Interactions audiovisuelles pour l'analyse de scènes auditives

Aymeric Devergie

N° 283 - 2010 Année 2010

thèse de l'université de Lyon

délivrée par

l'université claude bernard lyon 1

École doctorale M.E.G.A

diplôme de doctorat
(Arrêté du 07 août 2006)
mention : Acoustique

soutenue publiquement le 10 décembre 2010
par

Aymeric DEVERGIE

Interactions Audiovisuelles pour l'Analyse de Scènes Auditives

Directeur de thèse : Pr. Jean-Louis Guyader
Co-directeurs : Dr. Nicolas Grimault
Dr. Frédéric Berthommier

Jury : Dr. Laurent Demany, Rapporteur
Pr. Jean Vroomen, Rapporteur
Dr. Pascal Barone, Examinateur
Dr. François Pellegrino, Examinateur, Président du Jury
Pr. Jean-Louis Guyader
Dr. Nicolas Grimault
Dr. Frédéric Berthommier

Remerciements

Prenant soin de la partie de ma thèse qui sera probablement la plus lue, voici mes remerciements.

Je tiens tout d'abord à remercier Messieurs les membres du jury d'avoir accepté d'évaluer ce travail de thèse. Merci pour ces échanges et discussions relatives à ce thème de recherche passionnant et complexe.

Ce travail a été réalisé dans deux laboratoires rhône-alpins. Mes remerciements voyageront tout comme moi entre Lyon et Grenoble.

Merci à Rémi Gervais de m'avoir accueilli au sein du laboratoire NSCC à Lyon pendant ces trois années. NSCC fût un environnement propice aux échanges tant scientifiques que humains et ont contribués à rendre ce travail passionnant. Les demi-journées des doctorants furent extrêmement bénéfiques et intéressantes. Merci également à Gérard Bailly, directeur du Département Parole et Cognition du Gipsa-LAB à Grenoble de m'avoir accueilli lors de mes déplacements dans la capitale des Alpes.

Merci à Jean-Louis Guyader d'avoir accepté de diriger cette Thèse. Malgré l'apparente distance de la thématique de recherche, vous avez su prêter une oreille attentive et experte vis à vis de ce travail. Les discussions que nous avons pu avoir ont toujours été enrichissantes.

Merci à Barbara Tillmann et à Hélène Loevenbruck pour l'ensemble des discussions, conseils et collaborations que nous avons partager ces trois années.

J'adresse un merci collectif à l'ensemble de l'équipe CAP qui a su conserver sa dynamique à l'occasion des si nombreuses réunions d'équipe.

Merci à Nicolas Grimault pour toutes ces qualités tant scientifiques que humaines. Parmis elles, je citerais son pragmatisme, sa rigueur et sa

décontraction qui ont rendu ce travail passionnant et fortement enrichissant. Instauré un cadre aussi propice à la recherche fondamentale ainsi qu'une ambiance si agréable mérite d'être souligné de manière appuyée.

Merci à Frédéric Berthommier pour les échanges scientifiques intarrissables sur la parole audiovisuelle. C'est avec beaucoup de pédagogie que vous m'avez fait prendre conscience de la complexité inhérente à ce thème de recherche.

Je tiens à vous remercier tous les deux car grâce à cette collaboration tri-partite, ce travail n'a cessé d'être enrichi d'idées et de points de vue complémentaires.

Les amis, c'est ici que je parle de vous.

Ayant vécu deux époques, ici, à NSCC je commencerais par l'époque faste où l'on comptait deux thésards au mètre carré. Merci à Etienne pour m'avoir accepter dans son bureau. De grandes affinités sont apparues dès le début. Le fait que nous restions toujours en contact est une preuve supplémentaire de notre bonne entente. Merci pour tous les soutiens que tu m'as apporté. Tu es un grand.

Merci à Carine pour m'avoir fait découvert le 42 et le café Moka. Merci pour ces grands éclats de rire et ces longues discussions chez Mégane. Merci aux psychologues Fred Marmel et Lisianne pour m'avoir fait partager leur monde étrange où l'on parle cortex, neurones ou encore psycholinguistique.

Merci aux métalleux Johan et Nicolas le Teigneux pour l'immersion immédiate dans cet univers musical rempli de poésie. Merci pour ces instants de franche camaraderie.

Merci aux exilés, Tristan, Seb et Germain pour leur passage à NSCC qui aura laissé une trace indélébile.

Merci à la nouvelle génération d'étudiants qui assurera la relève. Merci Charlotte, Lauranne et Tatiana pour avoir supporté ma présence et accepté de partager le bureau 201 avec moi.

Merci Jérôme, Ben, Philippe et Carlos. Après mon départ, vous serez la mémoire de cette époque CAP désormais révolue.

Merci également à Pauline et Amandine qui seront les mémoires de l'ère

Poncelet et du fameux bureau où le café coula à flot et où les comptes à bâtons remplissaient des pans entiers de mur.

Enfin, une thèse sans famille n'aurait pu arriver à son terme.

Merci Patrick et Jutta pour votre confiance dans les choix que j'ai fait. Merci pour votre présence malgré la distance géographique récente. Merci à Camille et Anthony pour m'avoir gratifié du statut de Tonton récemment avec l'arrivée d'Axel.

Merci Chantal et Francis pour m'avoir accepté en tant que résident quasi-permanent lors de mes déplacements dans le Nord de la France (au dessus de Lyon).

Enfin, gardant le meilleur pour la fin, je remercie du fond du coeur Marie pour n'avoir cessé de croire en moi et avoir été présente dans les nombreux moments de doutes qui ont émaillés ce long travail de thèse. Merci d'avoir accepter mon choix, merci pour tout. A présent, et après ces années studieuses, j'espère de tout coeur partager avec toi bien plus que tout ce que nous avons vécu jusqu'ici.

Après tous ces remerciements, je vous souhaite une bonne lecture...

Lyon, le 13 décembre 2010

Résumé

Percevoir la parole dans le bruit représente une opération complexe pour notre système perceptif. Pour réaliser cette analyse de la scène auditive, nous mettons en place des mécanismes de ségrégation auditive. Nous pouvons également lire sur les lèvres pour améliorer notre compréhension de la parole. L'hypothèse initiale, présentée dans ce travail de thèse, est que ce bénéfice pourrait en partie reposer sur des interactions entre l'information visuelle et les mécanismes de ségrégation auditive. Les travaux réalisés montrent que lorsque la cohérence audiovisuelle est importante, les mécanismes de ségrégation précoce peuvent être renforcés. Les mécanismes de ségrégation tardive, quant à eux, font intervenir des processus attentionnels. Ces processus attentionnels pourraient être renforcés par la présence d'un indice visuel lié au stimulus auditif. Il apparaît que ce liage entre un flux de voyelles auditives, par exemple, et un indice visuel élémentaire est possible. Ce liage est renforcé lorsque l'indice visuel possède un contenu phonétique. Pour finir, les résultats présentés dans ce travail suggèrent que les mécanismes de ségrégation auditive puissent être influencés par un indice visuel pour peu que la cohérence audiovisuelle soit importante comme dans le cas de la parole.

Abstract

Perceive speech in noise is a complex operation for our perceptual system. To achieve this auditory scene analysis, we involve mechanisms of auditory streaming. We can also read lips to improve our understanding of speech. The initial hypothesis, presented in this thesis, is that visual benefit could be partly based on interactions between the visual input and the auditory streaming mechanisms. Studies conduced here show that when the audiovisual coherence is strong, primary streaming mechanisms can be strengthened. Late segregation mechanisms, meanwhile, may involve attentional processes. These attentional processes could therefore be strengthened by the presentation of a visual cue linked to the auditory signal. It appears that binding between a stream of vowels and a elementary visual cue can occur. This

binding is weaker than when the visual cue contained phonetic information. In conclusion, the results presented in this work suggest that the mechanisms of auditory streaming can be influenced by a visual cue as long as the audiovisual coherence is important as in the case of speech.

Mots clefs

Analyse de scènes auditives, Interactions audiovisuelles, Perception de la parole, Ségrégation auditive

Cette thèse a été réalisée au Laboratoire Neurosciences Sensorielles, Comportement et Cognition, UMR 5020, CNRS - Universtité Lyon 1 à Lyon ainsi qu'au Département Parole et Cognition du GIPSA-LAB, UMR 5216, Grenoble-INP - Université Stendhal, UJF à Grenoble. Cette thèse a été financée par le cluster 11 "Handicap, Vieillissement, Neuroscience" de la région Rhônes-Alpes Auvergne.

Table des matières

Introduction

Nous percevons la parole avec nos oreilles mais également avec nos yeux. La parole est par nature fondamentalement audiovisuelle. En effet, pour produire un son de parole, nous devons bouger nos lèvres. Cette relation entre le son et le mouvement des lèvres nous est très utile pour percevoir un signal de parole. En effet, nous pouvons améliorer notre compréhension de la parole dans le bruit si l'on exploite cette information visuelle. C'est ce que l'on appelle la lecture labiale. Chaque individu est capable d'utiliser cette aptitude avec plus ou moins de succès.

A présent, imaginez vous à une soirée animée avec un grand nombre de convives. Dans ce type de situation, il devient rapidement difficile de pouvoir converser avec notre interlocuteur même s'il est proche de nous. C'est l'effet *Cocktail Party* qui a été décrit en 1953 par Cherry. L'ensemble des sources acoustiques qui compose cette scène doit être analysé par notre système auditif. Pour percevoir distinctement ce que dit notre interlocuteur, nous pouvons utiliser notre capacité d'analyse de scènes auditives que l'on appelle également ségrégation auditive. Bregman en 1990 a proposé de transposer les connaissances acquises pour la perception visuelle et a ainsi formalisé ce thème de l'analyse de scènes auditives (ASA). Ainsi, l'ASA représente l'ensemble des mécanismes perceptifs et cognitifs qui nous permettent de séparer les différents événements acoustiques et de les regrouper pour former les sources sonores.

Par ailleurs, dans ces situations où les sons de parole se mélangent, Sumby et Pollack ont démontré en 1954 que, si l'auditeur pouvait observer le mouvement de lèvres de son interlocuteur, l'intelligibilité de son propos était améliorée de 40% environ. Le bénéfice, mis en évidence dans cette étude,

suggère que les informations acoustiques et visuelles peuvent alors interagir pour améliorer la compréhension de la parole. Quelques années plus tard, McGurk et MacDonald (1976) ont apporté une preuve supplémentaire en faveur de ce type d'interaction audiovisuelle pour la parole. Il s'agit de l'-effet *McGurk*. Si l'on présente une syllabe audiovisuelle composée par une syllabe /ba/ auditive et une syllabe /ga/ visuelle, l'auditeur rapporte dans la majorité des cas avoir perçu la syllabe /da/. Le /ba/ et le /ga/ ont fusionné pour former cette syllabe intermédiaire qu'est le /da/. La perception de cette syllabe audiovisuelle hybride a permis aux auteurs d'affirmer que les entrées auditives et visuelles de parole pouvaient interagir et venir perturber l'identification phonétique de l'événement audiovisuel de parole. Dès 1980, Massaro *et al.* ont proposé un modèle pour rendre compte de cette interaction que l'on peut supposer phonétique. C'est le modèle *FLMP* pour *Fuzzy Logical Model of Perception*. Ce modèle se décompose en plusieurs processus qui se déroulent séquentiellement. Appliqué à la parole, le premier processus consiste en un traitement distinct des stimuli visuels d'un côté et auditifs de l'autre. A l'issue de cette analyse, le modèle produit une représentation phonétique de chaque stimulus. Le processus suivant est un processus de comparaison entre la combinaison de ces représentations avec les représentations stockées en mémoire. Le dernier processus est un processus de décision. Cette décision repose sur le meilleur compromis entre les représentations stockées et les représentations générées à partir des stimuli. Dans l'effet *McGurk*, le /ba/ audio et le /ga/ visuel sont identifiés séparément. Ensuite, notre système choisit le meilleur compromis entre la réunion de ces représentations et les syllabes stockées en mémoire. Ainsi, la syllabe qui représente le mieux les deux stimuli est la syllabe /da/.

À part cet effet *McGurk*, les interactions entre des stimuli visuels et auditifs peuvent provoquer un autre type d'illusion. Si les stimuli visuels et auditifs ne sont pas localisés au même endroit dans l'espace, nous pouvons être trompés par l'effet ventriloque. Pick *et al.* (1969) ont démontré qu'il était possible de tromper notre perception de la position réelle de la source sonore dans l'espace en direction de la source visuelle. Par exemple, le son de voix de la marionnette que manipule le ventriloque semble effectivement provenir de

la marionette elle même. En réalité, on le sait, c'est le ventriloque qui parle. Les interactions entre l'information visuelle et auditive sont suffisamment fortes pour induire cette illusion perturbant la localisation spatiale. Dans ce cas de figure, l'effet n'induit pas d'erreur d'identification phonétique, nous parlerons alors d'interaction de nature pré-phonétique.

