cohérence temporelle. Pour l'incohérence temporelle nous avons simplement décalé le signal auditif par rapport au signal visuel, différemment d'une syllabe à l'autre au sein d'une séquence de plusieurs syllabes, et ce en restant dans la fenêtre d'intégration audiovisuelle [-30, 170] ms (van Wassenhove et al, 2007). Nous avons également testé une condition qui intègre ces deux types d'incohérence. Nous avons montré que les deux types d'incohérence « pure », phonétique et temporelle, produisent un effet significatif (résultats de l'Expérience 5, Figure 102).

L'effet de l'incohérence phonétique pure est moins élevé que celui de l'incohérence complète, comme l'a montré notamment l'Expérience 6. Ceci confirme notre hypothèse que l'incohérence peut être décomposée en composantes élémentaires.

D'autre part le décrochage produit par l'incohérence phonétique est plus élevé que celui dû à l'incohérence temporelle (Figure 102). Cependant nous nous sommes cantonnés à de très faibles décalages temporels afin de rester dans la fenêtre d'intégration audiovisuelle. L'élément nouveau ici est qu'une succession de faibles décalages temporels, chacun susceptible

de ne produire aucune modification de l'effet McGurk, peut produire un effet global sur la fusion audiovisuelle.

Le décrochage produit par l'incohérence phonétique semble également rapide, puisque les résultats de l'Expérience 6 ne présentent pas de différence de diminution d'effet McGurk entre 1 et 5 syllabes. Ainsi, comme dans le cas de l'incohérence complète, il semble qu'une incohérence phonétique sur une seule syllabe suffise pour produire un décrochage.

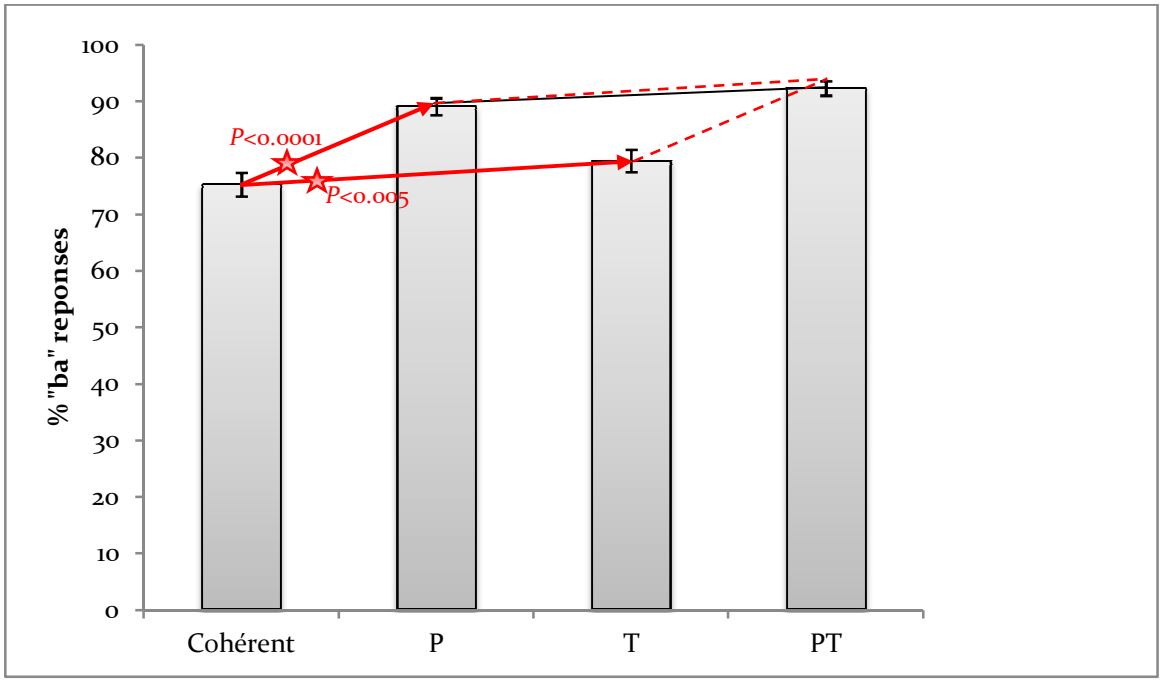


Figure 102 - Identification de cibles « McGurk » (pourcentage de réponses « ba »/ « ba » + « da ») en fonction du contexte, d'après l'Expérience 5

Nous sommes alors demandés si l'on pouvait remettre le niveau d'effet McGurk à son taux original après le décrochage dû à un contexte incohérent. Ainsi nous avons rajouté une séquence de contexte cohérent ou de pause après un contexte incohérent. Nous avons montré que l'introduction d'une séquence de trois syllabes cohérentes ramenait le taux de fusion au même niveau que dans le contexte cohérent de base (Figure 103). Ainsi il apparaît que, alors que le décrochage provoqué par le contexte incohérent est rapide, la restauration est plus lente et progressive.

Par contre, une simple pause allant jusqu'à une seconde et demie ne permet pas de remettre à niveau le taux de fusion. Ceci suggère qu'un retour à l'état initial n'est pas automatique et qu'il est nécessaire de fournir certaines conditions expérimentales (qui restent largement à définir, nous y reviendrons), pour provoquer un « reset ».

Ce résultat sur l'incapacité d'une pause à ramener le sujet dans son état initial est extrêmement intéressant. Il nous suggère en effet la possibilité de « geler » l'état cognitif d'un sujet dans un état « délié ». Ceci ouvre la voie à la construction de nouveaux paradigmes expérimentaux qui nous nous permettront de mieux étudier l'effet du contexte préalable, notamment dans le cas d'études en neuroimagerie et neurophysiologie, nous y reviendrons.

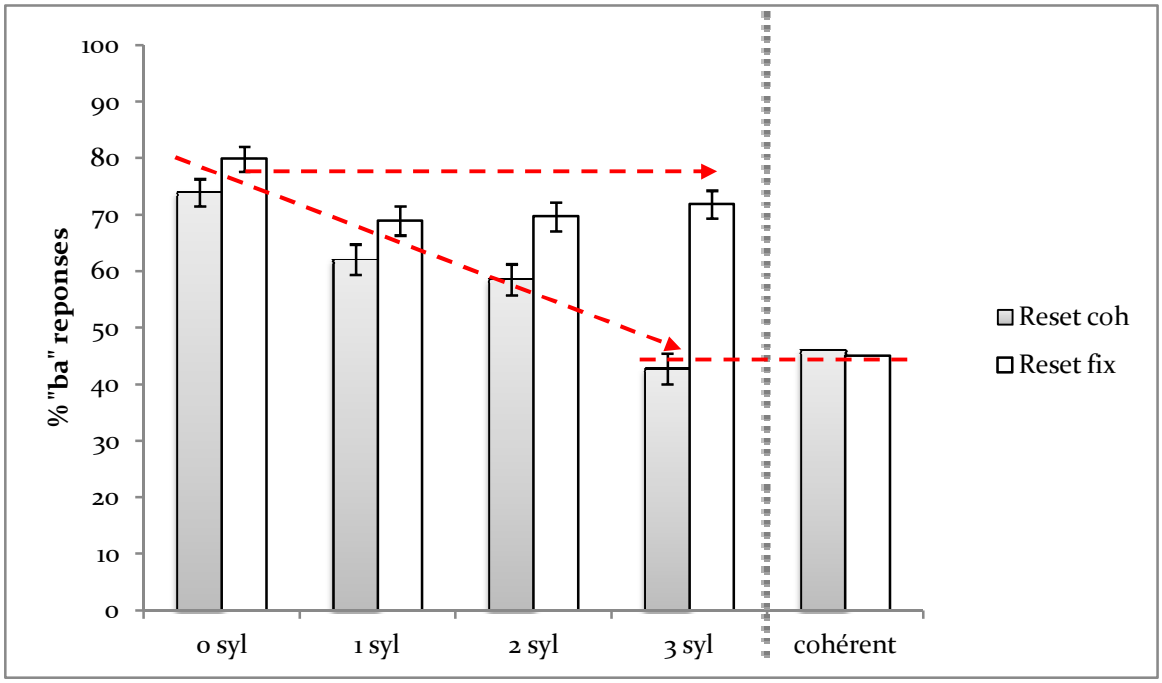


Figure 103 - Identification de cibles « McGurk » (pourcentage de réponses « ba »/ « ba » + « da ») en fonction du reset, d'après l'Expérience 7

Un autre volet important de nos études concerne l'analyse des temps de réponse. En effet, il apparaît que systématiquement, dans les expériences où nous avons pu mesurer précisément ces temps, les cibles McGurk impliquent des temps de réponse plus élevés que les cibles Ba (cet effet, significatif dans les Expériences 6 et 7, ne l'est cependant dans l'Expérience 4 que dans une analyse à effet sujet fixe). Rappelons que dans les Expériences 1, 2 et 5 utilisant un logiciel non adapté, nous ne disposons pas de ces résultats de temps de réponse. Dans le Tableau 36 nous donnons le résumé des analyses de temps de réponse.

Le point essentiel est que dans ces expériences où nous disposons de données fiables sur les temps de réponse, les facteurs autres que le facteur « cible » ne produisent pas d'effet significatif sur le temps de réponse, que ce soit le type de contexte, sa durée, ou le type de reset. Or ces facteurs produisent par contre des effets significatifs sur les réponses (variations du pourcentage de réponses « ba »). Cette forte divergence entre réponses et temps de réponse est un résultat important de nos expériences.

D'une manière générale la cohérence de nos résultats expérimentaux à travers les différentes expériences est forte et très rassurante : cohérence des résultats de diminution de l'effet McGurk en contexte incohérent à travers les expériences 1, 2, 4, 5, 6 et 7 – cohérence des données sur la rapidité de cette diminution à travers les expériences 1, 2, 4, 5, 6 et 7 ; cohérence de la possibilité d'obtenir un effet avec un simple contexte d'incohérence phonétique à travers les expériences 5 et 6 ; cohérence des résultats sur les temps de réponse à travers les expériences 4, 6 et 7. Un travail récent nous a permis également de retrouver les résultats sur les mécanismes de reliage plus lents, dans des proportions extrêmement proches de celles de l'Expérience 7 (voir (Nahorna et al, 2013)).

Dans la section suivante nous allons entreprendre d'interpréter les phénomènes observés en relation avec la littérature et dans la construction progressive d'un modèle cognitif qui permettrait de fournir un cadre cohérent à nos interprétations.

Tableau 36 – Les résultats d'analyse de temps de réponse à travers les expériences.

Expérience	Facteur	Significativité	Moyenne « Ba », en ms	Moyenne « McGurk », en ms	Ecart, en ms
4. Validation de l'effet contexte	Contexte : *coh, incoh*	*NS*			
	Cible : *Ba, McGurk*	*S*, en sujet fixe	550.5	577.5	27
6. Dynamique temporelle	Contexte : *coh, incoh, phonétique*	*NS*			
	Cible: *Ba, McGurk*	*S*	597.3	676.7	79.3
7. Reliage	Reset: *coh, fix*	*NS*			
	Durée du reset: *1, 2, 3 syllabes*	*NS*			
	Cible : *Ba , McGurk*	*S*	599	676.5	77.5

13.2 Interprétation des résultats

13.2.1 Mise en évidence d'un mécanisme de liage qui module la fusion audiovisuelle

L'ensemble de nos travaux montre une influence du contexte sur la fusion audiovisuelle, mesurée par l'effet McGurk. L'introduction d'une incohérence dans le contexte préalable provoque une diminution du rôle de l'information visuelle dans la parole multimodale. Ainsi nous observons une modulation de l'effet McGurk qui dépend du contexte préalable.

Il est important de rappeler que l'effet McGurk est plutôt résistant aux nombreuses incongruences, telles que celles qui concernent la localisation spatiale (Bertelson et al, 1994), l'asynchronie temporelle jusqu'à 200 ms (McGrath & Summerfield, 1985), le montage d'un visage de femme avec une voix de homme (Green et al, 1991), etc. Par contre, il existe des indices dans la littérature que l'effet McGurk peut être réduit par certains types d'incohérence phonétique tels que le conflit vocalique entre les flux auditif et visuel (Munhall et al, 1996) ou par l'incohérence de la vitesse ou des styles d'articulation entre le signal auditif et le signal visuel (Munhall et al, 1996), (Tanaka et al, 2009) et autres. Ainsi l'effet McGurk peut-être jusqu'à un certain point modulé (quoique assez faiblement) par ces propriétés internes. Dans ce manuscrit nous avons dédié un chapitre entier aux propriétés de l'effet McGurk, démontrées dans la littérature. Pour plus d'information nous renvoyons le lecteur au paragraphe 1.3. Dans nos expériences il est important de rappeler que nous avons contrôlé les propriétés des stimuli (voir Partie II) pour assurer que les effets de diminution contextuelle viennent bien du contexte lui-même et pas de différences de stimuli cibles.