Dans une toute autre perspective de recherche, les méthodes d'imagerie cérébrale récentes comme l'électroencéphalographie ou la magnétoencéphalographie ont permis de mettre en évidence des corrélâts neuronaux associés à des interactions audiovisuelles pré-phonétiques. Sur le plan physiologique, ces interactions pré-phonétiques sont réputées apparaître précocément au cours du traitement des stimuli par le cerveau. Cette précocité peut également impliquer des structures corticales primitives. Ce terme de précocité recouvre une notion de précocité temporelle et structurelle (corticale). Calvert *et al.* (1997) et Pekkola *et al.* (2005) ont montré, par exemple, qu'il était possible de moduler l'activité du cortex auditif primaire en présentant des stimuli visuels. Le cortex auditif primaire est une structure dans laquelle se déroulent des mécanismes précoces de l'analyse. L'activation observée laisse donc penser qu'il existe des interactions précoces. De plus, certains des mécanismes de la ségrégation auditive sont également réputés être précoces. Micheyl *et al.* (2007) ont mis en évidence des corrélâts associés à l'organisation perceptive de séquences, i.e. l'état de ségrégation auditive dans le cortex auditif primaire. Pressnitzer *et al.* (2008) sont parvenus à mettre en évidence des corrélâts associés à la ségrégation dans les fibres nerveuses du noyau cochléaire. Les mêmes structures cérébrales semblent donc être impliquées dans les mécanismes de ségrégation et les interactions audiovisuelles. Ceci laisse entrevoir une nouvelle perspective pour expliquer le bénéfice de la lecture labiale. Ce bénéfice pourrait reposer en partie sur des interactions entre les mécanismes de ségrégation et la présence d'un indice visuel associé à la parole. Ce flux d'informations visuelles pourrait venir affecter ou moduler les mécanismes de ségrégation.

Ce manuscrit sera donc consacré à l'étude des interactions entre les mécanismes de ségrégation auditive et le signal visuel de parole. Dans le premier chapitre, nous présenterons les concepts et publications qui vont

nous permettre d'appréhender notre problématique. Le choix des publications présentées n'est pas exhaustif mais fait état des éléments qui, à nos yeux, apportent un éclairage pertinent vis-à-vis de cette thématique de recherche. Le choix est également restreint dans la mesure où ce thème est relativement vaste et recouvre des champs de recherche aussi variés que la psychoacoustique ou la neurophysiologie. Dans la première section de ce chapitre introductif (1.1), nous aborderons les interactions audiovisuelles dites préphonétiques et leur contribution dans les mécanismes de ségrégation auditive. Ensuite, les deux sections suivantes (1.2, 1.3) seront consacrées aux interactions audiovisuelles tardives. Les trois chapitres qui suivront, rédigés sous forme d'articles, présenteront les trois études qui ont été réalisées au cours de cette thèse. Chacune d'elle aborde un des thèmes développé dans les sections introductives. Enfin, le dernier chapitre de ce manuscrit propose un rappel des résultats rapportés dans chacune des études ainsi qu'une mise en perspective de ces résultats. Les perspectives de recherche pouvant être mises en place à court et moyen terme seront également présentées.

Chapitre 1

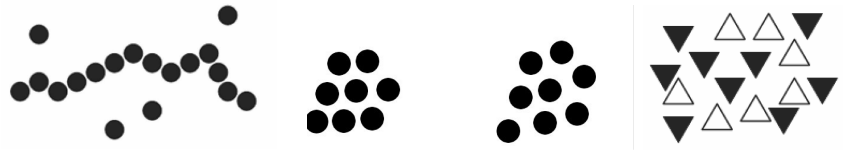
Analyse de scènes audiovisuelles

Le bénéfice de lecture labiale mis en évidence par Sumby et Pollack (1954) pourrait donc être en partie dû à des interactions entre le flux d'informations visuelles et les mécanismes de ségrégation auditive. Afin de présenter les phénomènes qui peuvent être impliqués dans ces interactions audiovisuelles, ce chapitre introductif sera scindé en trois sections. Les mécanismes de ségrégation auditive sont de deux natures selon van Noorden (1975). Il existe des mécanismes précoces et des mécanismes tardifs de ségrégation. Afin de respecter cette distinction, la première section (1.1) sera consacrée aux interactions entre le flux visuel et ces mécanismes précoces. Avant d'aborder les interactions audiovisuelles et les mécanismes de ségrégation tardive dans la troisième section (1.3), nous allons nous arrêter sur les mécanismes tardifs à proprement parler (section 1.2). En effet, un vif débat anime la communauté scientifique à propos de l'intervention des mécanismes liés à l'attention et leur incidence sur ces mécanismes de ségrégation. Nous aborderons ce thème pour la modalité auditive seule car, ensuite, nous intégrerons cet élément dans notre réflexion sur les interactions audiovisuelles.

1.1 Interactions audiovisuelles et ségrégation irrépressible

1.1.1 Analyse de scènes auditives

La mixture sonore qui atteint nos oreilles doit être interprétée par notre système auditif. Cette onde sonore complexe contient très souvent plusieurs sources sonores qui composent la scène auditive. Notre système auditif doit parvenir à séparer ces différentes sources afin de pouvoir les identifier. Cette opération est appelée Analyse de Scènes Auditives (ASA). Bregman (1990) a proposé de formaliser ce thème de l'analyse de scènes auditives en adaptant à la modalité auditive les connaissances établies pour la modalité visuelle. Koehler (1967) a quant à lui proposé une théorie : la théorie de la Forme (*Gestalt theorie*) énonçant les principes qui gouvernent notre capacité d'analyse de scènes visuelles. Certains de ces principes sont représentés dans la figure 1.1. Ces principes rendent compte de la manière dont notre système visuel regroupe les différents éléments pour former des objets visuels cohérents. Le signal visuel est un signal qui requiert un temps d'analyse important. Notre système doit intégrer différents éléments (ou primitives) comme la couleur, l'intensité, le mouvement ou encore l'orientation. Toutes ces primitives permettent d'établir des contours, des formes qui mènent à la formation d'objets. Notre cerveau dispose de différentes structures pour traiter ces éléments. Les primitives sont traitées par des structures relatives primaires. Des structures supérieures effectuent des traitements de plus en plus complexes. Treisman et Gelade (1980) ont proposé la théorie d'intégration des attributs (*Feature Integration Theory*) pour rendre compte de la manière dont notre cerveau lie ces primitives. Les principes de regroupement et le liage des primitives nous permet de réaliser l'analyse de la scène visuelle. L'image 1.2 représente une scène visuelle dans laquelle figure un dalmatien. La connaissance de la forme du dalmatien nous permet d'extraire le contour et ainsi la forme dans un contexte visuel complexe. Notre système visuel est parvenu à analyser la scène visuelle.



(a) Continuité (b) Proximité (c) Similarité

FIGURE 1.1 – Principes de Théorie de la Forme : continuité (a), proximité(b) et similarité (c)

FIGURE 1.2 – Scène visuelle complexe dans laquelle est représentée un dalmatien

Pour étudier les mécanismes de l'analyse de scènes auditives, Bregman (1990) a suggéré de distinguer deux familles de mécanismes. Selon lui, il est important de distinguer les mécanismes de ségrégation attentants aux événements purement simultanés des mécanismes attenants aux événements purement séquentiels. Dans la réalité, ces cas sont peu représentés. En effet, il existe généralement un recouvrement partiel entre les différentes sources acoustiques.

1.1.2 Ségrégation simultanée

La communauté scientifique qui s'est intéressée à l'analyse de scènes auditives a porté une grande partie de ces efforts pour appréhender les mécanismes de ségrégation simultanée. En effet, ces mécanismes étaient considérés comme reflétant le mieux notre capacité à comprendre la parole dans le bruit. Cependant dans l'absolu, la stricte superposition de deux sources au cours du temps est un fait relativement rare. Pour étudier la ségrégation simultanée, Darwin (1984) a proposé un paradigme devenu classique : le *paradigme des doubles voyelles*. Dans ces séries d'expériences, les auditeurs devaient identifier deux voyelles audio mixées ensemble et alignées temporellement. Ce paradigme a permis d'étudier le rôle d'indices acoustiques comme la fréquence fondamentale ou encore le contenu formantique (définition ci-dessous) (de Cheveigné, 1999). Meddis et Hewitt (1992) ont étudié le rôle de la fréquence fondamentale dans une tâche d'identification de doubles voyelles. La figure 1.3 montre les performances d'identification en fonction de la différence de fréquence fondamentale. Les performances d'identification sont optimales lorsque la différence de fréquence fondamentale atteint un demi-ton.

Le contenu fréquentiel de chaque voyelle nous permet de les distinguer les unes des autres. C'est ce que l'on appelle le contenu formantique. Chaque voyelle possède des pics d'énergie dans son spectre de fréquence, ces pics sont appelés les formants. Les deux premiers formants, appelés sobrement F1 et F2, suffisent pour discriminer les différentes voyelles. Chaque voyelle peut être placée sur un graphique dont les deux dimensions sont les formants F1 et F2 (figure 1.4). Les voyelles sont situées dans un triangle, le triangle formatique. Assmann et Summerfield (1989) ont étudié l'effet de manipulations du contenu formantique sur la ségrégation de doubles voyelles. Pour contrôler de manière optimale le contenu formantique des stimuli qu'ils ont utilisé, les voyelles ont été générées avec l'algorithme de Klatt (Klatt, 1980). Cet algorithme produit des voyelles avec trois formants contrôlés sans les fluctuations d'intensité que l'on trouve pour des voyelles naturelles. Les deux voyelles mixées, proposées dans cette étude, possédaient la même fréquence fonda-

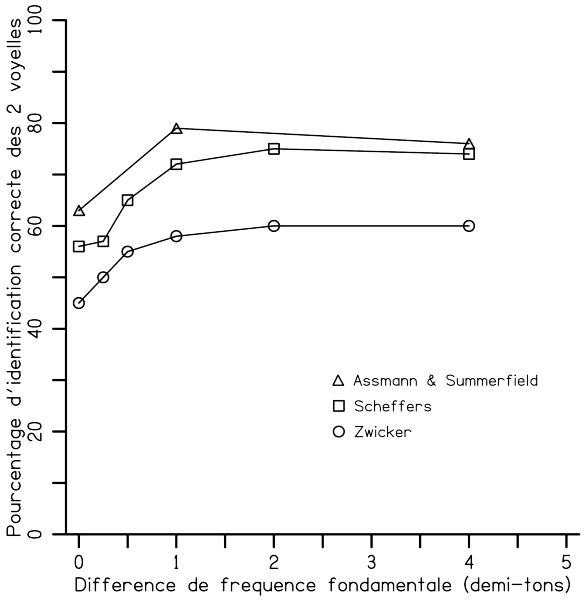


FIGURE 1.3 – Scores d'identification de doubles voyelles en fonction de la différence de fréquence fondamentale. D'après Meddis et Hewitt (1992) reprenant les résultats de Assmann et Summerfield (1990), Scheffers (1983) et Zwicker (1984)

mentale. En faisant varier le contenu formatique et uniquement celui-ci, les auteurs ont montré que le contenu formantique pouvait être un indice acoustique utile pour ségréger les deux voyelles. Les auteurs ont également montré que si les formants variaient conjointement au cours du temps, la ségrégation était renforcée (Assmann, 1994) (figure 1.5). L'étude des mécanismes de ségrégation simultanée a par ailleurs permis d'approfondir les connaissances concernant la perception de la hauteur de sons complexes (pour une revue cf. de Cheveigné (2005)).

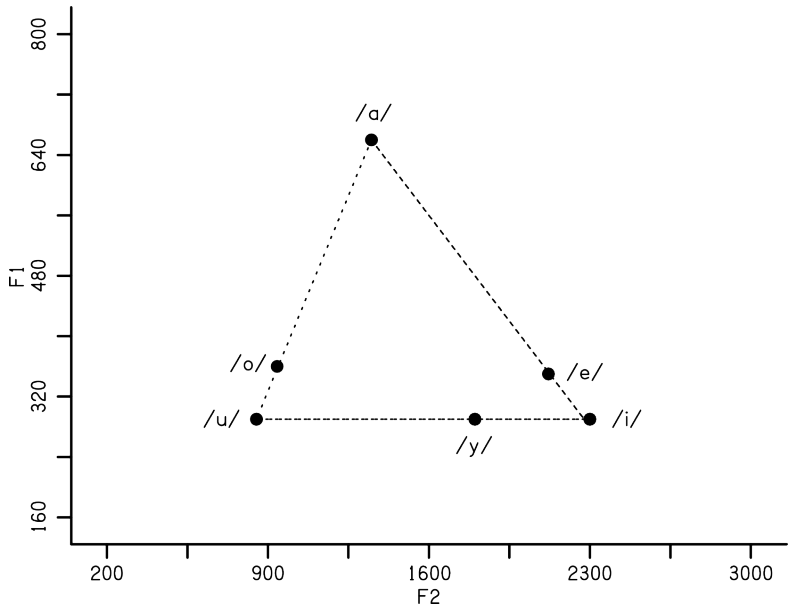


FIGURE 1.4 – Triangle formantique. Les voyelles /a/, /i/, /u/, situées aux sommets, correspondent aux mouvements articulatoires les plus "extrêmes"

1.1.3 Ségrégation séquentielle

Pour observer la ségrégation simultanée sur la base de la fréquence fondamentale, il faut introduire une différence d'un demi-ton. En revanche, pour observer la ségrégation séquentielle, il faut introduire une différence de plusieurs demi-tons. Les mécanismes de ségrégation semblent donc bien distincts. De plus, les mécanismes de ségrégation séquentielle pourraient refléter