Nous interprétons nos résultats dans le cadre d'un modèle à deux étages, qui a été introduit dans la littérature depuis longtemps dans le domaine de la perception des scènes auditives. Dans ce modèle, le premier étage a pour fonction de grouper ensemble les

composantes auditives d'une source, avant de passer à une étape de catégorisation (Bregman, 1990). Notre proposition consiste à élargir ce modèle vers le traitement des scènes audiovisuelles. Ainsi nous proposons d'introduire l'existence d'un processus de liage préalable à la fusion audiovisuelle.

Le liage de la parole audiovisuelle nécessite une association correcte des flux auditifs et visuels, comme c'est le cas dans l'effet cocktail party. Nous savons que les enfants de 4 mois sont capables de lier correctement la parole avec le visage (Kuhl & Meltzoff, 1982), (Kuhl & Meltzoff, 1984). Lorsque deux visages sont présentés à des sujets sur un écran, l'attention visuelle spatiale est capable de choisir entre les visages celui qui correspond à la lecture labiale cohérente avec le son (Andersen et al, 2009). La présentation de visages interférents modifie la réponse électrophysiologique à un visage dans les aires visuelles (Senkowski et al, 2008). Des effets de liage-déliage audiovisuel apparaissent également dans des tâches de lecture labiale, où l'interférence d'un stimulus auditif incohérent avec l'image ne produit d'effet significatif que dans le cas d'un démarrage synchrone des flux auditif et visuels et sous certaines conditions de cohérence phonétique (Brungart & Simpson, 2005). La modalité visuelle intervient également dans les effets de segmentation et multistabilité (Sato et al, 2007) et plus globalement d'analyse audiovisuelle de scènes de parole (Basirat et al, 2012). Tout ceci est cohérent avec notre hypothèse d'un processus de liage en charge d'évaluer la cohérence entre les flux auditifs et visuel, et produisant, dans le cas d'incohérence suffisante, un décrochage de l'effet McGurk.

Notre proposition d'introduire un processus préalable à la fusion modifie la vision classique sur la fusion audiovisuelle de la parole, considérée être automatique ou, lorsque cette vision est modifiée (voir section 1.3), faisant intervenir une surcharge d'un mécanisme de contrôle attentionnel général, mais non un processus de calcul de cohérence lié à l'information incidente. Nous proposons donc de modifier le modèle classique vers un modèle à deux étages. Dans les paragraphes suivants nous allons discuter le modèle cognitif proposé et le fonctionnement possible du processus de liage.

13.2.2 Architecture à deux étages

Revenons vers les modèles de fusion audiovisuelle existant dans la littérature (voir §2.1). Les modèles à un étage considèrent que la décision phonétique apparaît automatiquement et produit un percept intégré qui combine les signaux auditif et visuel sous l'influence possible de mécanismes attentionnels généraux (Figure 104).

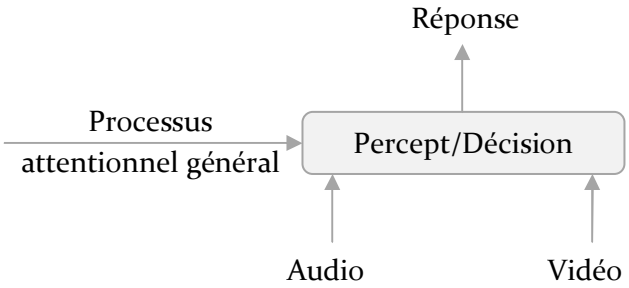


Figure 104 – Modèle à un étage de la fusion audiovisuelle dans la perception de la parole

Les résultats présentés dans le paragraphe précédent impliquent selon nous qu'un étage supplémentaire est nécessaire avant la décision (D) (Berthommier, 2004). Cet étage mettrait en jeu l'évaluation en ligne de la cohérence (C) des inputs auditifs et visuels (Figure 105). La cohérence de deux sources permettrait aux sujets de mieux traiter les deux flux et d'extraire une information adéquate. L'impact de l'information visuelle est bien mentionné dans la littérature dans le cas de la détection (Kim & Davis, 2004), (Grant & Seitz, 2000), et de la compréhension de la parole AV (Schwartz et al, 2004). Son influence est notable dans les situations bruitées (Sumby & Pollack, 1954).

Les sujets sont capables de percevoir et estimer une divergence entre l'image et le son et néanmoins de fusionner ces deux inputs dans un même percept (Manuel et al, 1983), (Summerfield & McGrath, 1984), (Soto-Faraco & Alsius, 2007), (Soto-Faraco & Alsius, 2009). Ceci suggère un accès conscient vers la sortie de la boîte de calcul de cohérence, notée C sur la figure.

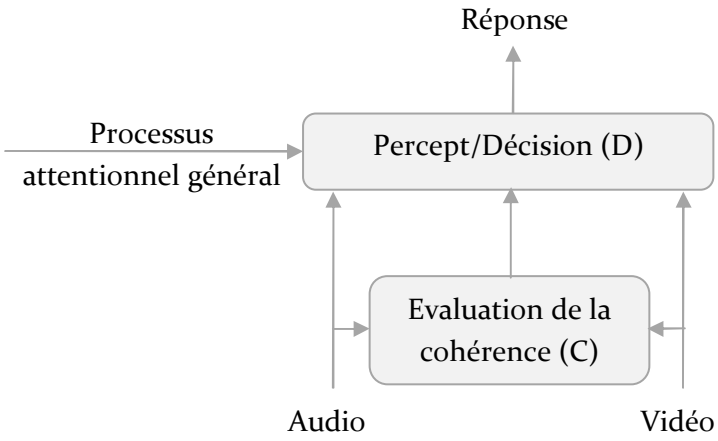


Figure 105 – Modèle à deux étages de la fusion audiovisuelle dans la perception de la parole

L'évaluation de la cohérence peut produire une diminution du poids du flux visuel dans le processus de décision lorsque les flux auditifs et visuels sont incohérents comme c'est le cas dans l'Expérience 4.

Dans ce modèle à deux étages le système de décision fusionne les flux auditif et visuel de façon conditionnée en prenant en compte les résultats du système de liage, que nous proposons dans ce travail. De nombreux travaux sur la fusion multisensorielle proposent que les données comportementales s'accordent de façon optimale avec un modèle Bayésien d'intégration des signaux (dans le cas de la fusion visuo-haptique (Ernst & Banks, 2002), dans le cas de la fusion audiovisuelle pour la localisation (Alais & Burr, 2004), dans le cas de la fusion visuo-vestibulaire de la perception de direction (Angelaki et al, 2011)). L'intégration statistique optimale des signaux est basée sur un processus probabiliste, qui les fusionne en pondérant par des facteurs inversement proportionnels aux variances de chaque modalité. Ainsi le signal avec une variance faible est plus fiable, dont il a plus de poids. Ces modèles sont optimaux dans le sens où ils produisent une évaluation multisensorielle avec la variance minimale. Le modèle FLMP (Fuzzy-Logical Model of Perception) est un exemple de ce type de modèles (Massaro, 1987), (Massaro, 1989) (§2.1.2). Il propose de calculer un produit de probabilités unisensorielles et par conséquent implicitement il intègre un mécanisme de diminution du rôle des inputs ambigus.

Selon nos données, pour le même input sensoriel (les mêmes cibles McGurk), le résultat peut varier selon le contexte. L'explication possible est que le processus de décision contient une composante de pondération contrôlée par la sortie du processus de liage, qui suppose que la composante auditive fournit la base du processus de décision, et que la composante vidéo se voit attribuée un poids moins élevé dans la fusion si l'incohérence est élevée.

Plusieurs modèles de pondération de la fusion ont été proposés dans la littérature, faisant dépendre les poids du bruit (Teissier et al, 1999), (Berthommier, 2001), du sujet (Schwartz, 2010), ou de l'attention (Schwartz et al, 2010). La sortie du processus de liage pourrait ainsi être directement intégrée dans le processus de décision dans le cadre Bayésien. Dans ce cas la décision dépendrait à la fois des deux flux individuels et des données en faveur de la cohérence ou de l'incohérence (Yu et al, 2009), (Noppeney et al, 2010).

13.2.2.1 Relations entre les étages d'évaluation et de décision

Comme nous l'avons vu précédemment, le bloc C peut intervenir directement dans le processus de décision. Nous avons montré que des contextes totalement incohérents influencent la perception d'effet McGurk. (Expérience 4). Dans les Expériences 5 et 6, nous avons montré que l'incohérence phonétique pure suffit à moduler la fusion. Or le contenu d'une telle incohérence semble impliquer l'accès à l'étage de décision pour estimer la cohérence. Pour cette raison, nous proposons d'introduire dans notre modèle un retour de la boîte de décision vers la boîte de calcul de la cohérence (Figure 106).

Une autre observation nous renforce dans cette proposition. Dans l'Expérience 6 (Figure 88), il est apparu que le pattern temporel du déliage dans le contexte phonétique est différent de celui du contexte incohérent : le déliage phonétique n'apparaît maximum qu'à partir de la deuxième syllabe, tandis que le déliage est maximal dès la première syllabe pour le contexte incohérent. Ceci suggère qu'un retour de l'étage de décision après la première syllabe est nécessaire pour que le déliage soit pris en compte. Certes, ces différences de dynamique, faibles, ne sont pas significatives. La question est néanmoins posée de différences de dynamiques temporelles entre les différents effets de déliage, et elle pourrait motiver de nouvelles expériences, nous y reviendrons.

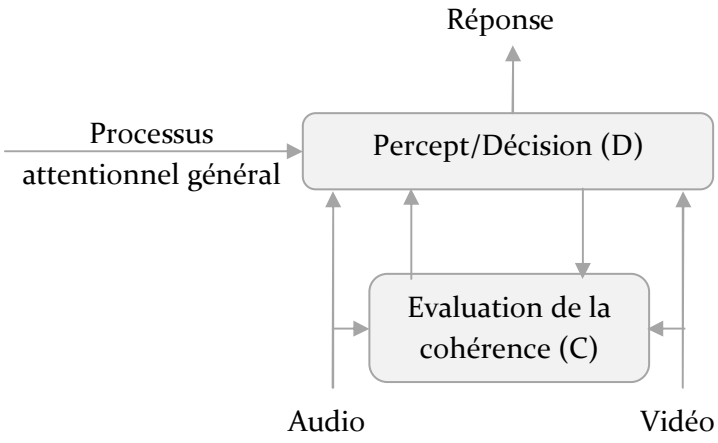


Figure 106 – Relations entre (C) et (D) dans le modèle à deux étages de la fusion audiovisuelle dans la perception de la parole

13.2.2.2 Evaluation de la cohérence entre flux auditif et visuel

Nous allons maintenant tenter de mieux caractériser le fonctionnement du bloc C d'« évaluation de la cohérence ».

Ce processus peut procéder en accord avec les principes gestaltistes de "destin commun" qui décrivent un des mécanismes possibles d'analyse des scènes perceptives. Ce principe se base sur une co-évolution des propriétés de l'input pour grouper les composantes correspondantes dans le processus de liage (Bregman, 1990). Dans notre cas la co-évolution pourrait logiquement s'appuyer sur la co-modulation audiovisuelle correspondant à la corrélation partielle dans le temps entre signal acoustique (typiquement une énergie instantanée, une enveloppe globale, ou une enveloppe dans des bandes spectrales spécifiques) et vidéo (typiquement les paramètres des lèvres ou visage). Plusieurs articles ont mis en évidence ce type de corrélation (Munhall, 1998), (Yehia et al, 1998), (Barker & Berthommier, 1999), (Jiang et al, 2002), (Chandrasekaran et al, 2009) et le lien avec la détection de la parole audiovisuelle (Grant & Seitz, 2000), (Kim & Davis, 2004).

Par ailleurs, nous avons proposé de décomposer l'incohérence totale selon les dimensions phonétique et temporelle, ce qui contribue à la description du calcul de la cohérence. L'effet de l'incohérence temporelle dans l'Expérience 5 est faible mais l'incohérence elle-même est réduite dans cette expérience, aussi ce résultat n'est pas surprenant. Par contre l'effet net de l'incohérence phonétique est plus surprenant. En effet, dans le cas d'une incohérence phonétique la comodulation temporelle entre flux auditif et visuel est largement respectée, si ce n'est bien sûr en termes de contenu phonétique fin. Ceci suggère que la comodulation temporelle ne suffit pour assurer un liage parfait. L'effet modulateur du contexte phonétiquement incohérent suggère, nous l'avons dit, que le contenu phonétique de chaque flux est déterminé et exploité dans le processus de liage. Dans le terme de l'architecture cognitive, ceci signifie que le processus de liage audiovisuel reçoit l'information de la caractérisation phonétique auditive et visuelle, comme indiqué dans le schéma (Figure 106). Ceci peut-être compatible notamment avec l'architecture dite à « identification séparée », ou processus de fusion tardive (Schwartz et al, 1998), où l'indentification séparée des flux auditif et visuel précède la fusion audiovisuelle. Le modèle FLMP serait donc compatible avec cette architecture, pourvu qu'on lui adjoigne un mécanisme de pondération de la fusion en fonction du contexte et de la cohérence qu'il implique.

13.2.2.3 L'état par défaut

La question suivante que nous nous posons est celle de l'état par défaut, dans laquelle on suppose que le système se trouve avant toute sorte d'influence contextuelle. Bien sûr nous pouvons argumenter que ce type de situation n'arrive jamais, qu'il n'y a pas d'état par défaut et que donc le système reste toujours dans l'état qui est un résultat de l'histoire des influences récents. Comme la cohérence audiovisuelle est la situation le plus probable, l'état présent généralement doit être considéré comme correspondant à la situation de la cohérence : ainsi ce serait l'état lié. On peut aussi supposer qu'il n'existe pas d'état par défaut mais plutôt (en termes bayésiens) un « prior » où les flux auditif et visuel sont supposés cohérents et ainsi

doivent être liés ensemble. Cette hypothèse d'un état par défaut (ou d'un prior) lié est compatible avec le fait que le paradigme classique de l'effet McGurk doit conduire vers la fusion, ce qui est effectivement en général le cas. Ceci est aussi compatible avec le « biais de compatibilité » décrit dans diverses expériences portant sur la fusion d'indices contradictoires, mono ou multisensoriels, et qui montrent que les sujets partent toujours d'une hypothèse initiale de non conflit, donc de cohérence des entrées, avant de réviser peu à peu leur jugement (Yu et al, 2009), (Noppeney et al, 2010). Les auteurs parlent ainsi de « biais de compatibilité » (là encore, au sens d'un « prior » perceptif) avant l'évidence d'un conflit qui conduit progressivement les sujets à choisir une entrée sensorielle plutôt que l'autre.

Enfin, ce biais initial vers un percept cohérent est aussi une donnée forte dans les théories gestaltistes du groupement perceptif comme celles de Bregman, nous l'avons vu. Ainsi, dans leurs travaux sur les effets de multistabilité auditive ou visuelle, Hupé et Pressnitzer (Hupé & Pressnitzer, 2012) montrent bien que l'hypothèse initiale des sujets est toujours une hypothèse de cohérence, que ce soit pour des effets de multistabilité auditive (« effet Van Noorden ») ou visuels (plaids dynamiques).

Des fluctuations pourraient apparaître autour de cet état par défaut même sans contexte incohérent récent. On peut ainsi se demander si les variations interindividuelles de l'effet McGurk (Schwartz, 2010) pourraient être associées à de telles fluctuations, et si les sujets qui ne présentent pas beaucoup d'effet McGurk sont au moment de l'expérience dans un état délié ou imparfaitement lié. Cependant, ceci est peu probable, en prenant en compte que dans nos données plusieurs sujets ne perçoivent pas d'effet McGurk même après de longues périodes de cohérence. Enfin notons que l'hypothèse d'un état par défaut lié ne signifie pas nécessairement que les sujets perçoivent « da ». En effet le percept est un résultat de la fusion des indices sensoriels disponibles, auditifs et visuels, qui peuvent conduire vers toutes sortes de réponses différentes, telles que « ba », « da », « ga », « tha », « bda », « bga » etc., et il n'y a pas de raison pour que le « da » soit systématiquement le vainqueur de cette compétition perceptive. Ceci dépend en réalité de nombreux facteurs, qui incluent la nature du système phonologique des sujets, les variations culturelles, etc. (Schwartz, 2010).

13.2.2.4 Dynamique d'évaluation du calcul de la cohérence

Nous voyons le liage comme un processus dynamique, qui évalue l'état de cohérence des deux flux à chaque instant. Notre Expérience 6 montre que lorsque le niveau d'incohérence est élevé, le décrochage est rapide, la durée d'une syllabe d'incohérence suffisant pour un déliage maximal. Les données sont semblables quoique moins claires dans le cas d'incohérence phonétique.

Nous avons également vu dans l'Expérience 6 qu'à partir de la troisième syllabe nous observons une diminution de déliage dans le cas d'incohérence forte (Figure 101). Notre hypothèse est que l'incohérence totale provoque une rupture rapide de l'état lié, avec déliage maximal, puis le système se stabilise vers un niveau de déliage stable. Ce phénomène correspondrait à un effet « d'adaptation au déliage ».

Si le déliage est rapide, notre Expérience 7 nous a montré qu'il faut en revanche accumuler un minimum d'évidences pour pouvoir relier les deux modalités après déliage (Figure 103). Ainsi le bloc C doit accumuler des éléments de présence ou absence de cohérence selon une dynamique complexe, puis transmettre cette information vers l'étape de décision.

Nous avons vu également que qu'une pause (silence et image fixe) ne suffit pas pour restaurer le liage. Ceci est une fois encore cohérent avec les connaissances sur les mécanismes d'analyse de scènes auditives, pour lesquels un changement soudain dans les propriétés acoustiques du signal peut réinitialiser le mécanisme d'allocation des sources plus rapidement qu'un simple silence. Le système traite un changement comme un indice qui manifeste un nouvel événement sonore, et ainsi probablement réinitialise l'état du système (Bregman, 1990).

Nous pouvons ainsi décrire un processus de liage comme un processus calculatoire en ligne, qui évalue la cohérence de deux flux à chaque instant et cumule l'information calculée. L'existence de l'effet McGurk nous montre qu'une mise en évidence instantanée de l'incohérence ne suffit pas à décrocher les flux auditifs et visuels (sauf, on l'a vu, les situations où l'incohérence phonétique consonantique caractéristique de l'effet McGurk était cumulée à un conflit vocalique entre les flux auditif et visuel (Munhall et al, 1996) ou à une incohérence de vitesse et/ou d'articulation entre le signal auditif et visuel (Munhall et al, 1996), (Tanaka et al, 2009) et autres) et qu'il faut adjoindre au minimum un contexte d'une syllabe incohérente pour produire un décrochage. Ensuite la dynamique évolue en fonction du signal présenté (Expériences 5, 6 et 7). Nous proposons de généraliser ce processus sous la formule :

$$Cohérence = C_0 + \int_0^t C(\tau)d\tau$$

où C_0 est la cohérence initiale, et $C(\tau)$, la cohérence instantanée à l'instant τ.

Ceci nous conduit à modifier notre modèle, en séparant la boîte C en deux composantes, l'une d'évaluation instantanée, qui estime la cohérence audiovisuelle à l'instant présent, notamment dans le cas de l'effet McGurk isolé sans contexte, et l'autre d'inférence d'un état global, qui est un résultat des évaluations cumulées.

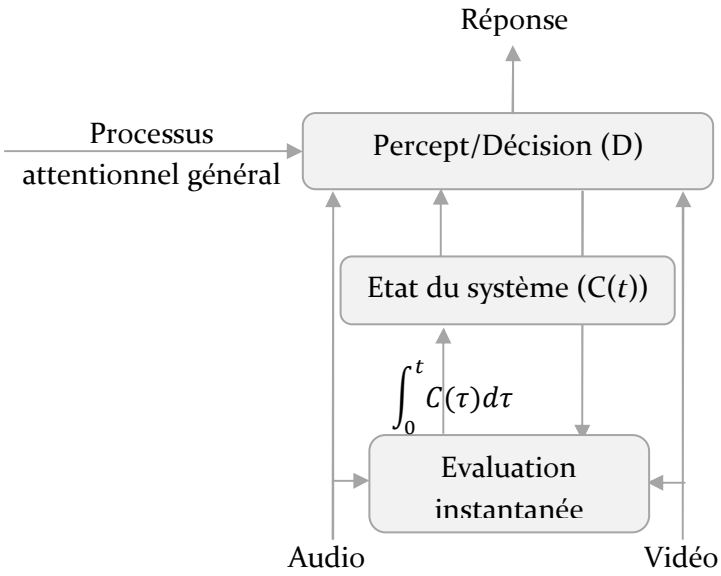


Figure 107 – Calcul de la cohérence dans le modèle à deux étages de la fusion audiovisuelle dans la perception de la parole

13.2.2.5 Temps de réponse

On sait classiquement que le temps de réaction est plus court pour traiter les stimuli congruents que pour les stimuli incongruents dans différentes sortes de tâches (Gondan et al, 2005), y compris les stimuli McGurk. L'interprétation est généralement que les stimuli sont ambigus, conduisant à un temps de traitement plus élevé à cause de la difficulté à prendre une décision (Massaro & Cohen, 1983). Le résultat de nos expériences est différent : certes les réponses sur les cibles Ba sont systématiquement plus rapides que sur des cibles McGurk (Tableau 36), mais il n'apparaît pas d'influence des autres facteurs tels que contexte, reset, durée du contexte ou du reset, qui pourtant produisent tous un effet significatif et en général quantitativement important sur le taux de perception d'effet McGurk.

Une interprétation possible est que la réponse des sujets aux cible McGurk est ralentie par la prise d'information instantanée sur l'existence d'un conflit entre les entrées sensorielles, mais qu'elle ne prend pas en compte la globalité de l'information de cohérence sur l'état du système, ou en tout cas pas suffisamment pour que cela produise des fluctuations visibles. La proposition que l'on pourrait faire est donc qu'il y a un lien direct entre la sortie du processus d'évaluation instantanée et le temps de réponse : l'allongement significatif et régulier des temps de réponse des cibles McGurk, indépendamment des effets globaux de contexte (qu'il soit « déliant » comme dans les Expériences 4 et 6 ou « reliant » comme dans l'Expérience 7) serait la trace d'un effet de ce bloc de calcul instantané C(t) sur la prise de décision du sujet.

Pour prendre en compte la donnée classique, rappelée précédemment, qu'une situation incertaine demande plus de temps pour prendre une décision, ce qui est également conforme au ralentissement associé aux stimuli McGurk, nous proposons que les temps de réponse dépendent également du processus de décision. En résumé, deux flèches conduisent dans la Figure qui suit vers l'établissement du temps de réponse : la sortie du processus d'évaluation instantanée, détectant un conflit immédiat, et la décision globale du système, ralentissant la réponse en cas de stimulus ambigu.

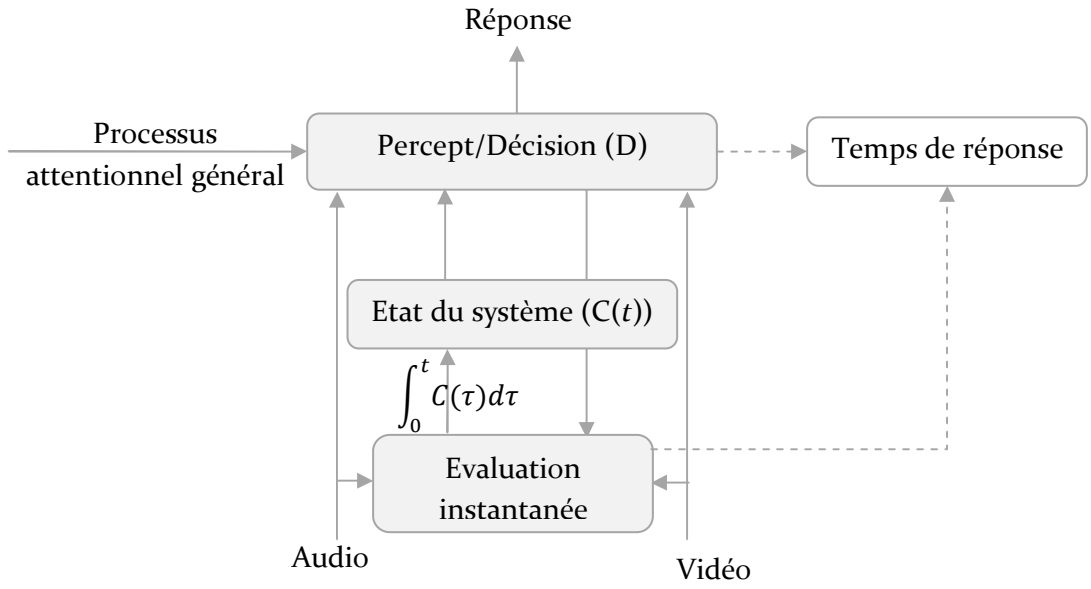


Figure 108 – Détermination des temps de réponse dans le modèle à deux étages de la fusion audiovisuelle dans la perception de la parole

13.3 Corrélats neuroanatomiques et neurophysiologiques

Nous allons maintenant discuter les possible corrélats neuroanatomiques et neurophysiolgiques d'un système de liage audiovisuel. Ce système doit permettre le calcul de la caractérisation de la cohérence audiovisuelle et ainsi son activité neuronale doit être d'une manière ou d'une autre liée à la cohérence entre les flux auditif et visuel. Par ailleurs il doit fournir un dispositif qui permette à une entrée visuelle de moduler l'activité dans les régions auditives, fournissant ainsi la base d'effets tels que l'amélioration de détection de la parole audiovisuelle (Grant & Seitz, 2000), (Kim & Davis, 2003), (Kim & Davis, 2004) et les interactions électrophysiologiques précoces dans le cortex auditif (van Wassenhove et al, 2005). Les aires mentionnées dans cette discussion sont représentées dans la Figure 109.