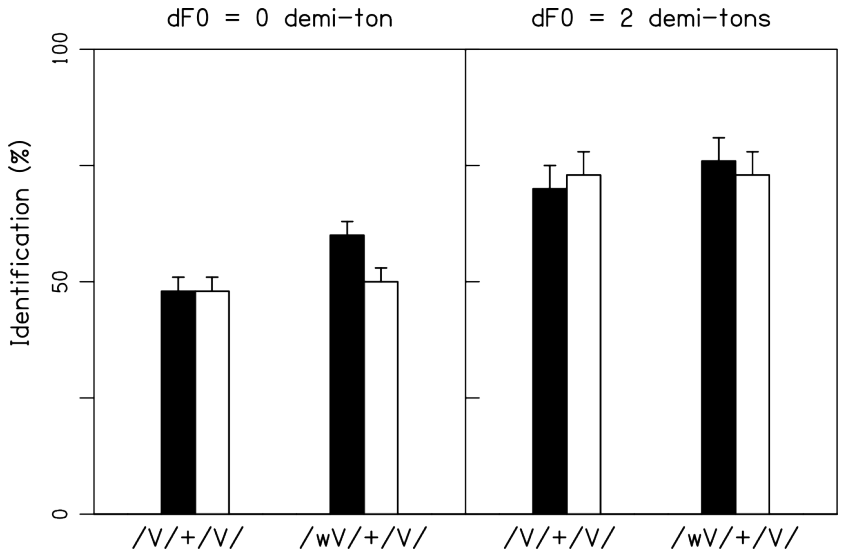


FIGURE 1.5 – Scores d'identification de doubles voyelles sans différence de fréquence fondamentale (panel de gauche) et avec une différence de deux demi-tons (panel de droite). Les barres /V/+/V/ représentent le mélange de deux voyelles dont le contenu formatique est stable. Les barres /V/+/wV/ représentent le mélange de deux voyelles dont les formants co-variaient dynamiquement

notre faculté de compréhension de la parole dans le bruit. En effet, une étude récente a établi une corrélation entre compréhension de la parole dans le bruit et la ségrégation séquentielle (Grimault et Gaudrain, 2006). La capacité de ségrégation simultanée viendrait, selon cette même hypothèse, compléter notre processus d'analyse de scènes auditives. Les études proposées dans ce manuscrit seront exclusivement consacrées aux mécanismes de ségrégation séquentielle.

En 1950, Miller et Heise se sont intéressés aux mécanismes de ségrégation séquentielle. Ils présentaient dans leur étude des séquences de sons purs. Ces séquences contenaient deux sons de fréquence différente (un son A et un son B) alternés de la manière suivante : A-B-A-B-... . La différence de fréquence pouvait induire la perception d'une seule séquence contenant les sons A et B ou bien de deux séquences, une de sons A et une de sons B. Grâce à ce paradigme, Miller et Heise ont définit un seuil ; le seuil de trille (*thrill threshold*). Ce seuil correspond à la différence de fréquence que l'on doit introduire pour induire la ségrégation des sons A et des sons B. En dessous du seuil, on perçoit l'alternance des sons A et B, donc on perçoit le trille. Au dessus de ce seuil, la séquence est ségrégée, on ne perçoit plus le trille.

Cette première étude a ouvert la voie et inspiré van Noorden (1975). En reprenant ce principe de construction des séquences, van Noorden a proposé, dans son manuscrit de thèse non publiée, une étude fondamentale pour la communauté scientifique qui s'est penchée sur les mécanismes de ségrégation auditive. Les séquences générées avaient la structure suivante : A-B-A-A-B-A-... , A et B étaient deux sons purs de fréquence variable. van Noorden a manipulé deux paramètres : la différence de fréquence entre les sons A et B et le rythme de présentation des sons. En fonction de ces deux paramètres, la séquence pouvait être alors perçue comme un seul flux avec un rythme de galop (instauré par la répétition du son A) ou comme deux séquences (séquence A et séquence B) ayant leur propre rythme (figure 1.6).

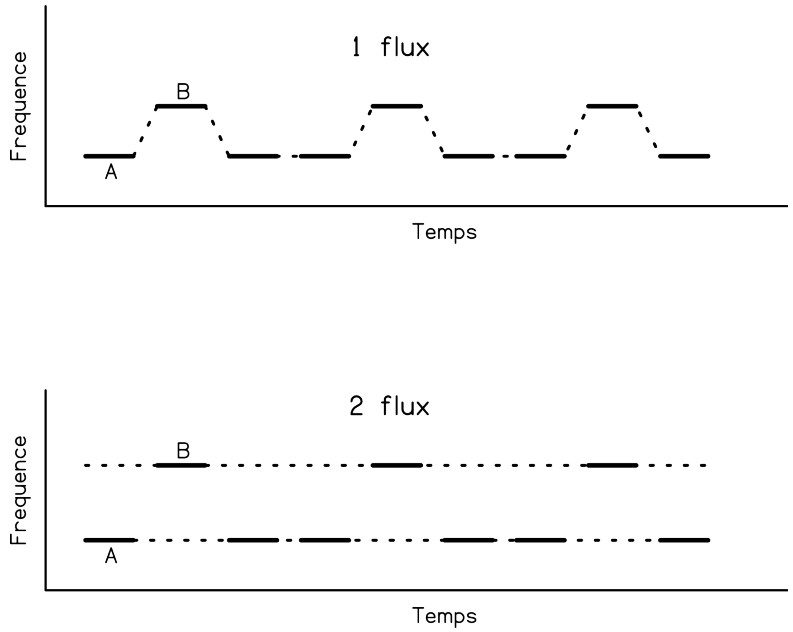


FIGURE 1.6 – Représentation schématique d'une séquence de sons purs alternant entre deux fréquences. Le panel du haut représente la séquence intégrée. Le panel du bas représente la séquence ségrégée

1.1.4 Mécanismes de la ségrégation

Grâce à ce paradigme, van Noorden a défini deux seuils gouvernant la perception des séquences. Le premier seuil est le seuil de cohérence temporelle (*TCB : Temporal Coherence Boundary*). En dessous de ce seuil, la séquence est perçue comme une séquence intégrée unique. Au dessus de ce seuil, il devient impossible de maintenir l'ensemble des sons dans une séquence unique. Cette séquence se scinde en deux de manière irrépressible. La différence de fréquence utilisée dans cette expérience permet d'induire cette ségrégation irrépressible. Ce caractère irrépressible permet, selon Bregman (1990), d'affirmer que les mécanismes impliqués dans cette situation reposent sur des différences d'attributs acoustiques des stimuli (Moore et Gockel, 2002). On parle alors de mécanismes orientés par les stimuli (*stimuli-driven*). van Noorden a montré également que ce seuil de cohérence temporelle était sensible au rythme de présentation de la séquence. Bregman *et al.* (2000) ont eux aussi confirmé l'effet du rythme de présentation sur ce seuil de cohérence.

L'élément déterminant, selon Bregman *et al.*, c'est l'intervalle qui sépare deux sons successifs que l'on appelle intervalle inter stimuli (ISI). Plus l'ISI est court, plus la ségrégation est importante. Un rythme de présentation rapide favorise donc la ségrégation irrépressible (figure 1.7).

Van Noorden définit à l'aide de ce paradigme un autre seuil, le seuil de fission (*FB : Fission boundary*). Ce terme peut prêter à confusion. La séquence est initialement perçue comme composée de deux séquences (A et B) distinctes. Lorsque l'on diminue la différence de fréquence entre les sons, et que l'on atteint ce seuil de fission, il devient impossible de percevoir les deux séquences A et B séparément. Pour résumé, au dessus du seuil, il y a fission et en dessous du seuil, il y a fusion. Ce seuil n'est pas sensible au rythme de présentation des sons. Selon l'auteur, l'estimation de ce seuil de fission permet d'accéder à un autre type de mécanisme. Ces mécanismes sont basés sur les schémas (*schema-based*). Si l'on y réfléchit de plus près, le fait de percevoir distinctement chaque séquence A et B permet d'en construire des représentations (des schémas) que l'on stocke en mémoire. Il devient alors difficile pour ces deux schémas de fusionner car nous avons acquis une connaissance relativement stable de chaque schéma. C'est d'ailleurs également pour cette raison que le seuil de fission est plus bas que le seuil de cohérence temporelle (figure 1.7). Ces travaux ont permis de mettre en évidence deux types de mécanismes impliqués dans l'analyse de scènes auditives : les mécanismes irrépressibles (automatiques) et les mécanismes basés sur les schémas.

Un dernier point fondamental qui caractérise les mécanismes de ségrégation séquentielle est ce que l'on nomme le phénomène de construction ou phénomène de *build-up*. Quelques secondes sont nécessaires lorsque l'on écoute une séquence pour que le percept se stabilise et que l'on perçoive la séquence comme intégrée ou bien ségrégée (Bregman, 1978; Anstis et Saida, 1985). Selon des termes probabilistes, les deux états peuvent être perçus à chaque instant. L'état stable est alors celui qui exprime la plus grande probabilité d'être perçu, l'autre état ayant malgré tout une probabilité non nulle d'être perçu. Raisonner ainsi permet d'expliquer le cas pour lequel le percept bascule d'un état à l'autre. Le fait d'avoir une probabilité non nulle pour l'autre état autorise la bascule vers celui-ci. On parle alors de bi-stabilité ou

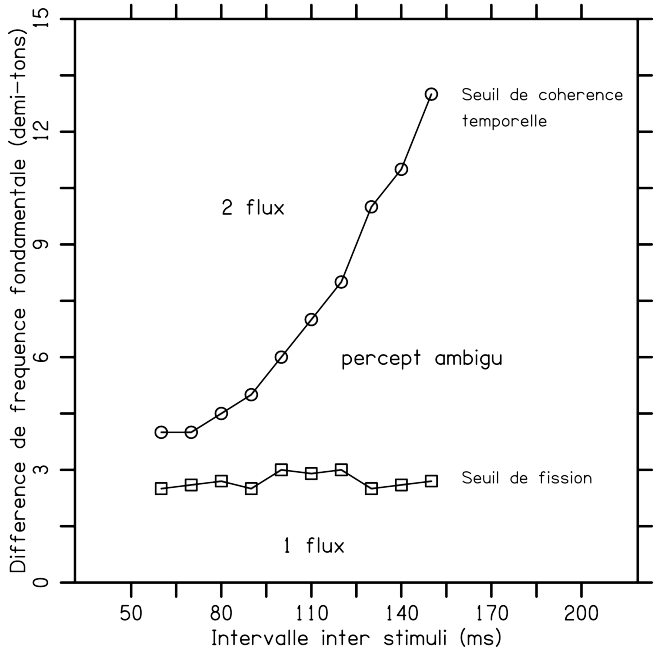
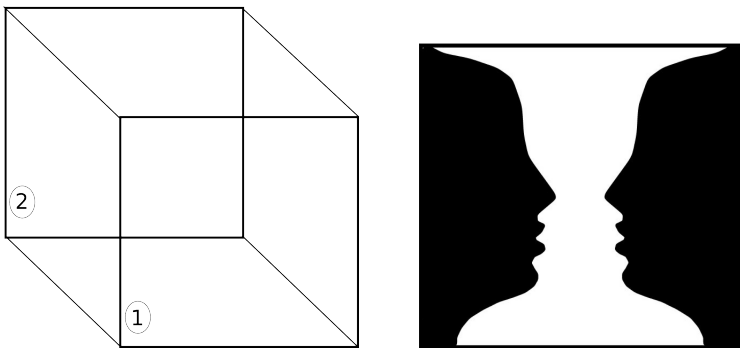


FIGURE 1.7 – Graphique représentant l'état de ségrégation en fonction du rythme de présentation de la séquence (en abscisse) et de la différence de fréquence (en ordonnée), d'après van Noorden (1975)

de multi-stabilité si plus de deux états peuvent être perçus. En observant la figure 1.7, on se rend compte qu'il existe une zone incluse entre le seuil de fission et le seuil de cohérence temporelle qui correspond à cette bistabilité. Dans cette zone, le percept est dit ambigu selon les termes de van Noorden. Il est possible, d'un point de vue expérimental d'induire ce type de bascule entre les états. Par exemple, on peut demander aux auditeurs de rapporter en temps réel la perception qu'ils ont d'une séquence présentée pendant plusieurs secondes. Après la présentation des séquences, on estime un ratio entre le temps pendant lequel ils ont perçu la séquence comme intégrée et le temps pendant lequel ils l'ont perçu comme ségrégée. Cette méthode subjective a l'avantage de ne pas contraindre l'auditeur à percevoir la séquence d'une manière ou d'une autre. En effet, si la tâche proposée contraint l'auditeur à intégrer la séquence, nous estimons le seuil de cohérence temporelle. En revanche, si la tâche contraint à ségréger la séquence, nous estimons le seuil de fission. Pour faire une analogie avec la perception d'une scène visuelle, la figure 1.8 représente des stimuli visuels bistables. Le cube de Necker peut être perçu comme ayant la face 1 ou la face 2 devant. Le vase de Rubin peut être perçu comme un vase ou bien comme deux visages de profil s'observant l'un l'autre.



(a) Cube de Necker (b) Vase de Rubin

FIGURE 1.8 – Exemples de figures bistables : Cube de Necker (a), Vase de Rubin (b)

1.1.5 Lecture labiale

Comme, nous l'avons mentionné plus haut dans l'introduction, la parole est audiovisuelle. Ainsi, pour améliorer notre compréhension de la parole dans le bruit nous pouvons exploiter l'information visuelle en plus des mécanismes de ségrégation purement auditifs.