Il apparaît de plus en plus clairement que les influences inter-modales sont susceptibles d'intervenir dès le niveau du cortex sensoriel primaire qui était auparavant supposé être spécifique à une modalité sensorielle (Driver & Noesselt, 2008). Des influences cross-modales en perception de la parole audiovisuelle ont été mises en évidence par neuroimagerie fonctionnelle (fMRI) dans les cortex primaires auditif et visuel (Calvert et al, 1997), (Calvert et al, 1999) et nous savons que ces influences peuvent apparaître assez tôt dans les processus perceptifs (Besle et al, 2004), (Colin et al, 2002). S'il est possible d'envisager le rôle de relais sous-corticaux (colliculus supérieur ou relais thalamiques) et des liens horizontaux (qui relient directement les contextes sensoriels), le cortex associatif hétéromodal dans le sillon temporal supérieur (STS) est souvent mentionné comme un candidat majeur dans ce processus (Calvert et al, 2000), (Ghazanfar & Schroeder, 2006). Le fait que le processus de liage repose au moins pour partie sur des processus phonétiques spécifiques (comme démontré par (Bernstein et al, 2004), (Schwartz et al, 2004) et le rôle des incohérences phonétiques dans nos propres résultats de l'Expérience 5), discréditent largement les processus sous-corticaux comme le site unique ou majeur des mécanismes de liage de la parole audiovisuelle. Le rôle de STS (plus précisément de la partie postérieure de STS, pSTS) est considéré souvent comme crucial, en particulier pour le traitement des corrélations entre les stimuli auditifs et visuels, qui est très certainement un ingrédient de base dans le liage de la parole audiovisuelle (Campbell, 2008).

Ainsi, dans une étude combinant électrophysiologie (MEG) et neuroimagerie (fMRI), **(Arnal et al, 2009)** suggèrent l'existence de deux routes séparées associant les cortex auditif et visuel. Dans l'architecture qu'ils proposent, la voie rapide cortico-corticale, qui ne serait pas sensible à l'incongruence audiovisuelle, relierait directement les paramètres de mouvement visuel au cortex auditif et permettrait de faire des prédictions et de moduler l'activité auditive à court terme. Un lien plus lent conduirait vers le STS qui serait le centre de l'estimation du degré d'incohérence entre les entrées auditive et visuelle. Le message serait envoyé en retour de STS aux cortex auditif et visuel. Il faut noter que, dans la proposition des auteurs, la voie plus lente, estimant le niveau d'incohérence audiovisuelle, modifierait la réponse neuronale au fil du temps. Ceci pourrait fournir un corrélat de l'augmentation de temps de réponse pour les cibles McGurk par rapport aux cibles congruentes « Ba » dans nos expériences. Enfin, dans une étude électrophysiologique récente (MEG) **(Keil et al, 2012)** ont tenté de relier le rôle des fluctuations temporelles de l'activité cérébrale avec des variabilités de l'effet McGurk. Ils montrent une corrélation entre le taux de fusion perceptive et l'état de connexion entre le gyrus temporal supérieur gauche et le réseau distribué des régions frontales et temporales. Ceci fournit une mise en évidence intéressante que l'effet McGurk est un processus dynamique

relié à l'état d'aires corticales spécifiques (proposées être localisées dans le cortex temporal dans cette étude).

Plus précisément le rôle majeur de l'intégration intermodale est souvent attribué, nous l'avons dit, à la partie postérieure du sillon temporal supérieur (pSTS). Dans un paradigme de stimulation magnétique transcranienne (TMS) (Beauchamp et al, 2010) ont réussi à diminuer l'effet McGurk en stimulant cette zone. Enfin, dans une étude explorant l'effet McGurk à l'aide de l'imagerie par résonance magnétique fonctionnelle (fMRI), (Nath & Beauchamp, 2011) ont montré une corrélation entre l'activité de la région STS gauche et l'importance de l'effet McGurk, les sujets fournissant une réponse de type fusion présentant une activité plus élevée dans cette région.

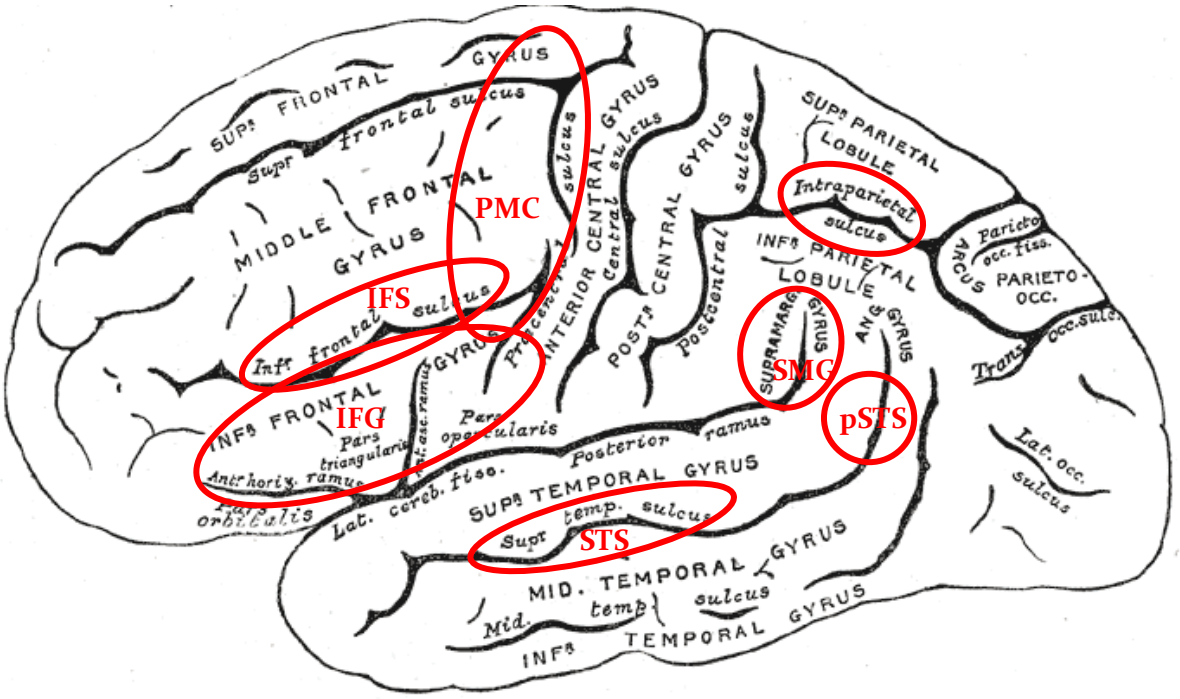


Figure 109 – Localisation neuroanatomique des principales régions impliquées potentiellement dans les processus de liage audiovisuel en perception de parole

En remontant vers le cortex pariétal, le gyrus supra-marginal (Supra-Marginal Gyrus, SMG) a été proposé par (Bernstein et al, 2008) comme un lieu possible pour l'analyse des incongruences audiovisuelles. Dans leur étude, les auteurs ont présenté des stimuli de parole audiovisuelle avec des niveaux variables d'incongruence entre les flux auditif et visuel, et les données IRMf ont montré que seule une région corticale démontrait une sensitivité différenciée selon le niveau d'incongruence, cette région appartenant à SMG.

En remontant encore et en allant vers les régions antérieures du cortex, la « voie dorsale » qui relie les aires sensorielles et motrices (Hickok & Poeppel, 2004) semble également impliquée, à la fois dans les processus d'organisation de la parole auditive (Sato et al, 2004), (Kondo & Kashino, 2007) et dans la perception de la parole audiovisuelle (sillon intra-pariétal et gyrus inférieur frontal (Miller & D'Esposito, 2005); planum temporale postérieur Spt ; cortex

prémoteur dorsal et ventral (Okada & Hickok, 2009), en particulier pour le liage des stimuli incohérents (Jones & Callan, 2003). La voie dorsale est proposée par (Campbell, 2008) comme fournissant le lieu naturel pour traiter des propriétés de corrélation de la parole audiovisuelle. La théorie de la perception pour le contrôle de l'action (PACT, (Schwartz et al, 2012)) considère l'hypothèse que la voie dorsale joue un rôle dans le liage de la parole audiovisuelle, ce qui est également compatible avec un rôle prédictif associé à ce circuit par (Kilner et al, 2007), et ses applications au traitement de la parole audiovisuelle (Skipper et al, 2007). Ce cadre serait sans doute compatible avec le résultat de Colin et al. (Colin et al, 2002) montrant l'existence d'une réponse EEG de type MMN (MisMatch Negativity) à des stimuli McGurk sur l'électrode Fz correspondant à une région frontale.

Enfin, il est intéressant de noter l'étude de (Noppeney et al, 2010) concernant le traitement de l'information audiovisuelle incongruente sur des stimuli non langagiers (son et image d'actions sur des outils ou des instruments de musique). Dans cet article, qui combine approche psychophysique et données de neuroimagerie IRMf, les auteurs étudient la façon dont les indices d'incongruence sont accumulés au cours du temps, conduisant les sujets à s'écarter progressivement d'un état initial influencé par le biais de compatibilité mentionné précédemment vers un état délié dans lequel ils rejettent l'information incongruente et ainsi considérée par eux comme non pertinente pour la tâche perceptive. Ce paradigme est naturellement proche de notre hypothèse de l'existence d'un système de liage qui évalue la cohérence entre les flux auditif et visuel et module en conséquence le processus de décision (Figure 107). Selon les données fMRI dans cette étude le sillon frontal inférieur (Inferior Frontal Sulcus IFS) présente un profil d'accumulateur audiovisuel cohérent avec le pattern de temps de réaction observé dans l'étude. Par ailleurs les données montrent une inhibition par IFS des activations temporales supérieures dans le cortex auditif pour un input auditif incongruent et considéré comme non pertinent. Les auteurs concluent : « to form decisions that guide behavioral responses, the IFS may accumulate audiovisual evidence by dynamically weighting its connectivity to auditory and visual regions according to sensory reliability and decisional relevance. » Même si la tâche n'est pas exactement la même, la correspondance avec l'architecture que nous avons proposée dans la Figure 107 est assez claire.

En conclusion, on peut proposer que les liens direct cross-modaux entre les cortex sensoriels, une intervention de la région STS (et notamment de sa partie postérieure pSTS) par feedback vers les régions primaires, et une modulation attentionnelle pariéto-frontale associée avec les processus perceptuo-moteurs, pourraient conjointement jouer un rôle dans la fusion (Senkowski et al, 2008).

13.4 Perspectives

13.4.1 Perspectives expérimentales

Dans ce travail nous avons proposé un nouveau paradigme expérimental pour mettre en évidence et explorer la nature d'un système de liage de la parole audiovisuelle. Dans les expériences présentées nous avons commencé à caractériser ce processus. Bien évidemment ce travail d'exploration ne fait que commencer et il nous manque de nombreuses briques pour proposer une théorie complète. Nos perspectives les plus proches consistent donc à tenter de continuer la caractérisation pour mieux comprendre comment fonctionne ce système.

Une première question qui nous semble importante est de savoir si des *paramètres non-phonétiques* concernant par exemple la localisation spatiale, l'identité d'un locuteur, son genre, etc. peuvent jouer un rôle dans le processus du liage. Tandis que ces facteurs semblent jouer peu de rôle dans l'effet McGurk proprement dit (Green et al, 1991), (Bertelson et al, 1994), il serait intéressant de déterminer si des séquences contextuelles présentant une divergence spatiale ou une différence de sexe entre le son et l'image peuvent moduler l'effet McGurk de façon significative. Plus généralement on peut se demander si le liage audiovisuel de la parole est constitué exclusivement de processus spécifiques à la parole ou s'il n'est qu'une composante d'un système général *d'analyse de scènes multisensorielles*. C'est dans ce cadre qu'il conviendra de poursuivre la démarche de caractérisation des types d'incohérence susceptibles de produire un déliage et de tenter de décomposer l'incohérence maximale selon des composantes élémentaires.