En 1954, Sumby et Pollack ont réalisé une expérience dans laquelle ils proposaient aux participants des listes de mots. Ces listes de mots étaient présentées dans différentes conditions de bruit. Deux situations étaient comparées : une situation auditive seule et une situation dans laquelle les participants pouvaient voir la personne articuler les mots. Le bénéfice de cet indice visuel a été montré comme étant équivalent à une réduction du bruit de fond de 40% environ (figure 1.9). Ce bénéfice est appelé lecture labiale. Nous sommes tous plus ou moins habiles dans cette aptitude. Il existe une forte variabilité inter-individuelle. Ludman *et al.* (2000) avancent une hypothèse pour expliquer cette variabilité. Celle-ci serait dûe à des différences d'activations des aires corticales impliquées dans l'intégration de stimulations audiovisuelles.

Pour comprendre l'origine de cette variabilité dans notre capacité de lecture labiale, des travaux ont été réalisés auprès de différentes populations : des populations ayant des surdités développées avant l'acquisition du langage (pré-linguales), après l'acquisition du langage (post-linguales), des personnes dyslexiques et des personnes normo-entendantes. Mohammed *et al.* (2006) ont montré que la capacité de lecture labiale était corrélée à la faculté de lecture seule pour les personnes malentendantes et dyslexiques. Cette corrélation est absente chez les personnes normo-entendantes. Suh *et al.* (2009) ont comparé les performances de lecture labiale entre des personnes normo-entendantes, malentendantes pré-linguales et post-linguales. Par ordre de performances de lecture labiale des meilleures aux moins bonnes, les malentendants post-liguaux sont devant les normo-entendants qui sont eux même devant les malentendants pré-linguaux. Les auteurs ont suggéré que ces différences étaient dûes à l'expérience du langage, ce qui peut rejoindre l'aptitude de la lecture. Par ailleurs, ils ont montré que les performances étaient

liées à la latence d'activation du cortex auditif. Ils soulignent une corrélation positive entre la rapidité d'activation du cortex auditif et les performances de lecture labiale. Ludman *et al.* (2000) avaient également suggéré que cette variabilité de lecture labiale pouvait être la conséquence de modifications neurophysiologiques. Enfin, selon Rouger *et al.* (2007, 2008), il semblerait que les personnes atteintes par des pathologies auditives seraient de meilleurs intégrateurs audiovisuels que les personnes normo-entendantes. Retenons pour l'instant que nous sommes tous capable de tirer profit de l'information visuelle fournie par les mouvements de lèvres dans le but d'améliorer notre compréhension de la parole. Ceci semble d'autant plus important lorsque les conditions d'écoute sont défavorables.

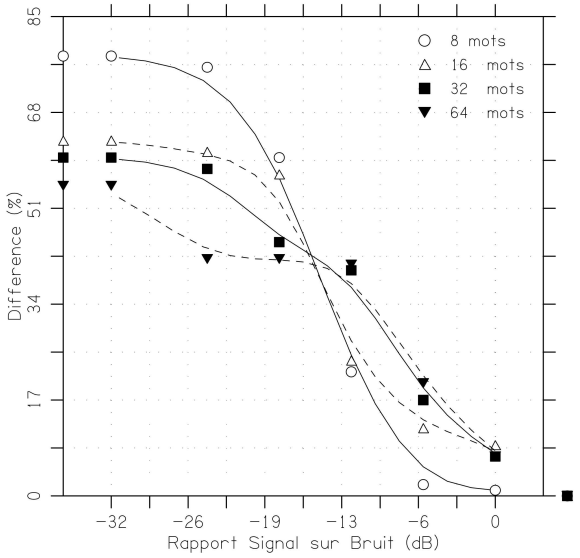


FIGURE 1.9 – Différences entre les scores d'intelligibilité pour la condition auditive seule et la condition audiovisuelle pour différents rapports de signal sur bruit

1.1.6 Détection de la parole dans le bruit

La lecture labiale améliore l'intelligibilité de la parole dans le bruit (Sumby et Pollack, 1954). En 2000, Grant et Seitz, par exemple, suggèrent que ce bénéfice de lecture labiale serait dû à une amélioration de la détection de la parole dans le bruit. Ainsi, comme la détection est plus facile, l'identification serait elle aussi facilitée. Plusieurs travaux ont été réalisés pour tester cette hypothèse (par. ex. Grant et Seitz, 2000; Grant, 2001; Kim et Davis, 2003). Grant et Seitz (2000) mettent en évidence que la détection de phrases prononcées dans le bruit est meilleure quand les participants peuvent observer le film montrant le locuteur en train de parler. Les auteurs parviennent également à montrer que ce bénéfice dépend du niveau de corrélation entre les fluctuations d'amplitude du signal auditif avec les variations de l'aire délimitée par les lèvres (l'aire aux lèvres). Pour approfondir cette hypothèse, Bernstein *et al.* (2004) ont proposé des indices visuels différents des vraies lèvres filmées. Les indices visuels conservaient cependant un niveau de corrélation avec le signal auditif. Les auteurs ont souhaité, en réalisant ces simplifications, accéder aux contraintes nécessaires que devaient partager les stimuli auditifs et visuels pour observer ce bénéfice de détection. Le type de simplification qui a été proposé dans cette étude a directement inspiré notre démarche pour la génération des signaux visuels de notre troisième étude (chapitre 4).

Bernstein *et al.* (2004) ont proposé deux stimuli auditifs présentés seuls ou doublés par un indice visuel. L'un de ces deux stimuli contenait une syllabe. Les participants devaient détecter la présence de la syllabe auditive. Quatre indices visuels étaient proposés (figure 1.10). L'indice visuel *AVS* (AudioVisual Speech) consistait en des mouvements de lèvres filmées. L'indice visuel *AVL* (AudioVisual Lissajous) était une aire délimitée par une courbe de Lissajous représentant l'aire aux lèvres. L'indice visuel AVR (AudioVisual Rectangle) était un rectangle dont l'extension verticale varie. Enfin, l'indice visuel AVSR (AudioVisual Steady Rectangle) était un rectangle qui apparaissait et disparaissait graduellement en fonction de l'intensité du signal auditif. La cohérence entre la variation de l'indice visuel et la stimulation auditive

était contrôlée pour toutes les conditions. La figure 1.11 représente les performances de détection en fonction de l'indice visuel exprimé sur une échelle de rapport signal sur bruit. Les performances de détection sont supérieures à toutes les autres conditions dans la condition *ASV* qui correspond aux vrais mouvements de lèvres. Les performances de détection sont également meilleures dans les conditions *AVL*, *AVR* et *AVSR* par rapport à la condition auditive seule (*AO*, Audio Only). Selon les auteurs, le bénéfice de détection dû à la présence de l'indice visuel est observé même si la corrélation entre les fluctuations d'amplitude du signal auditif et les variations de l'indice visuel n'est pas précise. Une corrélation *grossière* est suffisante.

Dans ce type d'expérience, on peut raisonnablement considérer que la tâche de détection ne nécessite pas l'identification phonétique des événements présents. Par conséquent, un bénéfice de détection dû à la présence d'indices visuels de parole peut être interprêté comme la conséquence d'interactions audiovisuelles non phonétiques. On parle dans ce cas d'interaction pré-phonétique. En respectant le cadre théorique suggéré par le modèle FLMP, nous devons cependant émettre l'hypothèse que des interactions pré-phonétiques puissent exister. Cette hypothèse n'invalide pas le modèle ; des interactions tardives phonétiques sont toujours possibles.

1.1.7 Corrélâts neurophysiologiques

La littérature concernant les corrélâts neuronaux des interactions entre les modalités visuelle et auditive est relativement conséquente. Ainsi, la section qui suit fait état de travaux récents ou plus anciens qui, selon nous, apportent des éléments clefs permettant d'étayer l'hypothèse selon laquelle les interactions entre ces modalités seraient relativement précoces. Cette précocité peut également être appuyée par les études sur la détection de la parole dans le bruit comme celles présentées ci-dessus.

Sams *et al.* (1991) ont enregistré l'activité magnétoencéphalographique (MEG) de l'hémisphère gauche (réputée traiter de manière privilégiée la parole) au cours de la perception de syllabes auditives ou visuelles de parole. Ils ont montré que le traitement du mouvement articulatoire visuel pouvait

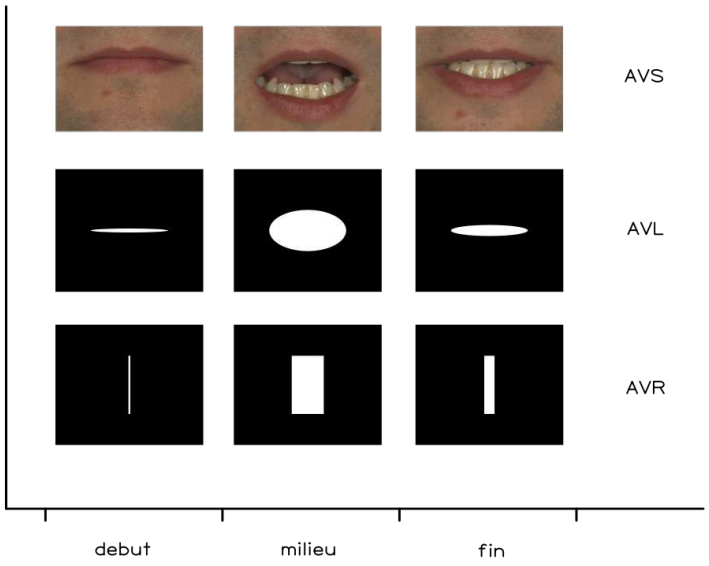


FIGURE 1.10 – Représentations conformes aux stimuli audiovisuels utilisés dans l'expérience d'après Bernstein *et al.*, 2004. Les acronymes signifient, pour AVS : Audio Visual Speech ; pour AVL : Audio Visual Lissajous et pour AVR : Audio Visual Rectangle

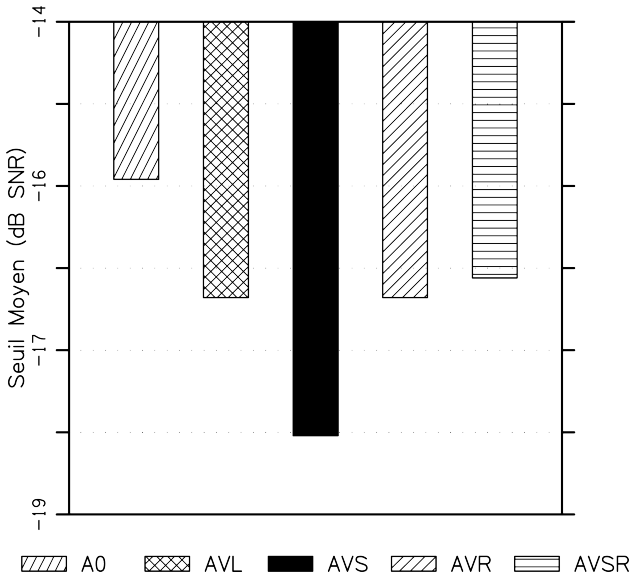


FIGURE 1.11 – Seuils moyens de détection d'une syllabe /ba/ prononcée dans le bruit en fonction de différents indices visuels

être réalisé dans le cortex auditif. De manière similaire, Moettoenen *et al.* (2002) ont observé qu'un changement dans le stimulus visuel pouvait activer le cortex auditif, cette fois-ci, de façon bilatérale. À noter également qu'ils ont rapporté une activation plus tardive en modalité visuelle seule qu'en modalité audiovisuelle. Selon eux, ceci suggère un effet d'interaction entre les modalités qui permettraient de traiter plus rapidement un changement visuel dans le cas d'un signal de parole. Calvert *et al.* (1997) vont encore plus loin en montrant que le cortex auditif primaire peut être activé en présentant des indices visuels langagiers seuls, i.e. sans stimulation auditive de parole. Cette activation du cortex auditif primaire n'est pas observée lorsque les stimuli visuels sont non langagiers (Pekkola *et al.*, 2005). L'activation du cortex auditif par une stimulation visuelle semble donc se produire uniquement lorsqu'il s'agit d'un signal visuel de parole. Giraud et Truy (2002) ont montré la dépendance inverse, à savoir qu'une activation des aires visuelles primaires pouvait se produire en présentant des sons de paroles. Cette sensibilité des aires primaires auditives et visuelles à des stimulations provenant d'une autre modalité et ce à un niveau relativement précoce soutient l'existence d'interactions audiovisuelles elles mêmes précoces.

Certains mécanismes de la ségrégation auditive ont également été démontrés comme étant précoces. Musacchia *et al.* (2006); Pressnitzer *et al.* (2008) sont parvenus, en enregistrant l'activité unitaire de fibres nerveuses du noyau cochléaire (figure 1.12), qui est une structure sous-corticale, à mettre en évidence des corrélâts représentant l'état de ségrégation auditive. À titre d'exemple, une étude réalisée sur les fibres nerveuses d'une structure similaire chez l'insecte a également révélé des corrélâts associés à l'état de ségrégation auditive (Schul et Sheridan, 2006). La ségrégation auditive apparaît donc comme une fonction élémentaire que nous pourrions partager avec des espèces beaucoup moins développées. Cette observation souligne le caractère précoce mais également automatique de la ségrégation auditive.