La *décomposition de l'incohérence en composantes élémentaires* nous a ainsi conduit au cours de cette thèse à proposer deux dimensions essentielles, l'une portée par les comodulations de la dynamique globale des signaux (associée à l'incohérence temporelle dans l'Expérience 5) et l'autre par le contenu phonétique (Expériences 5 et 6). En ce qui concerne le déliage provoqué par l'incohérence temporelle, qui devrait être une composante essentielle, nous nous sommes limités à des décalages très faibles, restant à l'intérieur de la fenêtre d'intégration audiovisuelle pour une syllabe (van Wassenhove et al, 2007). Ainsi, il conviendrait d'étudier quels seraient les résultats en dehors de la fenêtre de l'intégration audiovisuelle, quelle sera la dynamique de liage contextuelle dans ce cas, et à partir de quel moment, en augmentant le décalage temporel, nous devrions obtenir une rupture quasi totale de la fusion du signal visuel comme c'est le cas probablement dans les films doublés.

La question de la dynamique temporelle de liage conduit à une autre question, qui n'a pas été traitée dans ce manuscrit, concernant le rôle potentiel de la *rythmicité*. On sait que la rythmicité d'une source d'information perceptive permet de détecter et d'utiliser les relations entre les événements pour structurer l'environnement dynamique. Les interactions rythmiques entre la structure du temps dans le traitement humain et la structure de l'environnement en mouvement provoquent des mécanismes de synchronisation et d'attente temporelle. Ainsi lors d'événements très cohérents temporellement une forte anticipation comportementale apparaît et conditionne les capacités de traitement de l'information (Jones, 1976), (Jones & Boltz, 1989). Le rôle prédictif du rythme a été également démontré dans les interactions audiovisuelles (en dehors de la parole) (Kösem & van Wassenhove, 2012). Or, tous les stimuli de nos expériences comportent de fortes caractéristiques rythmiques, avec une cadence assez régulière de production des syllabes dans les flux auditif ou visuel selon les cas. Ainsi nous pouvons nous demander si les effets de liage obtenus proviennent purement d'une rythmicité des stimuli, induisant une forte *prédictibilité* sur les cibles (« Ba » ou McGurk). Cependant nous observons que le fait de rajouter une pause entre le contexte et la cible (reset fixe, Expérience 7) maintient le même niveau de déliage, alors que pourtant la prédictibilité temporelle de la cible est ici éliminée. Ceci montre que les effets de modulation de l'effet McGurk que nous avons obtenu ne sont pas purement dus à des mécanismes de prédiction temporelle. Il reste que l'existence d'une cohérence rythmique dans nos stimuli contextuels pourrait faire partie des composantes qui contribuent à la cohérence entre les inputs auditifs et visuels. Pour étudier le rôle de la rythmicité il conviendra de mettre en place une nouvelle expérience avec une

désynchronisation aléatoire des syllabes du contexte, tout en conservant la cohérence audiovisuelle, afin d'obtenir des stimuli contextuels cohérents mais non rythmiques.

Une autre question, que nous avons déjà commencé à étudier, concerne la possibilité de fournir un réel mécanisme de *reset* permettant de quitter un état délié. Nos essais de reset au cours de l'Expérience 7 ne nous ont pas réellement permis de trouver un mécanisme de reset permettant de rebasculer immédiatement le sujet dans un état lié. L'alerte audiovisuelle ne produit aucun effet (Expérience 2), le reset cohérent produit une évolution progressive d'un système vers l'état lié et le reset fixe préserve l'état délié (expérience 7). Selon Bregman (Bregman, 1990) l'arrivée d'un nouvel événement produit un reset instantané et fait basculer un état cognitif. Nous pouvons donc imaginer une expérience où on change complètement la scène lors du contexte. Par exemple on peut introduire dans le reset et la cible un locuteur différent du contexte, déterminer la dynamique de reliage et la comparer avec les résultats de l'Expérience 7, afin de voir si l'on obtient un reliage plus rapide. On peut également se demander si dans ce cas un sujet peut parvenir à mémoriser plusieurs flux cognitifs en même temps ? Ainsi, si l'on introduit un contexte incohérent avec un locuteur 1, poursuivi par un reset cohérent avec un locuteur 2 et se terminant par une cible McGurk du locuteur 1, est-ce que le sujet groupe la cible avec son contexte incohérent ? Enfin, si le changement de locuteur ne suffit pas à faire basculer perceptivement le sujet dans un changement radical de scène permettant un réel reset, nous pourrions introduire une scène totalement différente (retransmission d'un événement sportif, clip musical, etc.) afin de déterminer comment le sujet peut réellement « remettre à zéro » son compteur perceptif.

Le fait de pouvoir « geler » l'état cognitif d'un sujet dans un mode « délié » nous ouvre des perspectives très intéressantes pour des expérimentations électrophysiologiques ou neuroanatomiques avec des techniques d'EEG et d'IRMf. Notre collègue Marc Sato nous a proposé de tenter d'exploiter notre paradigme expérimental pour tester l'effet du déliage sur l'influence du contexte visuel sur les potentiels évoqués auditifs N1-P2 (Figure 110). Rappelons que si l'on compare les potentiels évoqués fronto-centraux pour des stimulations de parole auditive vs. audiovisuelle, la composante visuelle produit une modulation de l'onde N1 (premier pic négatif vers 100 ms après le début du signal), impliquant à la fois une diminution d'amplitude et une petite anticipation temporelle (§3.1.1). L'hypothèse de Marc Sato est que le mécanisme de déliage pourrait réduire et faire disparaître cet effet. Nous avons ainsi mis en place une expérience préliminaire sur cette base (Figure 110), expérience hélas non achevée faute de temps dans le contexte de cette thèse, et qui est actuellement réalisée par un nouveau doctorant, en collaboration avec M. Sato et C. Vilain.

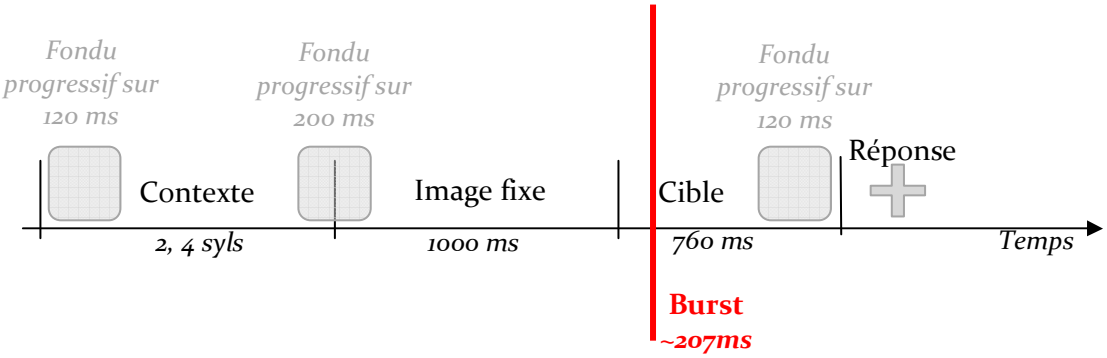


Figure 110 - Proposition d'un paradigme expérimental, qui vise à étudier un corrélat neurophysiologique possible du système de liage

13.4.2 Perspectives applicatives

Au-delà de ses enjeux expérimentaux et théoriques, ce travail pourrait à terme déboucher sur la mise en œuvre d'un modèle calculatoire qui pourrait trouver des *applications* par exemple dans des systèmes de *reconnaissance automatique de la parole multimodale*, notamment dans le cas de traitement automatique de contenus vidéos ou de l'interaction homme-machine. Dans des situations de dialogue impliquant plusieurs personnes dans une scène audiovisuelle complexe, l'implication d'un système de liage semble requise pour déterminer quelle est la personne source, comment se font les associations entre composantes auditives et visuelles dans une superposition de locuteurs, afin de pouvoir traiter l'information de manière efficace.

D'autre part les connaissances sur le système de liage pourraient conduire à des applications dans le domaine du *diagnostic* et de la *rééducation des troubles cognitifs* chez les patients malvoyants, malentendants, sourds implantés cochléaire etc., dans le but de mieux caractériser le liage, ses éventuelles perturbations et ses possibles améliorations. D'autres questions de recherche peuvent être ouvertes, visant notamment à savoir si la capacité de lier s'apprend dans les premiers temps de la vie, et se dégrade dans le vieillissement, et s'il existe des troubles spécifiques de liage.

L'obtention en 2012 d'une allocation de recherche de la Région Rhône-Alpes et le recrutement en thèse de Ganesh Attigodu Chandrashekara permettra sans doute de répondre pour partie à certaines de ces questions.

-

Liste des publications associées à cette thèse

Nahorna, O., Berthommier, F. & Schwartz, J.-L.,2012. Binding and unbinding the auditory and visual streams in the McGurk effect. *Journal of the Acoustical Society of America*, 132, p. 1061-1077.

Nahorna, O., Berthommier, F. & Schwartz, J.-L.,2012. Unbinding and rebinding the McGurk effect: Further experiments on audiovisual binding in speech perception. *En préparation.*

Chandrashekara, G.A., Berthommier, F., Nahorna, O. & Schwartz, J.-L., 2013. Effect of context, rebinding and noise on audiovisual speech fusion . In *Proceedings of the Interspeech 2013, Lyon, France.*

Nahorna, O., Chandrashekara, G.A., Berthommier, F. & Schwartz, J.-L., 2013. Modulating fusion in the McGurk effect by binding processes and contextual noise. In *Proceedings of the 11th International Conference on Auditory-Visual Speech Processing (AVSP 2013), St Jorioz, France.*

Nahorna, O., Berthommier, F. & Schwartz, J.-L., 2012. Dynamique temporelle du liage dans la fusion de la parole audiovisuelle. In *Actes des XXIXèmes Journées d'Étude sur la Parole (JEP2012), 2012,* Grenoble, France

Nahorna, O., Berthommier, F. & Schwartz, J.-L., 2011. Binding and unbinding the McGurk effect in audiovisual speech fusion: Follow-up experiments on a new paradigm. In *Proceedings of the 10th International Conference on Auditory-Visual Speech Processing (AVSP 2011), Volterra,Italy.*

Nahorna, O., Berthommier, F. & Schwartz, J.-L., 2011. Liage et fusion audiovisuelle en perception de la parole: on peut «débrancher» l'effet McGurk par un contexte audiovisuel incohérent. In *Rencontres Jeunes Chercheurs en Parole (RJCP2011), 2011, Grenoble, France.*

Nahorna, O., Berthommier, F. & Schwartz, J.-L., 2010. Binding and unbinding in audiovisual speech fusion: Removing the McGurk effect by an incoherent preceding audiovisual context. In *Proceedings of the 9th International Conference on Auditory-Visual Speech Processing (AVSP 2010), Hakone, Japon.*

Nahorna, O., Berthommier, F. & Schwartz, J.-L., 2012. Liage et fusion audiovisuelle en perception de la parole: on peut «débrancher» l'effet McGurk par un contexte audiovisuel incohérent. In *Actes des XXVIIIèmes Journées d'Étude sur la Parole (JEP2010), 2010, Mons, Belgique*

Travaux cités

Alais, D. & Burr, D., 2004. The ventriloquist effect results from near-optimal bimodal integration. *Current Biology*, 14(3), p. 257-262. Available at: http://dx.doi.org/10.1016/j.cub.2004.01.029.

Alsius, A., Navarra, J., Campbell, R. & Soto-Faraco, S., 2005. Audiovisual integration of speech falters under high attention demands. *Current Biology*, 15(9), p. 839-843. Available at: http://dx.doi.org/10.1016/j.cub.2005.03.046.

Alsius, A., Navarra, J. & Soto-Faraco, S., 2007. Attention to touch weakens audiovisual speech integration. *Experimental Brain Research*, 183(3), p. 399-404. Available at: http://dx.doi.org/10.1007/s00221-007-1110-1.

Andersen, T.S., Tiippana, K., Laarni, J., Kojo, I. & Sams, M., 2009. The role of visual spatial attention in audiovisual speech perception. *Speech Communication*, 51(2), p. 184-193. Available at: http://www.sciencedirect.com/science/article/pii/S016763930800126X.