Les corrélâts neuronaux que nous avons mentionné ici, soutiennent l'hypothèse selon laquelle une partie du bénéfice de lecture labiale pourrait reposer sur une interaction audiovisuelle relativement précoce. Cette interaction viendrait donc renforcer la ségrégation auditive à un niveau précoce.

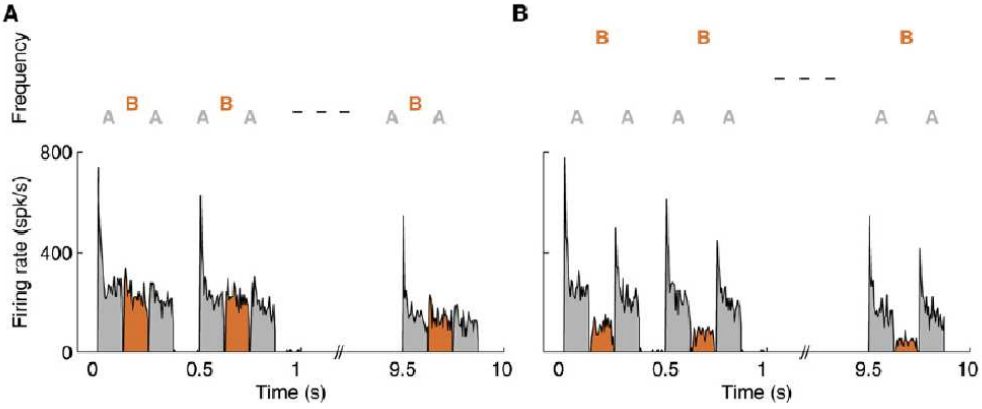
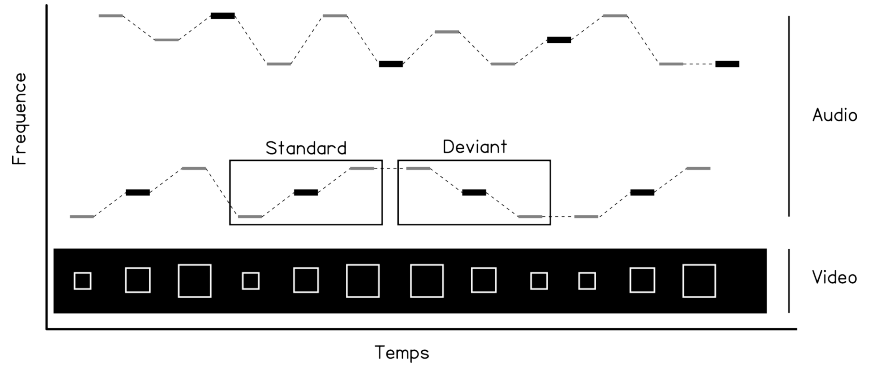


FIGURE 1.12 – Réponse unitaire d'une fibre nerveuse du noyau cochléaire chez le cobaye (à gauche, la réponse à une séquence intégrée, à droite la réponse à une séquence ségrégée) d'après Pressnitzer *et al.* (2008)

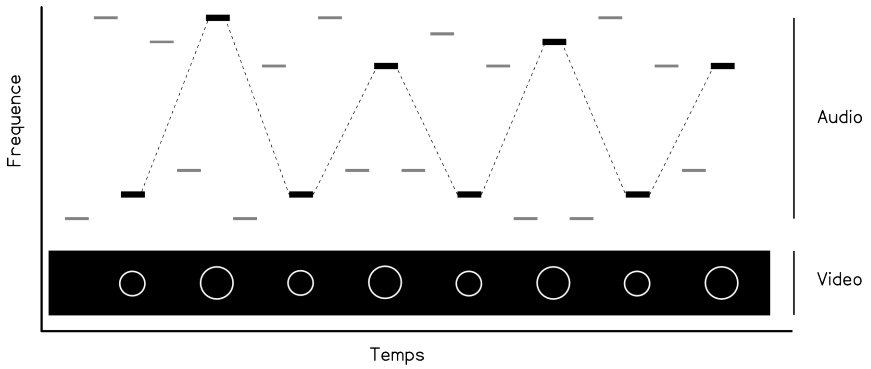
1.1.8 Influence d'un indice visuel sur la ségrégation irrépressible

Une seule série de travaux, à notre connaissance, a proposé un protocole expérimental permettant d'étudier l'influence d'un indice visuel sur les mécanismes de ségrégation auditive (Rahne *et al.*, 2007, 2008; Rahne et Böckmann-Barthel, 2009). Dans ces études, Rahne *et al.* ont enregistré les potentiels évoqués électroencéphalographiques (EEG) au cours de la perception de séquences de sons purs accompagnés simultanément par deux indices visuels élémentaires (cercles ou carrés). Les stimuli utilisés sont représentés dans la figure 1.13. Les stimuli auditifs sont des séquences composées de six sons purs dans deux gammes de fréquences différentes. La fréquence des sons alterne entre une valeur basse et une valeur haute. Les sons aigus sont joués dans un ordre aléatoire. Les sons graves forment un pattern ascendant de trois sons répétés en boucle. Ce pattern est appelé *pattern standard*. Parfois, ce pattern est remplacé par un pattern descendant. Ce pattern est alors appelé *pattern déviant*. Par ailleurs, un son sur trois pour l'ensemble de la séquence a une intensité de 15dB supérieure aux autres sons. La ségrégation auditive peut donc se faire sur la base de la différence de fréquence ou bien de l'intensité. Deux indices visuels élémentaires sont synchronisés soit avec le pattern

variant en fréquence (carrés) ou bien avec le pattern variant en intensité (cercles). Lorsque les auditeurs détectent ce pattern déviant, cela se traduit sur le plan EEG, par l'apparition d'une onde négative de dissemblance ou MMN (Mismatch Negativity). Cette onde est un corrélât neurophysiologique permettant d'affirmer que les auditeurs ont bien détecté le déviant et par conséquent sont parvenus à ségréger la séquence.



(a) Indice visuel calé sur la fréquence



(b) Indice visuel calé sur l'intensité

FIGURE 1.13 – Protocole expérimental de Rahne *et al.* (2007). Le panel du haut (a) représente la condition visuelle synchronisée avec la variation de la fréquence. Le panel du bas (b) représente la condition visuelle synchronisée avec la variation de l'intensité

Dans les deux études EEG, Rahne *et al.* (2007); Rahne et Böckmann-Barthel (2009) mettent en évidence une MMN quand l'indice visuel est synchronisé avec le pattern variant en fréquence, indiquant que l'indice visuel a permis de ségréger les sons graves des sons aigus. Ce protocole permet d'étudier uniquement la ségrégation sur la base de la fréquence. Pour

compléter ces études, Rahne *et al.* (2008) ont proposé une étude comportementale avec un matériel expérimental similaire. Ils demandent aux auditeurs d'indiquer quand ils perçoivent la séquence comme une séquence intégrée. Cette instruction permet d'évaluer le seuil de cohérence temporelle et d'accéder aux mécanismes de ségrégation irrépressible. En estimant le temps passé dans l'état intégré, les auteurs ne rapportent pas d'influence de l'indice visuel lorsque celui-ci est synchronisé avec le pattern variant en fréquence (figure 1.14). En revanche, ils montrent un effet de l'indice visuel synchronisé avec le pattern variant en intensité. Ces résultats permettent de suggérer que l'indice de fréquence est suffisamment fort pour induire un état de ségrégation stable. Cependant, lorsque la ségrégation repose sur un indice acoustique plus faible comme l'intensité, l'indice visuel peut venir renforcer la ségrégation.

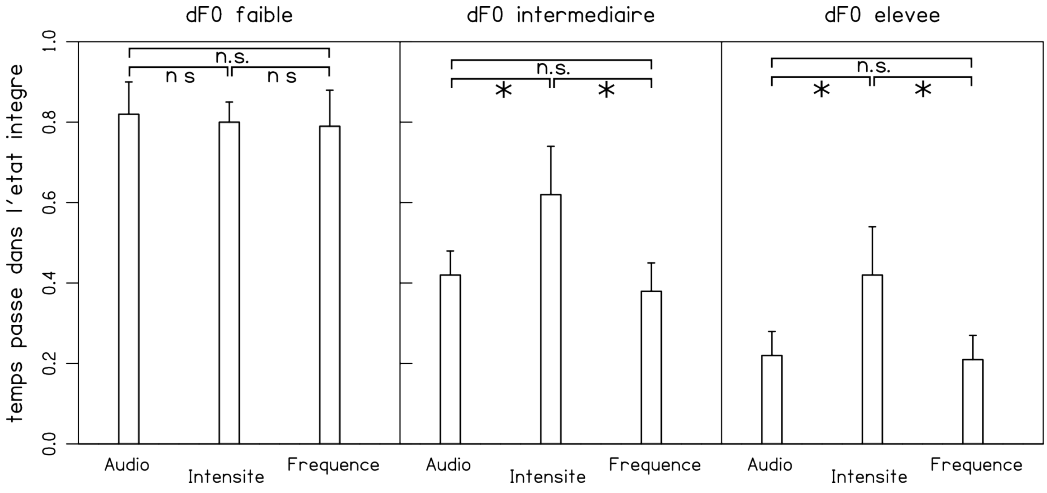


FIGURE 1.14 – De gauche à droite, chaque panel représente le temps passé dans l'état intégré pour une différence de fréquence entre les deux flux de la plus faible à la plus élevée. L'indice visuel est synchronisé soit avec la variation de fréquence (Frequence) soit avec la variation d'intensité (Intensite). Ces deux situations sont comparées avec la situation sans indice visuel (Audio). D'après Rahne *et al.* (2008)

Ces études montrent qu'il est possible d'influencer la ségrégation auditive précoce avec un indice visuel élémentaire comme une forme géométrique. L'étude 1 (chapitre 2) a été réalisée dans le but de tester cette influence

audiovisuelle dans un contexte plus écologique impliquant des signaux de parole.

1.2 Ségrégation basée sur les schémas

Dans la section précédente, nous nous sommes intéressés aux mécanismes précoces de la ségrégation auditive. À présent, nous allons aborder les mécanismes de ségrégation basée sur les schémas. Il s'agit, selon van Noorden (1975), du second type de mécanisme impliqué dans l'analyse de scènes auditives. Avant d'aborder la question des interactions audiovisuelles pour ces mécanismes de ségrégation, nous allons préciser la nature de ces mécanismes et développer le rôle de l'attention vis-à-vis de ces mécanismes. En effet, un vif débat persiste dans la communauté scientifique à propos de l'attention.

1.2.1 Ségrégation basée sur les schémas

Un schéma, selon Bregman (1990), est une représentation stockée en mémoire, comme peut l'être une mélodie qui nous est familière. Il s'agit ici de connaissances acquises sur le long terme. Selon Jones (1984), les schémas sont des représentations qui peuvent être construites au cours de la perception des stimuli. Les régularités acoustiques ou structurelles peuvent nous aider à créer ces représentations. En terme de processus, la définition d'un schéma pour Jones rejoint une description de type ascendante (*bottom-up*) tandis que celle de Bregman est d'avantage descendante (*top-down*).

Afin d'étudier les mécanismes de ségrégation basée sur les schémas Bregman et Rudnicky (1975) et Jones *et al.* (1981) ont proposé un paradigme de jugement d'ordre. Dans ce protocole, il s'agissait de juger l'ordre de présentation d'une paire de sons A et B ayant deux fréquences différentes. Cette paire de sons est tout d'abord présentée seule. Quelques secondes après, la paire est présentée dans une séquence construite de la manière suivante : C-C-C-E-A-B-E-C-C-C (figure 1.15). Les sons C (capteurs) et les sons E (encadrants) ont une fréquence constante différente des sons A et B. Bregman et Rudnicky (1975) ont manipulé la différence de fréquence entre les sons

C et E. Ils ont montré que les performances de jugement d'ordre des sons A et B sont meilleures quand la différence de fréquence entre les sons C et E est faible. Dans ce cas de figure, les sons E sons captés par les sons C, ce qui permet aux participants de juger plus facilement la paire A-B perçue isolément. En plus d'une différence de fréquence entre les sons C et E, Jones *et al.* (1981) ont introduit des différences de rythme de présentation. Trois configurations étaient proposées. Dans une condition, tous les sons étaient présentés sur le même rythme (C1). Dans une seconde condition (C2), les intervalles avant et après A-B étaient raccourcis : C—C—C—E-A-B-E—C—C—C, isolant le quadruplet E-A-B-E. Dans une dernière condition (C3), les intervalles avant et après les sons E étaient raccourcis : C—C—C-E-A-B-E-C—C—C. Les performances sont les meilleures pour les conditions C1 et C2 (table 1.1). Dans la condition C2, la rupture dans le rythme a favorisé la ségrégation et a facilité le jugement d'ordre pour la paire A-B. Dans la condition C1, Jones *et al.* (1981), ont suggéré que la régularité du rythme a favorisé la création d'attentes. Selon Jones *et al.*, ces attentes permettent de traiter plus efficacement le stimulus. Dans la condition C3, ni la rupture de rythme ni la régularité n'a pu permettre aux auditeurs de ségréger la séquence E-A-B-E des sons C et donc faciliter le jugement d'ordre.

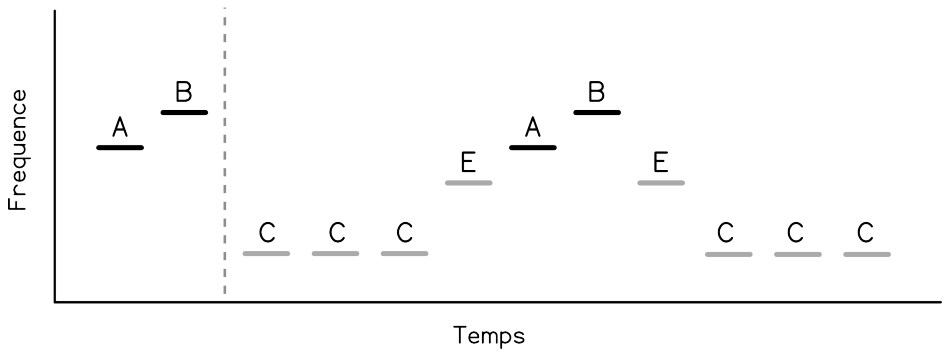


FIGURE 1.15 – Schéma du protocole expérimental proposé par Jones *et al.*, 1981

distance en F0			
condition	proche	eloigne	moyenne
C1	0.482	0.601*	0.542
C2	0.619*	0.601*	0.610
C3	0.531	0.537	0.534
contrôle	0.588*		0.588

TABLE 1.1 – Résultats de l'expérience de Jones *et al.*, 1981

1.2.2 Attention rythmique

Suite à cette série de travaux, Jones *et al.* (1981) ont proposé la théorie de l'attention dynamique (*Dynamic attention theory*). Selon cette théorie, nous serions capables d'extraire des régularités rythmiques présentes dans une séquence afin de développer des attentes. Ces attentes nous permettraient par la suite de traiter plus efficacement le stimulus attendu (Boltz, 1993; Schmuckler et Boltz, 1994). Pour développer ces attentes, notre système perceptif serait en mesure d'activer un oscillateur dont la fréquence d'oscillation serait entraînée par le rythme de présentation du stimulus. Le support physiologique de cet oscillateur importe peu pour notre réflexion. Jones *et al.* (2002) ont montré que cet oscillateur se mettait en place progressivement. Dans une étude réalisée en 2002, Jones *et al.* ont testé ce concept d'oscillateur. Les auteurs ont demandé aux participants de juger la relation fréquentielle entre un son de référence (*ref*) et un son *test* (figure 1.16). Ces deux sons étaient séparés par une séquence de 8 sons de fréquence aléatoire. Le son *ref* et les sons intercalés avant le son *test* étaient espacés de 600ms. L'intervalle entre le dernier son de la séquence et le son *test* pouvait prendre une valeur comprise entre 572ms et 676ms. Les performances de jugement entre les sons *ref* et *test* diminuent à mesure que l'intervalle s'écarte de la valeur centrale de 600ms correspondant au rythme de la séquence intercalée (figure 1.17). Les résultats suggèrent que le jugement de hauteur est facilité si l'événement apparaît à un instant attendu. Cette capacité à produire un oscillateur attentionnel calé sur les régularités des signaux pourrait évoluer avec l'âge et l'expérience musicale (Drake *et al.*, 2000).

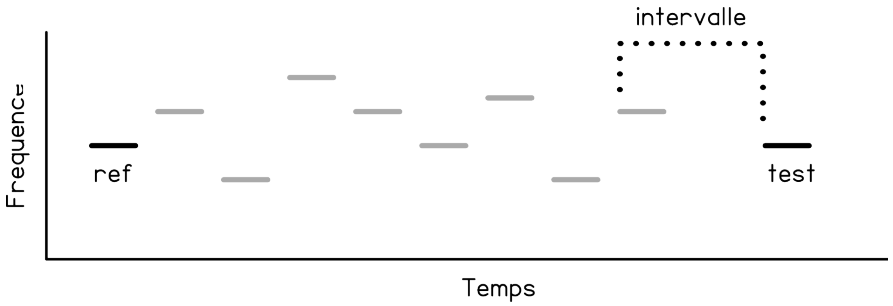


FIGURE 1.16 – Schéma du protocole expérimental proposé par Jones *et al.*, 2002

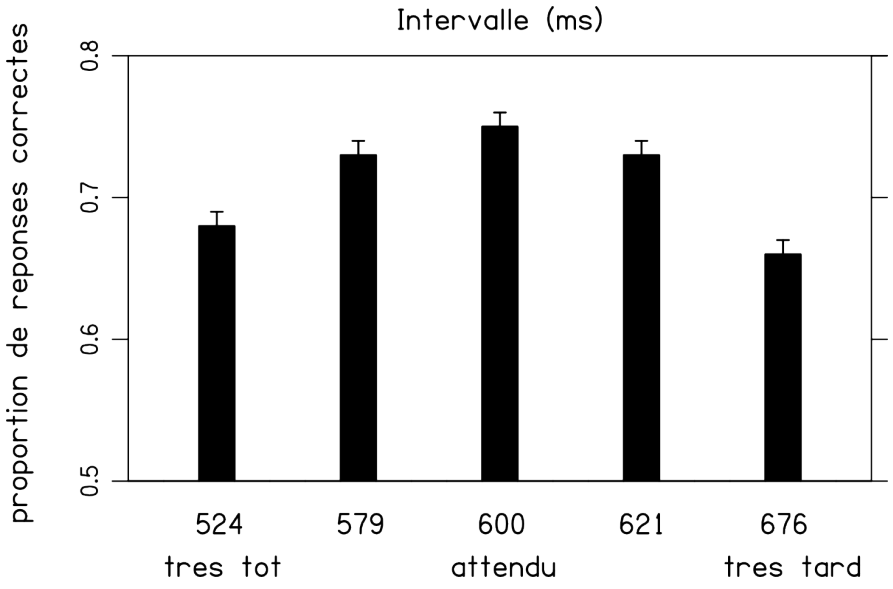


FIGURE 1.17 – Résultats de l'expérience de Jones *et al.*, 2002

1.2.3 Attention et ségrégation auditive

Le développement d'attentes calées sur le rythme de présentation des séquences nous permet de suggérer que notre attention peut être captée par des propriétés acoustiques du stimulus. Ces processus attentionnels pourraient donc influencer la ségrégation auditive. La contribution de l'attention dans les mécanismes de ségrégation reste cependant controversée.

Bregman (1978) a démontré que le processus de mise en flux (*le build-up*) est un processus qui dure quelques secondes. Durant ces premières secondes, le percept n'est pas stabilisé. Une fois le *build-up* achevé, le percept devient stable. En 2001, Carlyon *et al.* ont étudié l'interaction entre l'attention et ce *build-up*. Les auteurs ont proposé aux auditeurs un matériel sonore différent dans chaque oreille. Une séquence sonore susceptible d'être ségrégée ou intégrée était présentée dans une oreille. En détournant l'attention des participants vers le stimulus auditif présenté dans l'autre oreille avec une tâche secondaire, les auteurs ont montré que le processus de construction avait été réinitialisé. Selon eux, le *build-up* requiert donc de l'attention. Cette hypothèse a été étayée par une étude de Cusack *et al.* (2004). En interrompant le focus attentionnel pendant de courtes périodes, les auteurs sont parvenus à ré-initialiser le *build-up*. Ces premiers résultats ont été remis en cause plus tard par Macken *et al.* (2003). Selon Macken *et al.*, le protocole proposé par Carlyon *et al.* (2001) ne permettait pas d'exclure l'existence de mécanismes de ségrégation automatique indépendants de l'attention. Pour appuyer cette hypothèse, Sussman *et al.* (2007) ont proposé une étude EEG dans laquelle ils ont utilisé l'onde MMN comme marqueur de la ségrégation auditive. Cette onde observable sur des tracés électro-encéphalographiques à permis de révéler que même en l'absence de focus attentionnel dirigé vers les stimuli, la ségrégation pouvait se produire. Ainsi et selon cette observation, certains mécanismes de la ségrégation seraient non attentionnels.

1.2.4 Théories de l'attention

A la lumière des travaux mentionnés ci-dessus, il existerait donc deux mécanismes : un mécanisme de ségrégation automatique et un mécanisme attentionnel permettant de réaliser de la ségrégation (Alain et Arnott, 2000; Snyder *et al.*, 2006). Alain et Arnott (2000) parlent d'attention sélective. Ils précisent en 2008 que cette attention sélective pourrait être gouvernée par un premier mécanisme sensoriel qui améliore le traitement de l'information pertinente vis-à-vis de la tâche et un second mécanisme sensoriel qui atténue le traitement de l'information non pertinente. Botte *et al.* (1997) ont réalisé

une expérience dans laquelle ils montrent que le niveau sonore perçu de l'information non pertinente est moins fort que celui de l'information pertinente. Ceci vient étayer l'hypothèse selon laquelle il existerait bien un mécanisme attentionnel d'atténuation.

Fritz *et al.* (2007) ont également proposé une théorie similaire basée elle aussi sur deux mécanismes. Un premier mécanisme serait basé sur les propriétés acoustiques du signal. Les attributs de ce signal captent notre focus attentionnel et nous permettent de traiter l'information pertinente. Un second mécanisme reposerait sur l'utilisation de schémas nous permettant ainsi de sélectionner l'information pertinente. Les données présentées dans l'étude de Tan *et al.* (2008) viennent soutenir la théorie proposée par Fritz *et al.*. De plus, des études neurophysiologiques récentes (Chait *et al.*, 2010; Münte *et al.*, 2010; Snyder *et al.*, 2006) mettent en évidence des corrélâts neurophysiologiques cohérents avec les théories de Alain et Bernstein; Fritz *et al.*.

Pour résumé, il y aurait donc des mécanismes de ségrégation auditive purement automatiques et non attentionnels d'une part. D'autre part, il existerait deux mécanismes attentionnels. Le premier serait un mécanisme qui améliorerait le traitement de l'information pertinente qui peut être perceptive (timbre, hauteur) ou cognitive (schémas, connaissances). Le second mécanisme serait un mécanisme de suppression reposant sur nos connaissances.

1.2.5 Attention et ségrégation basée sur les schémas

Nous disposons à présent d'un cadre théorique supportant les processus attentionnels dans lequel les connaissances et les régularités perceptives des stimuli auraient un rôle à jouer vis-à-vis des mécanismes de ségrégation auditive. Afin de tester cette hypothèse, nous avons réalisé une seconde étude (chapitre 3). Dans cette étude, nous avons utilisé le paradigme des mélodies intercalées. Ce paradigme est classiquement utilisé en musique pour évaluer les effets d'apprentissage et d'utilisation de nos connaissances pour identifier des mélodies familières intercalées note à note avec des mélodies distractrices

inconnues.

Dowling *et al.* (1987) ont réalisé une expérience de mélodies intercalées (figure 1.18). Huit mélodies familières étaient présentées aux participants. Ces mélodies étaient ensuite présentées dans des séquences intercalant note à note les mélodies apprises et des mélodies inconnues. Dans cette expérience, les auteurs ont manipulé la différence entre les empans fréquentiels de la mélodie familière et de la mélodie distractrice. Dans une condition, les deux mélodies avaient le même empan. Dans une seconde condition, les deux mélodies avaient des empans différents. De plus, la séquence commençait soit par la note de la mélodie familière (*on-beat*) ou par la note de la mélodie distractrice (*off-beat*). La proportion de réponses correctes est meilleure quand les mélodies n'ont pas le même empan fréquentiel (89% contre 77.5% d'identification correcte). Cette différence perceptive permet de ségréger plus facilement les notes de la mélodie cible des notes de la mélodie distractrice. De plus, les performances sont meilleures pour la condition *on-beat* par rapport à la condition *off-beat* (87% contre 75.5%). Les participants captent la première note de la mélodie et enclenchent probablement un cycle attentionnel calé sur cette première note (Jones *et al.*, 1981). Les auteurs soulignent également que les performances sont meilleures que la chance même lorsque les mélodies partagent le même empan. La connaissance de la mélodie a permis aux participants de sélectionner les notes de la mélodie familière à extraire (Alain et Bernstein, 2008). Dans cette expérience, toutes les mélodies étaient présentées sur un rythme régulier et constant. Ceci peut constituer un indice perceptif suffisant pour faciliter le processus de reconnaissance des mélodies. Par ailleurs, Endress (2010) a montré qu'il était plus facile de suivre un contour mélodique défini par des notes non adjacentes quand celles-ci respectaient une règle tonale précise (comme les mélodies familières) que quand elles ne respectaient pas de règle (comme les mélodies distractrices). Ceci peut expliquer pourquoi, le pourcentage d'identification reste élevé malgré le fait d'avoir modifié le rythme original des mélodies.

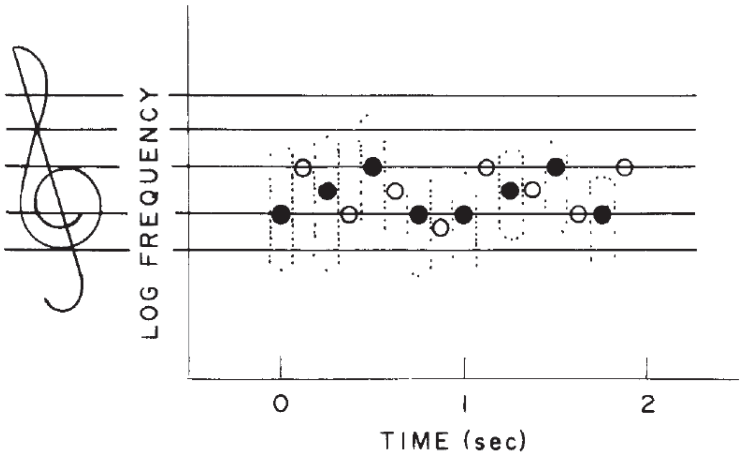


FIGURE 1.18 – Illustration de séquences intercalées d'après Dowling *et al.*, 1987

Bey (1999) a également suggéré que la force de la trace mnésique de la mélodie était déterminante pour réussir la tâche d'identification de mélodies intercalées. En effet, en présentant des nouvelles mélodies à apprendre juste avant d'effectuer la tâche d'identification, Bey et McAdams (2002) ne sont pas parvenus à répliquer les résultats obtenus par Dowling *et al.* (1987).