Angelaki, D.E., Gu, Y. & Deangelis, G.C., 2011. Visual and vestibular cue integration for heading perception in extrastriate visual cortex. *The Journal of Physiology*, 589(Pt 4), p. 825-833. Available at: http://dx.doi.org/10.1113/jphysiol.2010.194720.

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(43), p. 13445-13453. Available at: http://dx.doi.org/10.1523/JNEUROSCI.3194-09.2009.

Arnold, P. & Hill, F., 2001. Bisensory augmentation: a speechreading advantage when speech is clearly audible and intact. *British Journal of Psychology*, 92(Pt 2), p. 339-355.

Attina, V., 2005. *La Langue française Parlée Complétée : Production et Perception*. Ph.D. dissertation.

Barker, J. & Berthommier, F., 1999. Evidence of correlation between acoustic and visual features of speech. In *Proceedings of the 14th International Congress of Phonetic Sciences (ICPhS'99)*. San Francisco, USA.

Basirat, A., Schwartz, J.-L. & Sato, M., 2012. Perceptuo-motor interactions in the perceptual organization of speech: evidence from the verbal transformation effect. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 367(1591), p. 965-976. Available at: http://dx.doi.org/10.1098/rstb.2011.0374.

Beauchamp, M.S., Nath, A.R. & Pasalar, S., 2010. fMRI-Guided transcranial magnetic stimulation reveals that the superior temporal sulcus is a cortical locus of the McGurk effect. *The Journal of Neuroscience*, 30(7), p. 2414-2417. Available at: http://dx.doi.org/10.1523/JNEUROSCI.4865-09.2010.

Beautemps, D., Borel, P. & Manolios, S., 1999. Hyper-Articulated Speech: Auditory and Visual Intelligibility. In *Sixth European Conference on Speech Communication and Technology (EUROSPEECH'99)*. Budapest, Hungary.

Benoit, C., Mohamadi, T. & Kandel, S., 1994. Effects of phonetic context on audio-visual intelligibility of French. *Journal of speech and hearing research*, 37(5), p. 1195-1203.

Bernstein, L.E., Auer, E.T.J. & Moore, J.K., 2004. Audiovisual Speech Binding: Convergence or Association?: Cambridge, MA, US: MIT Press, xvii, 915 pp. Ch. Audiovisual Speech Binding: Convergence or Association? The handbook of multisensory processes. p. 203-223.

Bernstein, L.E., Auer, E.T.J. & Takayanagi, S., 2004. Auditory speech detection in noise enhanced by lipreading. *Speech Communication*, 44(1), p. 5-18. Available at: http://www.sciencedirect.com/science/article/pii/S0167639304001165.

Bernstein, L.E., Burnham, D. & Schwartz, J.-L., 2002. Special session: issues in audiovisual spoken language processing (when, where, and how?). In *7th International Conference on Spoken Language Processing ICSLP-2002*. Denver, Colorado, USA Adelaide, S. Aust. Causal Productions.

Bernstein, L.E., Demorest, M.E. & Tucker, P.E., 2000. Speech perception without hearing. *Perception and Psychophysics*, 62(2), p. 233-252.

Bernstein, L.E., Lu, Z.-L. & Jiang, J., 2008. Quantified acoustic-optical speech signal incongruity identifies cortical sites of audiovisual speech processing. *Brain Research*, 1242, p. 172-184. Available at: http://dx.doi.org/10.1016/j.brainres.2008.04.018.

Bertelson, P., Vroomen, J., Wiegeraad, G. & de Gelder, B., 1994. Exploring The Relation Between Mcgurk Interference And Ventriloquism. In *Third International Conference on Spoken Language Processing (ICSLP 94)*. Yokohama, Japan.

Berthommier, F., 2001. Audio-visual recognition of spectrally reduced speech. In *Proceedings of the 4th International Conference on Auditory-Visual Speech Processing (AVSP 2001)*. Scheelsminde, Danemark.

Berthommier, F., 2004. A phonetically neutral model of the low-level audio-visual interaction. *Speech Communication*, 44(1), p. 31-41. Available at: http://www.sciencedirect.com/science/article/pii/S016763930400113X.

Besle, J., Fort, A., Delpuech, C. & Giard, M.-H., 2004. Bimodal speech: early suppressive visual effects in human auditory cortex. *European Journal of Neuroscience*, 20(8), p. 2225-2234. Available at: http://dx.doi.org/10.1111/j.1460-9568.2004.03670.x.

Bloch, I., 1996. Information combination operators for data fusion: a comparative review with classification. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, 26(1), p. 52-67.

Bovo, R., Ciorba, A., Prosser, S. & Martini, A., 2009. The McGurk phenomenon in Italian listeners. *Acta Otorhinolaryngologica Italica : organo ufficiale della Societa italiana di otorinolaringologia e chirurgia cervico-facciale*, 29(4)(4), p. 203–208. Available at: http://ukpmc.ac.uk/abstract/MED/20161878.

Brancazio, L., Best, C.T. & Fowler, C.A., 2006. Visual influences on perception of speech and nonspeech vocal-tract events. *Language and Speech*, 49(Pt 1), p. 21-53.

Bregman, A.S., 1990. *Auditory scene analysis: The perceptual organization of sound.*: Cambridge, MA, US: The MIT Press. xiii 773 pp. Available at: http://books.google.fr/books?hl=fr&lr=&id=jI8muSpAC5AC&oi=fnd&pg=PR11&ots=SEnWM8AJAC&sig=tHkhVpwLskQLTNNprfY6KmVLvTU.

Bregman, A.S. & Pinker, S., 1978. Auditory streaming and the building of timbre. *Canadian Journal of Psychology*, 32(1), p. 19-31.

Bregman, A.S. & Rudnicky, A.I., 1975. Auditory segregation: stream or streams? *Journal of Experimental Psychology: Human Perception and Performance*, 1(3), p. 263-267.

Brüel & Kjaer, 2013. *Sonomètre. Sound Quality Head and Torso Simulator - Type 4100*. [Online]. Available at: http://www.bksv.fr/Products/transducers/ear-simulators/head-and-torso/4100.aspx?tab=overview

Brungart, D.S. & Simpson, B.D., 2005. Interference from audio distracters during speechreading. *Journal of the Acoustical Society of America*, 118(6), p. 3889-3902.

Burnham, D. & Dodd, B., 2004. Auditory-visual speech integration by prelinguistic infants: perception of an emergent consonant in the McGurk effect. *Developmental Psychobiology*, 45(4), p. 204-220. Available at: http://dx.doi.org/10.1002/dev.20032.

Burnham, D., Lau, S., Tam, H. & Schoknecht, C., 2001. Visual Discrimination of Cantonese Tone by Tonal but Non-Cantonese Speakers, and by Non-Tonal Language Speakers. In *Auditory-Visual Speech Processing (AVSP 2001)*. Aalborg, Denmark.

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(17), p. 2213-2218. Available at: http://dx.doi.org/10.1097/01.wnr.0000095492.38740.8f.

Calvert, G.A., Brammer, M.J., Bullmore, E.T., Campbell, R., Iversen, S.D. & David, A.S., 1999. Response amplification in sensory-specific cortices during crossmodal binding. *Neuroreport*, 10(12), p. 2619-2623.

Calvert, G.A., Bullmore, E.T., Brammer, M.J., Campbell, R., Williams, S.C., McGuire, P.K., Woodruff, P.W., Iversen, S.D. & David, A.S., 1997. Activation of auditory cortex during silent lipreading. *Science*, 276(5312), p. 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(11), p. 649-657.

Campbell, R., 2008. The processing of audio-visual speech: empirical and neural bases. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 363(1493), p. 1001-1010. Available at: http://dx.doi.org/10.1098/rstb.2007.2155.

Cathiard, M.A., Schwartz, J.-L. & Abry, C., 2001. Asking a naive question about the McGurk Effect: why does audio [b] give more [d] percepts with visual [g] than with visual [d]? In *International Conference on Auditory-Visual Speech Processing, AVSP'01*. Aalborg, Danmark.

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A. & Ghazanfar, A.A., 2009. The natural statistics of audiovisual speech. *PLOS Computational Biology*, 5(7), p. e1000436. Available at: http://dx.doi.org/10.1371/journal.pcbi.1000436.

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(5), p. 975-979. Available at: http://link.aip.org/link/?JAS/25/975/1.

Colin, C., Radeau, M., Deltenre, P., Demolin, D. & Soquet, A., 2002. The role of sound intensity and stop-consonant voicing on McGurk fusions and combinations. *European Journal of Cognitive Psychology*, 14(4), p. 475-491. Available at: http://www.tandfonline.com/doi/abs/10.1080/09541440143000203.

Colin, C., Radeau, M., Deltenre, P. & Morais, J., 2001. Rules of intersensory integration in spatial scene analysis and speechreading. *Psychologica Belgica*, 41(3), p. 131-144.

Colin, C. & Radeau, M., 2003. Les illusions McGurk dans la parole : 25 ans de recherches. *L'année psychologique*, 103(3), p. 497-542. Available at: http://www.persee.fr/web/revues/home/prescript/article/psy_0003-5033_2003_num_103_3_29649.

Colin, C., Radeau, M., Soquet, A. & Deltenre, P., 2004. Generalization of the generation of an MMN by illusory McGurk percepts: voiceless consonants. *Clinical Neurophysiology*, 115(9), p. 1989-2000. Available at: http://dx.doi.org/10.1016/j.clinph.2004.03.027.

Colin, C., Radeau, M., Soquet, A., Demolin, D., Colin, F. & Deltenre, P., 2002. Mismatch negativity evoked by the McGurk-MacDonald effect: a phonetic representation within short-term memory. *Clinical Neurophysiology*, 113(4), p. 495-506.

Correa, A., Lupiáñez, J., Madrid, E. & Tudela, P., 2006. Temporal attention enhances early visual processing: a review and new evidence from event-related potentials. *Brain Research*, 1076(1), p. 116-128. Available at: http://dx.doi.org/10.1016/j.brainres.2005.11.074.

Coull, J.T. & Nobre, A.C., 1998. Where and when to pay attention: the neural systems for directing attention to spatial locations and to time intervals as revealed by both PET and fMRI. *The Journal of Neuroscience*, 18(18), p. 7426-7435.

Davis, C. & Kim, J., 2001. Repeating and remembering foreign language words : implications for language teaching systems. *Artificial Intelligence Review*, 16(1), p. 37-47. Available at: http://handle.uws.edu.au:8081/1959.7/10640.

de Gelder, B. & Vroomen, J., 2000. The perception of emotions by ear and by eye. *Cognition and Emotion*, 14(3), p. 289-311. Available at: http://www.tandfonline.com/doi/abs/10.1080/026999300378824.

Dekle, D.J., Fowler, C.A. & Funnell, M.G., 1992. Audiovisual integration in perception of real words. *Perception and Psychophysics*, 51(4), p. 355-362.

Driver, J., 1996. Enhancement of selective listening by illusory mislocation of speech sounds due to lip-reading. *Nature*, 381, p. 66-68.

Driver, J. & Noesselt, T., 2008. Multisensory interplay reveals crossmodal influences on 'sensory-specific' brain regions, neural responses, and judgments. *Neuron*, 57(1), p. 11-23. Available at: http://dx.doi.org/10.1016/j.neuron.2007.12.013.

Duran, A.F., 1995. Mcgurk effect in Spanish and German listeners: influences of visual cues in the perception of Spanish and German conflicting audio-visual stimuli. In *European Conference on Speech Communication and Technology, EUROSPEECH*. ISCA.

Engel, A.K., Fries, P., König, P., Brecht, M. & Singer, W., 1999. Temporal binding, binocular rivalry, and consciousness. *Consciousness and Cognition*, 8(2), p. 128-151. Available at: http://dx.doi.org/10.1006/ccog.1999.0389.

Erber, N.P., 1969. Interaction of audition and vision in the recognition of oral speech stimuli. *Journal of speech and hearing research*, 12(2), p. 423-425.

Ernst, M.O. & Banks, M.S., 2002. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, 415(6870), p. 429-433. Available at: http://dx.doi.org/10.1038/415429a.

Fowler, C.A., 1986. An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics*, 14, p. 3-28.

Fries, P., Nikolić, D. & Singer, W., 2007. The gamma cycle. *Trends in Neurosciences*, 30(7), p. 309-316. Available at: http://dx.doi.org/10.1016/j.tins.2007.05.005.

Gelder, B.d., Bertelson, P., Vroomen, J. & Chen, H., 1995. Inter-language differences in the McGurk effects for Dutch and Cantonese listeners. In *Fourth European Conference on Speech Communication and Technology, EUROSPEECH'95*. Madrid, Spain.

Geschwind, N., 1972. Language and the brain. *Scientific American*, 226(4), p. 76-83.

Ghazanfar, A.A., Chandrasekaran, C. & Logothetis, N.K., 2008. Interactions between the superior temporal sulcus and auditory cortex mediate dynamic face/voice integration in rhesus monkeys. *The Journal of Neuroscience*, 28(17), p. 4457-4469. Available at: http://dx.doi.org/10.1523/JNEUROSCI.0541-08.2008.

Ghazanfar, A.A. & Schroeder, C.E., 2006. Is neocortex essentially multisensory? *Trends in Cognitive Sciences*, 10(6), p. 278-285. Available at: http://dx.doi.org/10.1016/j.tics.2006.04.008.

Gondan, M., Niederhaus, B., Rösler, F. & Röder, B., 2005. Multisensory processing in the redundant-target effect: a behavioral and event-related potential study. *Perception and Psychophysics*, 67(4), p. 713-726.

Grant, K.W. & Greenberg, S., 2001. Speech Intelligibility Derived from Asynchronous Processing of Auditory-Visual Information. In *Proceedings of the 4th International Conference on Auditory-Visual Speech Processing (AVSP 2001)*. Aalborg, Denmark.

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(4), p. 2438-2450. Available at: http://dx.doi.org/10.1093/cercor/bhr125.

Grant, K.W. & Seitz, P.F., 2000. The use of visible speech cues for improving auditory detection of spoken sentences. *Journal of the Acoustical Society of America*, 108(3 Pt 1), p. 1197-1208.

Grassegger, H., 1995. McGurk effect in German and Hungarian listeners. In *Proceedings of the XIIIth International Congress of Phonetics Sciences*. Stockholm.

Green, K.P., Kuhl, P.K., Meltzoff, A.N. & Stevens, E.B., 1991. Integrating speech information across talkers, gender, and sensory modality: female faces and male voices in the McGurk effect. *Perception and Psychophysics*, 50(6), p. 524-536.

Helmholtz, H., 1877. *On the sensations of tone.*: New York, Dover.

Hickok, G., 2009. The functional neuroanatomy of language. *Physics of Life Reviews*, 6(3), p. 121-143. Available at: http://dx.doi.org/10.1016/j.plrev.2009.06.001.

Hickok, G. & Poeppel, D., 2007. The cortical organization of speech processing. *Nature Reviews Neuroscience*, 8(5), p. 393-402. Available at: http://dx.doi.org/10.1038/nrn2113.

Hickok, G. & Poeppel, D., 2004. Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language. *Cognition*, 92(1-2), p. 67-99. Available at: http://dx.doi.org/10.1016/j.cognition.2003.10.011.

Hisanaga, S., Sekiyama, K., Igasaki, T. & Murayama, N., 2009. Audiovisual speech perception in Japanese and English: inter-language differences examined by event-related potentials. In *Proceedings of the 8th International Conference on Auditory-Visual Speech Processing (AVSP 2009)*. University of East Anglia, Norwich, UK.

Hocking, J. & Price, C.J., 2008. The role of the posterior superior temporal sulcus in audiovisual processing. *Cerebral Cortex*, 18(10), p. 2439-2449. Available at: http://dx.doi.org/10.1093/cercor/bhn007.

Hupé, J.-M. & Pressnitzer, D., 2012. The initial phase of auditory and visual scene analysis. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 367(1591), p. 942-953. Available at: http://dx.doi.org/10.1098/rstb.2011.0368.

Iacoboni, M., Woods, R.P., Brass, M., Bekkering, H., Mazziotta, J.C. & Rizzolatti, G., 1999. Cortical mechanisms of human imitation. *Science*, 286(5449), p. 2526-2528.

Jiang, J., 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(11), p. 506945. Available at: http://asp.eurasipjournals.com/content/2002/11/506945.

Jones, M.R. & Boltz, M., 1989. Dynamic attending and responses to time. *Psychological Review*, 96(3), p. 459-491.

Jones, J.A. & Callan, D.E., 2003. Brain activity during audiovisual speech perception: an fMRI study of the McGurk effect. *Neuroreport*, 14(8), p. 1129-1133. Available at: http://dx.doi.org/10.1097/01.wnr.0000074343.81633.2a.

Jones, J. & Munhall, K., 1997. The effects of separating auditory and visual sources on audiovisual integration of speech. *Canadian Acoustics*, 25(4), p. 13-19.

Jones, M.R., 1976. Time, our lost dimension: toward a new theory of perception, attention, and memory. *Psychological Review*, 83(5), p. 323-355.

Keil, J., Müller, N., Ihssen, N. & Weisz, N., 2012. On the variability of the McGurk effect: audiovisual integration depends on prestimulus brain states. *Cerebral Cortex*, 22(1), p. 221-231. Available at: http://dx.doi.org/10.1093/cercor/bhr125.

Kilner, J.M., Friston, K.J. & Frith, C.D., 2007. Predictive coding: an account of the mirror neuron system. *Cognitive Processing*, 8(3), p. 159-166. Available at: http://dx.doi.org/10.1007/s10339-007-0170-2.

Kim, J. & Davis, C., 2003. Hearing foreign voices: does knowing what is said affect visual-masked-speech detection? *Perception*, 32(1), p. 111-120.

Kim, J. & Davis, C., 2004. Investigating the audio–visual speech detection advantage. *Speech Communication*, 44(1–4), p. 19-30. Available at: http://www.sciencedirect.com/science/article/pii/S0167639304001189.

Klatt, D.H., 1979. Speech perception: A model of acoustic-phonetic analysis and lexical access. *Journal of Phonetics*, 7, p. 279-312.

Kohler, E., Keysers, C., Umiltà, M.A., Fogassi, L., Gallese, V. & Rizzolatti, G., 2002. Hearing sounds, understanding actions: action representation in mirror neurons. *Science*, 297(5582), p. 846-848. Available at: http://dx.doi.org/10.1126/science.1070311.

Kondo, H.M. & Kashino, M., 2007. Neural mechanisms of auditory awareness underlying verbal transformations. *Neuroimage*, 36(1), p. 123-130. Available at: http://dx.doi.org/10.1016/j.neuroimage.2007.02.024.

Kösem, A. & van Wassenhove, V., 2012. Temporal structure in audiovisual sensory selection. *PLoS One*, 7(7), p. e40936. Available at: http://dx.doi.org/10.1371/journal.pone.0040936.

Kuhl, P.K. & Meltzoff, A.N., 1982. The bimodal perception of speech in infancy. *Science*, 218(4577), p. 1138-1141.

Kuhl, P.K. & Meltzoff, A.N., 1984. The Intermodal Representation of Speech in Infants. *Infant Behavior and Development*, 7(3), p. 361-381. Available at: http://www.sciencedirect.com/science/article/pii/S0163638384800508.

Lallouache, M., 1990. Un poste 'visage-parole'. Acquisition et traitement de contours labiaux. (A “face-speech” workstation. Acquisition and processing of labial contours. In *Actes des XVIIIèmes Journées d'Etude sur la Parole*. Montréal, Canada.

Lehar, S., 2003. *The World in Your Head. A Gestalt View of the Mechanism of Conscious Experience*. 1st ed.: Lawrence Erlbaum Associates Inc. (Mahwaw).

Liberman, A.M. & Mattingly, I.G., 1985. The motor theory of speech perception revised. *Cognition*, 21(1), p. 1-36. Available at: http://www.sciencedirect.com/science/article/pii/0010027785900216.

Lichtheim, L., 1885. On aphasia. *Brain*, 7, p. 433–484.

Lisker, L. & Rossi, M., 1992. Auditory and visual cueing of the [+/- rounded] feature of vowels. *Language and Speech*, 35 (Pt 4), p. 391-417.

MacDonald, J., Andersen, S. & Bachmann, T., 1999. Hearing by eye: visual spatial degradation and the mcgurk effect. In *Sixth European Conference on Speech Communication and Technology, EUROSPEECH*. Budapest, Hungary ISCA.

Manuel, S., Repp, B., Liberman, A. & Studdert-Kennedy, M., 1983. Exploring the "McGurk effect". In *24th annual meeting of the Psychonomic Society*. San Diego, California, USA.

Massaro, D.W. & Cohen, M.M., 1983. Evaluation and integration of visual and auditory information in speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, 9(5), p. 753-771.

Massaro, D.W., 1989. Multiple Book Review of Speech perception by ear and eye: A paradigm for psychological inquiry. *Behavioral and Brain Sciences*, 12(04), p. 741-755. Available at: http://dx.doi.org/10.1017/S0140525X00025619.

Massaro, D.W., 1987. *Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry*.: Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc. 320 pp. Available at: http://books.google.fr/books?hl=fr&lr=&id=8ryyuG-lxCUC&oi=fnd&pg=PA2&ots=2nihHPov-w&sig=Sw_G2vsncj_fiZKN9kBmdsNQndk.

McGrath, M. & Summerfield, Q., 1985. Intermodal timing relations and audio-visual speech recognition by normal-hearing adults. *Journal of the Acoustical Society of America*, 77(2), p. 678-685.

McGurk, H., 1981. The Cognitive Representation of Speech.: North Holland. Ch. Listening with eye and ear. p. 336-337.

McGurk, H. & MacDonald, J., 1976. Hearing lips and seeing voices. *Nature*, 264(5588), p. 746-748.

Miller, L.M. & D'Esposito, M., 2005. Perceptual fusion and stimulus coincidence in the cross-modal integration of speech. *The Journal of Neuroscience*, 25(25), p. 5884-5893. Available at: http://dx.doi.org/10.1523/JNEUROSCI.0896-05.2005.

Moore, B.C., 2003. *An introduction to the psychology of hearing*. 5th ed.: Amsterdam [etc.] : Academic press.

Munhall, K.G., Gribble, P., Sacco, L. & Ward, M., 1996. Temporal constraints on the McGurk effect. *Perception and Psychophysics*, 58(3), p. 351-362.

Munhall, K.G.V.-B.E., 1998. Hearing by eye II: Advances in the psychology of speechreading and auditory-visual speech.: Hove, England: Psychology Press/Erlbaum (UK) Taylor and Francis, xiv, 319 pp. Ch. The moving face during speech communication. p. 123-139.

Munhall, K., Jones, J.A., Callan, D.E., Kuratate, T. & Vatikiotis-Bateson, E., 2004. Visual Prosody and Speech Intelligibility: Head Movement Improves Auditory Speech Perception. *Psychological Science*, 15(2), p. 133-137. Available at: http://pss.sagepub.com/content/15/2/133.abstract.

Nahorna, O., Chandrashekara, G.A., Berthommier, F. & Schwartz, J.-L., 2013. Modulating fusion in the McGurk effect by binding processes and contextual noise. In *Proceedings of the 11th International Conference on Auditory-Visual Speech Processing (AVSP 2013), St Jorioz, France.*

Nath, A.R. & Beauchamp, M.S., 2011. Dynamic changes in superior temporal sulcus connectivity during perception of noisy audiovisual speech. *The Journal of Neuroscience*, 31(5), p. 1704-1714. Available at: http://dx.doi.org/10.1523/JNEUROSCI.4853-10.2011.

Noppeney, U., Ostwald, D. & Werner, S., 2010. Perceptual decisions formed by accumulation of audiovisual evidence in prefrontal cortex. *The Journal of Neuroscience*, 30(21), p. 7434-7446. Available at: http://dx.doi.org/10.1523/JNEUROSCI.0455-10.2010.

Ojanen, V., Möttönen, R., Pekkola, J., Jääskeläinen, I.P., Joensuu, R., Autti, T. & Sams, M., 2005. Processing of audiovisual speech in Broca's area. *Neuroimage*, 25(2), p. 333-338. Available at: http://dx.doi.org/10.1016/j.neuroimage.2004.12.001.

Okada, K. & Hickok, G., 2009. Two cortical mechanisms support the integration of visual and auditory speech: a hypothesis and preliminary data. *Neuroscience Letters*, 452(3), p. 219-223. Available at: http://dx.doi.org/10.1016/j.neulet.2009.01.060.

Ponton, C.W., Bernstein, L.E. & Auer, J.E.T., 2009. Mismatch negativity with visual-only and audiovisual speech. *Brain Topography*, 21(3-4), p. 207-215. Available at: http://dx.doi.org/10.1007/s10548-009-0094-5.

Pulvermüller, F., Huss, M., Kherif, F., Moscoso del Prado Martin, F., Hauk, O. & Shtyrov, Y., 2006. Motor cortex maps articulatory features of speech sounds. *Proceedings of the National Academy of Sciences USA*, 103(20), p. 7865-7870. Available at: http://dx.doi.org/10.1073/pnas.0509989103.