1.3 Interactions audiovisuelles et ségrégation basée sur les schémas

Nous avons vu dans la section précédente que l'attention pouvait vraisemblablement venir influencer les mécanismes de ségrégation auditive. Dans la suite de ce manuscrit, nous allons nous intéresser aux interactions entre la modalité visuelle et la modalité auditive. Nous pouvons alors penser que l'attention pourrait jouer un rôle dans les processus d'interactions audiovisuelles. Dans un premier temps, nous allons aborder la question de l'attention audiovisuelle. Ensuite, nous nous intéresserons aux interactions audiovisuelles à proprement parlé.

1.3.1 Attention et interactions audiovisuelles

La scène audiovisuelle que nous devons analyser à chaque instant est une scène complexe. Il est impossible de traiter simultanément l'ensemble des stimulations qui composent cette scène. Lavie (1995) a suggéré que nos ressources attentionnelles étaient limitées. Le fait de disposer de capacités limitées entraîne des effets d'interaction lorsque nous devons réaliser plusieurs tâches en même temps. La quantité d'information à traiter apparaît donc comme un facteur déterminant vis-à-vis de l'attention. Cela est d'autant plus vrai lorsque l'on parle d'attention audiovisuelle. Lavie montre également que lorsque nos ressources attentionnelles ne sont pas saturées, on observe des interactions au cours de l'éxécution des tâches. En revanche, quand nos ressources sont saturées, les interactions disparaissent. Enfin, il montre que même en demandant explicitement aux participants d'ignorer une partie des stimulations, celles-ci sont, malgré tout, traitées par notre système perceptif. Plus tard, Alais *et al.* (2006) ont proposé que nos ressources attentionnelles seraient réparties entre plusieurs sous-systèmes. Chaque sous-système serait dédié à une modalité. Cette séparation permettrait de maintenir l'indépendance des traitements. Selon cette proposition, nous pourrions traiter strictement simultanément une stimulation auditive et une stimulation visuelle par exemple. Ce point de vue est discuté par Spence *et al.* (2001) et Töllner *et al.* (2009). Selon eux nous ne pourrions traiter qu'une information à chaque instant. Le jeu consisterait ensuite à effectuer des bascules attentionnelles rapides entre les modalités. Ainsi, il apparaît que nous serions capables de traiter de manière attentionnelle (Spence *et al.*, 2001) ou non attentionnelle (Lavie, 1995) des stimuli complexes, comme les stimuli audiovisuels. Dans cette perspective, devons-nous porter notre attention pour que des interactions entre les stimuli visuels et auditifs se produisent ? L'intégration audiovisuelle requiert-elle notre attention ?

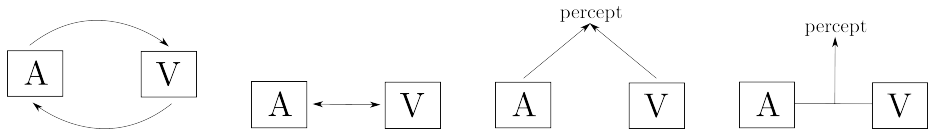
Terminologie

Avant de continuer, nous allons préciser dans ce paragraphe les différences que nous entendons entre les concepts d'interaction, de liage,

d'intégration et de fusion audiovisuelle. En effet, dans la littérature les mêmes termes sont employés dans des contextes différents, il est donc parfois difficile de s'y retrouver.

Le terme d'*interaction* audiovisuelle désignera les phénomènes que nous avons rapporté dans la section 1.1. Un indice visuel peut venir affecter des mécanismes réputés auditifs. Le terme *liage* désignera les situations pour lesquelles les stimuli de différentes modalités seront perçus comme émanant d'une seule et même source. Le liage pourra être considéré comme perceptif et/ou comme tardif (cognitif). Le terme *intégration* désignera l'ensemble des processus physiologiques et cognitifs conduisant à créer une représentation multimodale ou amodale (Hasson *et al.*, 2007) des stimuli auditifs et visuels formant l'objet perçu. Cela sous-entend que même si des structures primitives sont impliquées dans le processus d'intégration, l'intégration elle même suppose un certain niveau d'abstraction par rapport à ce que pourrait être une interaction de bas niveau. Enfin, le terme de *fusion* désignera un état perceptif ou cognitif dans lequel les différentes modalités perceptives sont associées très fortement entre elles pouvant conduire à des illusions perceptives comme l'effet de ventriloquisme ou bien l'effet McGurk. Il est donc sous-entendu que l'on ne peut plus accéder aux différentes entrées sensorielles quand la fusion audiovisuelle a pris place.

A présent, nous proposons quatre représentations schématiques illustrant les différents concepts (figure 1.19). Il ne s'agit là que d'une proposition pour clarifier le paragraphe précédent.



(a) Interaction (b) Liage (c) Intégration (d) Fusion

FIGURE 1.19 – Représentation schématique des concepts associés à la formation d'objets audiovisuels

Processus attentionnels et niveaux de traitement

Selon deux études récentes (Navarra *et al.*, 2010; Koelewijn *et al.*, 2010), l'attention pourrait affecter l'intégration audiovisuelle à différents niveaux de traitement. Par exemple, les travaux réalisés en neuro-imagerie de (Jäncke *et al.*, 1999) suggèrent que l'attention pourrait moduler l'activité des aires auditives primaires. Poghosyan et Ioannides (2008) mettent en évidence une modulation de l'activité similaire dans les aires visuelles primaires. Tiippana *et al.* (2001) suggèrent eux aussi que le traitement unimodal pourrait être modulé par l'attention.

Treisman et Gelade en 1980 avaient proposé quelques années plus tôt une théorie ; la théorie d'intégration des attributs (*Feature Integration Theory*). Cette théorie rend compte de la manière dont nous intégrons les différents attributs visuels pour former un objet visuel. Treisman et Gelade suggèrent dans cette théorie que certains traits visuels comme la forme ou la couleur pourraient être liés sans porter notre attention. Pour des traitements relativement primaires, l'attention n'est donc pas nécessaire selon les auteurs. Cette hypothèse rejoint l'hypothèse similaire avancée par Lavie (1995).

Le liage entre les stimuli de différentes modalités serait quant à lui tributaire de notre attention (Alsius *et al.*, 2005; Fairhall et Macaluso, 2009). Alsius *et al.* (2005) ont mis en évidence une interaction entre le niveau d'attention et la force du liage audiovisuel, en engagant l'attention des participants dans une tâche secondaire au cours de perception d'un matériel audiovisuel. Les auteurs présentent des syllabes de type McGurk. L'effet McGurk est alors réduit si les participants réalisent une tâche secondaire. Dans une autre expérience Fairhall et Macaluso (2009) ont étudié ces effets d'interaction entre attention et intégration audiovisuelle. Deux haut-parleurs et deux écrans étaient disposés devant les participants. Ces derniers devaient associer un flux sonore avec l'un des deux films présenté sur chaque écran. Les deux films étaient présentés dans des champs visuels distincts (i.e. droit ou gauche). L'activité corticale des participants était enregistrée au cours de l'expérience. Cette tâche implique l'attention spatiale. Fairhall et Macaluso ont observé que les aires dédiées à l'intégration audiovisuelle étaient d'avantage activées

lorsque l'attention était portée sur les stimuli à intégrer. L'intégration audiovisuelle semble donc impliquer l'attention.

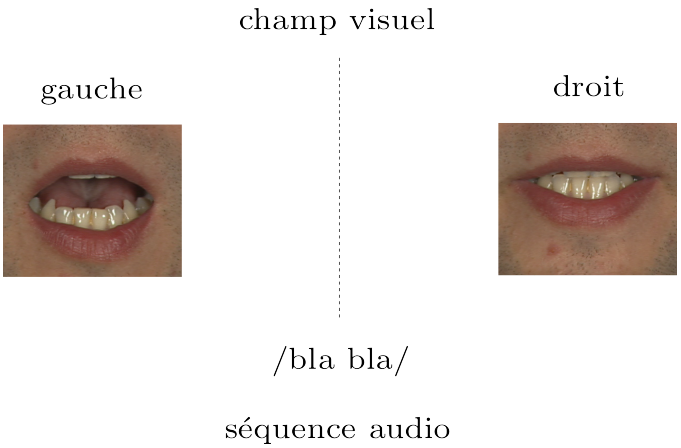


FIGURE 1.20 – Protocole expérimental proposé par Fairhall et Macaluso (2009)

Pour compléter cette discussion concernant l'implication de l'attention dans les processus d'intégration audiovisuelle, Scholl (2001) ont proposé une théorie, la théorie de l'attention basée sur les objets (*Object based theory*). L'attention audiovisuelle pourrait être orientée par ce que les auteurs appellent des objets. Ainsi, si l'on parvient à former à partir des stimuli un objet audiovisuel cohérent, notre attention peut être captée par celui-ci. Best *et al.* (2008) sont parvenus à une conclusion similaire avec un matériel uniquement sonore. Si les stimuli auditifs, utilisés par Best *et al.*, respectaient le principe de continuité (théorie de la Forme), l'attention des participants était alors orientée vers ces objets. La théorie de l'attention basée sur les objets suggère donc que l'attention est un processus qui peut être orienté par une représentation du stimulus auditif ou audiovisuel. Ce qui est important, c'est la possibilité de créer cet objet. Santangelo et Spence (2007) ont observé, que même si la tâche requiert d'importants traitements (appelée aussi charge cognitive), l'objet audiovisuel pouvait capter notre attention. La cohérence de cet objet audiovisuel apparaît donc comme un facteur déterminant dans les processus attentionnels.

L'objet audiovisuel semble, par ailleurs, bénéficier d'un traitement privilégié de la part de notre système perceptif. C'est le phénomène de *super-additivité*. Lorsque notre système perceptif est exposé à des stimuli audiovisuels, l'activité corticale associée est plus importante que si l'on somme l'activité associée au traitement du stimulus visuel et du stimulus auditif pris indépendamment. Notre système perceptif traite plus efficacement un stimulus audiovisuel qu'un stimulus unimodal. De plus, selon Besle *et al.* (2004), la présentation d'un stimulus visuel faciliterait le traitement du stimulus auditif associé. Selon van Wassenhove *et al.* (2005), un mouvement visuel de lèvres permettrait de préparer le traitement de la cible audio à venir. Enfin, ce phénomène de super-additivité ne se produirait que si l'on porte l'attention sur les deux modalités (Talsma *et al.*, 2007; Talsma et Woldorff, 2005).

Pour résumé, l'attention semble jouer un rôle important dans le processus d'intégration audiovisuelle. Une revue récente proposée par Talsma *et al.* (2010) soutient cette théorie. Nous proposons ici un modèle de déroulement temporel des processus d'intégration audiovisuelle qui tient compte de l'attention. Tout d'abord, l'intégration des traits perceptifs élémentaires d'une même modalité se fait de manière non attentionnelle. La salience perceptive d'une combinaison de traits capte notre attention. Une fois notre attention captée, l'intégration audiovisuelle de plus haut niveau peut se mettre en place. Ensuite, cette attention de plus haut niveau mise en œuvre, il est possible de créer un objet audiovisuel. La constitution de cet objet rend possible son utilisation pour anticiper ou *super-ajouter* les modalités entre elles et maintenir notre focus attentionnel. Cet objet audiovisuel peut être ensuite être ré-injecter dans les aires primaires (Watkins *et al.*, 2007).