Reisberg, D., 1978. Looking where you listen: visual cues and auditory attention. *Acta Psychologica (Amst)*, 42(4), p. 331-341. Available at: http://dx.doi.org/10.1016/0001-6918(78)90007-0.

Reisberg, D., Mclean, J. & Goldfield, A., 1987. Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli. In Dodd, B. & Campbell, R. eds. *Hearing by eye: The psychology of lip-reading*. Hillsdale, New Jersey: Lawrence Erlbaum Associates. p. 97-114.

Revonsuo, A. & Newman, J., 1999. Binding and consciousness. *Consciousness and Cognition*, 8(2), p. 123-127. Available at: http://dx.doi.org/10.1006/ccog.1999.0393.

Risberg, A. & Lubker, J.L., 1978. *Prosody and speechreading*. Quaterly Progress and Status Report. Speech Transmission Laboratory

Rizzolatti, G. & Craighero, L., 2004. The mirror-neuron system. *Annual Review of Neuroscience*, 27, p. 169-192. Available at: http://dx.doi.org/10.1146/annurev.neuro.27.070203.144230.

Rizzolatti, G., Fadiga, L., Gallese, V. & Fogassi, L., 1996. Premotor cortex and the recognition of motor actions. *Cognitive Brain Research*, 3(2), p. 131-141.

Rizzolatti, G., Fadiga, L., Matelli, M., Bettinardi, V., Paulesu, E., Perani, D. & Fazio, F., 1996. Localization of grasp representations in humans by PET: 1. Observation versus execution. *Experimental Brain Research*, 111(2), p. 246-252.

Rizzolatti, G., Fogassi, L. & Gallese, V., 2001. Neurophysiological mechanisms underlying the understanding and imitation of action. *Nature Reviews Neuroscience*, 2(9), p. 661-670. Available at: http://dx.doi.org/10.1038/35090060.

Sato, M., Baciu, M., Loevenbruck, H., Schwartz, J.-L., Cathiard, M.-A., Segebarth, C. & Abry, C., 2004. Multistable representation of speech forms: a functional MRI study of verbal transformations. *Neuroimage*, 23(3), p. 1143-1151. Available at: http://dx.doi.org/10.1016/j.neuroimage.2004.07.055.

Sato, M., Basirat, A. & Schwartz, J.-L., 2007. Visual contribution to the multistable perception of speech. *Perception and Psychophysics*, 69(8), p. 1360-1372.

Sato, M., Vallée, N., Schwartz, J.-L. & Rousset, I., 2007. A perceptual correlate of the labial-coronal effect. *Journal of Speech Language and Hearing Research*, 50(6), p. 1466-1480. Available at: http://dx.doi.org/10.1044/1092-4388(2007/101).

Schwartz, J.-L., Basirat, A., Ménard, L. & Sato, M., 2010. The Perception-for-Action-Control Theory (PACT): A perceptuo-motor theory of speech perception. *Journal of Neurolinguistics*, 25(5), p. 336-354. Available at: http://www.sciencedirect.com/science/article/pii/S0911604409000876.

Schwartz, J.-L., Berthommier, F. & Savariaux, C., 2004. Seeing to hear better: evidence for early audio-visual interactions in speech identification. *Cognition*, 93(2), p. B69--B78. Available at: http://dx.doi.org/10.1016/j.cognition.2004.01.006.

Schwartz, J.-L., Grimault, N., Hupé, J.-M., Moore, B.C.J. & Pressnitzer, D., 2012. Multistability in perception: binding sensory modalities, an overview. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 367(1591), p. 896-905. Available at: http://dx.doi.org/10.1098/rstb.2011.0254.

Schwartz, J.-L., 2010. A reanalysis of McGurk data suggests that audiovisual fusion in speech perception is subject-dependent. *Journal of the Acoustical Society of America*, 127(3), p. 1584-1594. Available at: http://dx.doi.org/10.1121/1.3293001.

Schwartz, J.-L., Robert-Ribes, J. & Escudier, P., 1998. Hearing by eye II: Advances in the psychology of speechreading and auditory-visual speech.: Hove, England: Psychology Press/Erlbaum (UK) Taylor \& Francis, xiv, 319 pp.. Ch. Ten years after Summerfield: A taxonomy of models for audio–visual fusion in speech perception. p. 85-108.

Schwartz, J.-L., Tiippana, K. & Andersen, T.S., 2010. Disentangling unisensory from fusion effects in the attentional modulation of Mcgurk effects: a Bayesian modeling study suggests that fusion is attention-dependent. In *Proceedings of the 9th International Conference on Auditory-Visual Speech Processing (AVSP 2010)*. Hakone, Kanagawa, Japon.

Sekiyama, K. & Burnham, D., 2008. Impact of language on development of auditory-visual speech perception. *Developmental Science*, 11(2), p. 306-320. Available at: http://dx.doi.org/10.1111/j.1467-7687.2008.00677.x.

Sekiyama, K., 1997. Cultural and linguistic factors in audiovisual speech processing: the McGurk effect in Chinese subjects. *Perception and Psychophysics*, 59(1), p. 73-80.

Sekiyama, K. & Tohkura, Y., 1993. Inter-language differences in the influence of visual cues in speech perception. *Journal of Phonetics*, 21(4), p. 427-444.

Sekiyama, K. & Tohkura, Y., 1991. McGurk effect in non-English listeners: Few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility. *The Journal of the Acoustical Society of America*, 90(4), p. 1797-1805. Available at: http://link.aip.org/link/?JAS/90/1797/1.

Senkowski, D., Saint-Amour, D., Gruber, T. & Foxe, J.J., 2008. Look who's talking: the deployment of visuo-spatial attention during multisensory speech processing under noisy environmental conditions. *Neuroimage*, 43(2), p. 379-387. Available at: http://dx.doi.org/10.1016/j.neuroimage.2008.06.046.

Senkowski, D., Schneider, T.R., Foxe, J.J. & Engel, A.K., 2008. Crossmodal binding through neural coherence: implications for multisensory processing. *Trends in Neurosciences*, 31(8), p. 401-409. Available at: http://dx.doi.org/10.1016/j.tins.2008.05.002.

Shalom, D.B. & Poeppel, D., 2008. Functional anatomic models of language: assembling the pieces. *Neuroscientist*, 14(1), p. 119-127. Available at: http://dx.doi.org/10.1177/1073858407305726.

Singer, W., 1999. Neuronal synchrony: a versatile code for the definition of relations? *Neuron*, 24(1), p. 49--65, 111-25.

Skipper, J.I., Nusbaum, H.C. & Small, S.L., 2005. Listening to talking faces: motor cortical activation during speech perception. *Neuroimage*, 25(1), p. 76-89. Available at: http://dx.doi.org/10.1016/j.neuroimage.2004.11.006.

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(10), p. 2387-2399. Available at: http://dx.doi.org/10.1093/cercor/bhl147.

Soto-Faraco, S. & Alsius, A., 2007. Conscious access to the unisensory components of a cross-modal illusion. *Neuroreport*, 18(4), p. 347-350. Available at: http://dx.doi.org/10.1097/WNR.0b013e32801776f9.

Soto-Faraco, S. & Alsius, A., 2009. Deconstructing the McGurk-MacDonald illusion. *Journal of experimental psychology. Human perception and performance*, 35(2), p. 580-587. Available at: http://dx.doi.org/10.1037/a0013483.

Stekelenburg, J.J. & Vroomen, J., 2007. Neural correlates of multisensory integration of ecologically valid audiovisual events. *Journal of Cognitive Neuroscience*, 19(12), p. 1964-1973. Available at: http://dx.doi.org/10.1162/jocn.2007.19.12.1964.

Stevenson, R.A., Zemtsov, R.K. & Wallace, M.T., 2012. Individual Differences in the Multisensory Temporal Binding Window Predict Susceptibility to Audiovisual Illusions. *The Journal of Experimental Psychology: Human Perception and Performance* Available at: http://dx.doi.org/10.1037/a0027339.

Sumby, W.H. & Pollack, I., 1954. Visual Contribution to Speech Intelligibility in Noise. *The Journal of the Acoustical Society of America*, 26(2), p. 212-215. Available at: http://link.aip.org/link/?JAS/26/212/1.

Summerfield, Q., 1987. Hearing by eye: The psychology of lip-reading.: Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc, x, 268 pp. Ch. Some preliminaries to a comprehensive account of audio-visual speech perception. p. 3-51.

Summerfield, Q. & McGrath, M., 1984. Detection and resolution of audio-visual incompatibility in the perception of vowels. *The Quarterly Journal of Experimental Psychology*, 36(1), p. 51-74.

Summerfield, Q., 1979. Use of visual information for phonetic perception. *Phonetica*, 36(4-5), p. 314-331.

Tallon-Baudry, C., Bertrand, O., Delpuech, C. & Permier, J., 1997. Oscillatory gamma-band (30-70 Hz) activity induced by a visual search task in humans. *The Journal of Neuroscience*, 17(2), p. 722-734.

Tallon-Baudry, C., Bertrand, O., Delpuech, C. & Pernier, J., 1996. Stimulus specificity of phase-locked and non-phase-locked 40 Hz visual responses in human. *The Journal of Neuroscience*, 16(13), p. 4240-4249.

Tanaka, A., Sakamoto, S., Tsumura, K. & Suzuki, Y., 2009. Visual speech improves the intelligibility of time-expanded auditory speech. *Neuroreport*, 20(5), p. 473-477. Available at: http://dx.doi.org/10.1097/WNR.0b013e3283279ae8.

Teissier, P., Robert-Ribes, J., Schwartz, J.-L. & Guerin-Dugue, A., 1999. Comparing models for audiovisual fusion in a noisy-vowel recognition task. *IEEE Transactions on Speech and Audio Processing*, 7(6), p. 629-642.

Thompson, L. & Ogden, W., 1995. Visible speech improves human language understanding: Implications for speech processing systems. *Artificial Intelligence Review*, 9, p. 347-358. Available at: http://dx.doi.org/10.1007/BF00849044.

Tiippana, K., Andersen, T.S. & Sams, M., 2004. Visual attention modulates audiovisual speech perception. *European Journal of Cognitive Psychology*, 16(3), p. 457-472. Available at: http://www.tandfonline.com/doi/abs/10.1080/09541440340000268.

Treisman, A.M. & Gelade, G., 1980. A feature-integration theory of attention. *Cognitive Psychology*, 12(1), p. 97-136.

Treisman, A., 1992. Introduction aux sciences cognitives.: Paris: Editions Gallimard. Ch. L'attention, les traits et la perception des objets. p. 153-191.

Treisman, A. & Schmidt, H., 1982. Illusory conjunctions in the perception of objects. *Cognitive Psychology*, 14(1), p. 107-141.

Van der Burg, E., Olivers, C.N.L., Bronkhorst, A.W. & Theeuwes, J., 2008. Pip and pop: nonspatial auditory signals improve spatial visual search. *Journal of Experimental Psychology: Human Perception and Performance*, 34(5), p. 1053-1065. Available at: http://dx.doi.org/10.1037/0096-1523.34.5.1053.

Van Noorden, L., 1975. *Temporal coherence in the perception of tone sequences.* Ph.D. dissertation.

van Wassenhove, V., Grant, K.W. & Poeppel, D., 2007. Temporal window of integration in auditory-visual speech perception. *Neuropsychologia*, 45(3), p. 598-607. Available at: http://dx.doi.org/10.1016/j.neuropsychologia.2006.01.001.

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 USA*, 102(4), p. 1181-1186. Available at: http://dx.doi.org/10.1073/pnas.0408949102.

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(1), p. 316-325. Available at: http://dx.doi.org/10.1016/j.neuroimage.2006.05.032.

Wilson, S.M., Saygin, A.P., Sereno, M.I. & Iacoboni, M., 2004. Listening to speech activates motor areas involved in speech production. *Nature Neuroscience*, 7(7), p. 701-702. Available at: http://dx.doi.org/10.1038/nn1263.

Wright, T.M., Pelphrey, K.A., Allison, T., McKeown, M.J. & McCarthy, G., 2003. Polysensory interactions along lateral temporal regions evoked by audiovisual speech. *Cerebral Cortex*, 13(10), p. 1034-1043.

Yehia, H., Rubin, P. & Vatikiotis-Bateson, E., 1998. Quantitative association of vocal-tract and facial behavior. *Speech Communication*, 26(1–2), p. 23-43. Available at: http://www.sciencedirect.com/science/article/pii/S016763939800048X.

Yu, A.J., Dayan, P. & Cohen, J.D., 2009. Dynamics of attentional selection under conflict: toward a rational Bayesian account. *The Journal of Experimental Psychology: Human Perception and Performance*, 35(3), p. 700-717. Available at: http://dx.doi.org/10.1037/a0013553.