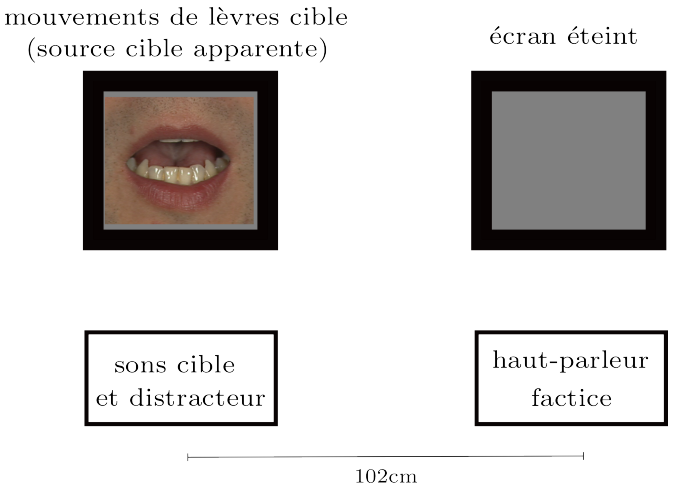
1.3.2 Liage perceptif

À présent, laissons l'attention de côté. Nous allons présenter, dans ce qui suit, le concept de liage audiovisuel. Les stimuli issus de différentes modalités doivent être liés entre eux pour que nous puissions créer une représentation cohérente de l'objet audiovisuel auquel nous sommes confrontés. Il est important de distinguer le liage perceptif du liage cognitif. Le liage perceptif, que

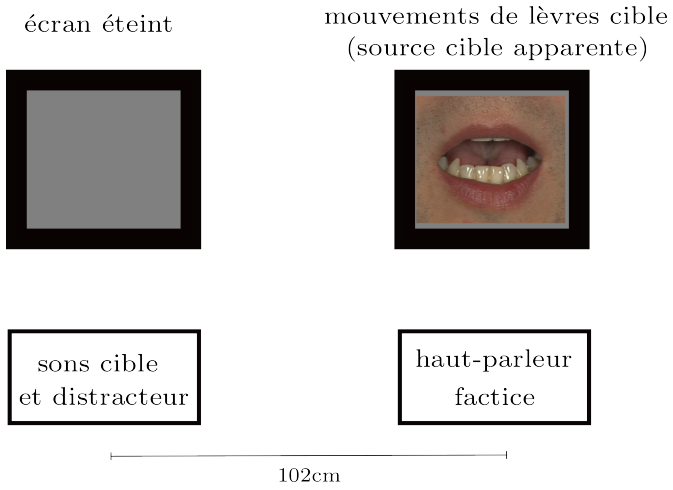
nous allons aborder ici, consiste à lier les différents stimuli sur la base d'une cohérence physique. La cohérence entre les traits peut être spatiale et/ou temporelle. Le liage cognitif est un liage tardif comme peut l'être l'intégration phonétique décrite dans l'étude de Massaro et Cohen (1983).

Les travaux, évoqués ci-dessous, ont utilisé l'effet ventriloque pour étudier le liage audiovisuel. L'effet ventriloque classique consiste à biaiser la localisation spatiale de la source en déplaçant la source visuelle de la source sonore réelle. La localisation que nous faisons repose sur l'indice visuel. On parle de ventriloquisme spatial. Ce biais de localisation est la conséquence d'un liage perceptif relativement fort entre ce que l'on voit et ce que l'on entend. Driver (1996) ont réalisé une expérience basée sur ce ventriloquisme spatial (figure 1.21). Deux flux de parole (cible et distracteur) étaient présentés sur l'un des deux haut-parleurs soit à droite, soit à gauche. Deux écrans étaient placés devant le participant. Un film articulant la phrase cible était présenté soit à droite soit à gauche. Le flux visuel et le flux auditif étaient présentés soit du même côté soit à des positions différentes. Dans cette deuxième configuration, l'effet ventriloque se produit, conduisant les participants à localiser le flux sonore à l'endroit où est diffusé le film. De plus, les auteurs parviennent à montrer que la compréhension de la parole cible est meilleure dans cette situation ventriloque que dans la situation cohérente (figure 1.22). Les auteurs émettent l'hypothèse que l'indice visuel a de manière illusoire séparé les deux flux de parole spatialement, conduisant à une amélioration de la compréhension. Cette expérience montre à quel point le liage audiovisuel peut être fort.

Un autre effet de ventriloquisme existe, il s'agit du ventriloquisme temporel (par ex. Bertelson et Aschersleben, 2003). Pour mettre en évidence ce type d'effet, les expériences consistent à présenter une série de sons purs de très courte durée (des bips) accompagnée d'un nombre différent de flashs visuels. Les participants sont alors invités à rapporter le nombre de flashs visuels qu'ils ont perçu (figure1.23). L'effet de ventriloquisme temporel conduit les participants à percevoir un nombre de flashs visuels égal au nombre de bips qu'ils ont entendu, même si physiquement des flashs étaient absents



(a) situation cohérente



(b) situation ventriloque

FIGURE 1.21 – Illustration adaptée du protocole expérimental de Driver (1996)

au moment des bips. Watkins *et al.* (2007) ont proposé une expérience identique. En enregistrant l'activité dans les aires visuelles primaires, les auteurs sont parvenus à mettre en évidence une activité identique dans le cas d'un flash présent et d'un flash absent. Selon les auteurs, le liage perceptif est si fort dans ce type d'effet qu'une représentation de l'objet audiovisuel formé redescend jusque dans les aires primaires.

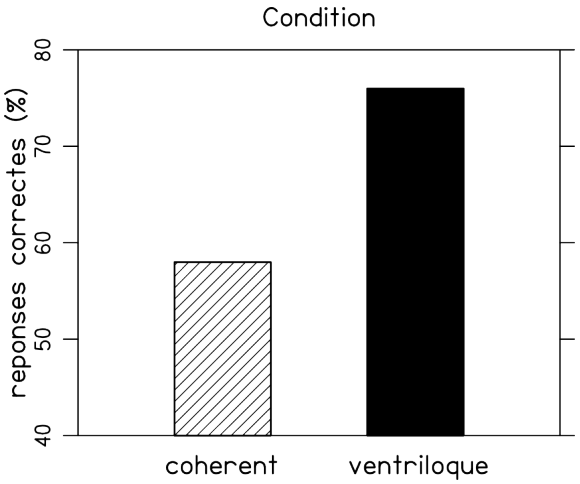


FIGURE 1.22 – Résultats de l'expérience de Driver (1996)

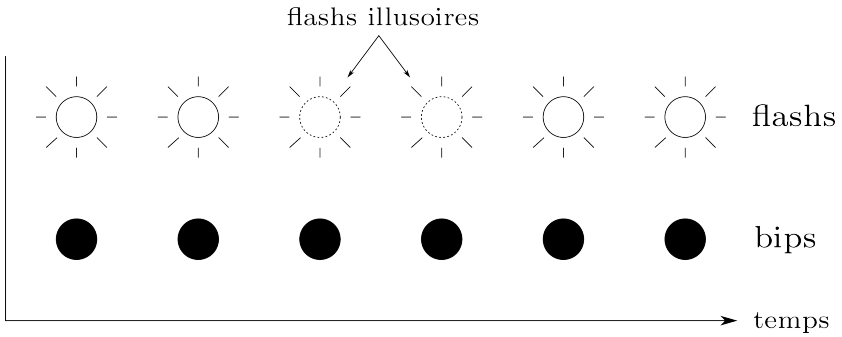


FIGURE 1.23 – Illustration de l'effet de ventriloquisme temporel

Le liage d'éléments visuels et auditifs peu complexes comme des sons purs et des flashs peut être suffisamment fort pour induire ces effets de ventriloquisme. Cependant, pour être liés, les stimuli doivent respecter une proximité spatiale et temporelle qui a été estimée dans l'étude réalisée par Conrey et Pisoni (2003).

1.3.3 Liage tardif

Nous pouvons cependant tolérer une certaine quantité d'incohérence spatiale et temporelle lorsque qu'il ne s'agit pas de signaux simples (bips et

flashs) comme des signaux de parole (par ex. Macaluso *et al.*, 2004). Lorsque nous sommes confrontés à des signaux de parole, nous pouvons tolérer une importante quantité de dé-synchronisation temporelle entre le signal visuel et le signal auditif. On parle alors de fenêtre d'intégration temporelle (Massaro *et al.*, 1996; van Wassenhove *et al.*, 2007). Nous pouvons tolérer au maximum une avance du signal visuel par rapport au signal auditif de 70ms et un retard du signal visuel de 130ms (figure 1.24). Ces valeurs sont indicatives. Dans la littérature, on retrouve plusieurs étendues pour cette fenêtre d'intégration. La largeur de cette fenêtre peut varier si l'on expose les participants à des signaux auditifs et visuels désynchronisés (Navarra *et al.*, 2005) pendant une phase préliminaire. Après cette exposition à du matériel dé-synchronisé, les participants ne perçoivent plus les dé-synchronisations auxquelles ils étaient sensibles avant cette phase. Ils tolèrent des dé-synchronisations plus importantes. Cette fenêtre d'intégration audiovisuelle n'est pas une fenêtre symétrique. Il nous est difficile d'admettre que le flux auditif et le flux visuel émanent d'une même source si le son est trop en avance par rapport à l'image. De plus, Stekelenburg et Vroomen (2007) ont proposé que le traitement du flux visuel était plus long que le traitement du flux sonore. Une perception synchrone des deux flux implique alors que les traitements respectifs soient terminés au moment de l'évaluation de la synchronie. Cela pourrait expliquer en partie pourquoi le flux visuel doit précéder le flux auditif.

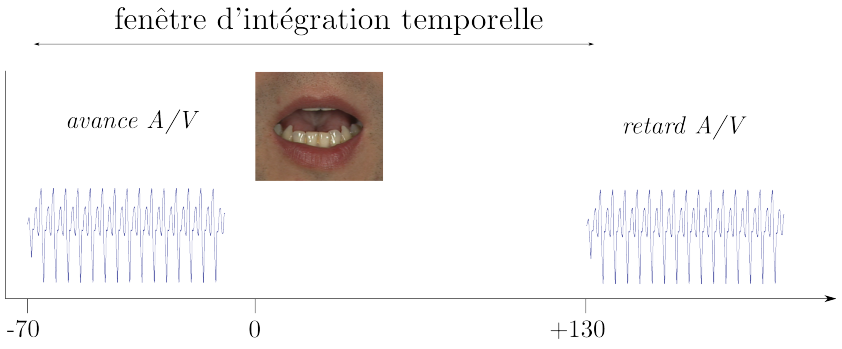


FIGURE 1.24 – Représentation d'une fenêtre d'intégration temporelle

Percevoir la synchronie entre une entrée visuelle et auditive de parole

ne repose donc pas exclusivement sur la synchronisation physique parfaite puisque nous pouvons tolérer des dé-synchronisations. Notre perception subjective de la synchronie est donc différente de la synchronisation physique. Cette suposition rejoint l'idée proposée par Efron (1970). Ce dernier établit une distinction entre temps perçu et temps de la stimulation. Dans cette même perspective, Pariyadath et Eagleman (2007) montrent que plus un stimulus est prédictible, plus son décours temporel perçu est stable. On comprend alors pourquoi la fenêtre d'intégration temporelle pour la parole est plus étendue que pour un stimulus non langagier. Le signal de parole est plus prédictible qu'une succession de bips et de flashs. Le temps perçu est donc plus stable. Notre acuité à discriminer des dé-synchronisations pour un signal de parole est plus faible que pour des signaux non-langagiers. La fenêtre est dons plus étendue.

Pour résumer, le liage tardif (cognitif) autorise une tolérance à la désynchronisation entre les signaux visuels et auditifs qui n'est pas observée avec des signaux non langagiers. L'intégration tardive audiovisuelle peut dépasser des incohérences spatiales et temporelles perceptives pour induire une représentation cohérente subjective. Les deux liages perceptif et cognitif peuvent collaborer pour créer une représentation des signaux de parole audiovisuelle. Soto-Faraco et Alsius (2009) ont également montré qu'il était possible d'accéder consciemment à ces deux types de liage.

1.3.4 Présomption d'unité

Le liage audiovisuel semble donc pouvoir s'établir à plusieurs niveaux de traitement. Pour rendre compte de l'importance de la force du liage qui est rapportée pour la parole, le concept de présomption d'unité a été suggéré dans plusieurs travaux. Selon ce principe, lorsque nous percevons des événements auditifs et visuels cohérents, cela nous permet d'accéder à une représentation de cet objet unique. Si la cohérence est importante, nous faisons l'hypothèse que les stimuli des différentes modalités émanent de la même source. Nous faisons l'hypothèse d'unité (*Unity assumption*). Vatakis et Spence ont largement contribué à l'étude de ce concept (Vatakis et Spence, 2006, 2007a,b;

Vatakis *et al.*, 2008a; Vatakis et Spence, 2008). Dès que nous acceptons que ce que nous percevons est un objet unique, nous sommes capables de tolérer de plus grandes incohérences spatiales et temporelles. Selon Kanai *et al.* (2007), la perception de la synchronie audiovisuelle serait directement liée à ce concept de présomption d'unité. La parole revêt donc une place particulière car nous admettons très facilement l'unité de l'objet audiovisuel de parole. Il est absolument indispensable, vis-à-vis de notre interaction sociale, que nous puissions percevoir tout signal de parole le plus efficacement possible.

Le concept de présomption d'unité ne nous apprend pas explicitement de quelle manière on accède à ce percept *unitaire*. Gardons à l'esprit ce concept, et tournons nous vers la notion de cohérence audiovisuelle. Finalement, pour établir le fait que les stimuli auditifs et visuels émanent de la même source et soient perçus comme un objet unique, il faut en évaluer la cohérence. C'est le niveau de cohérence reposant à la fois sur des éléments perceptifs mais également cognitifs qui va nous conduire à cette présomption d'unité. Plutôt que de parler de présomption d'unité, nous parlerons dans la suite de l'évaluation de cohérence audiovisuelle. Cette évaluation est un processus dynamique qui s'établit sur le long terme. La présomption d'unité s'appuie donc sur l'évaluation de cette cohérence.

Alpert *et al.* (2008) ont étudié, sur le plan neurophysiologique, la dynamique du traitement de l'information audiovisuelle. La connaissance de cette dynamique a permis à Arnal *et al.* (2009) de proposer un modèle d'intégration audiovisuelle (figure 1.25). Ce qui est particulièrement intéressant dans ce modèle, c'est la mise en place de connexions descendantes depuis les aires responsables des traitements tardifs vers les aires responsables des traitements précoces. Ce bouclage permettrait de rendre compte de la force du liage que l'on observe pour la parole. Si les traitements phonétiques (tardifs) spécifiques au traitement de la parole viennent moduler l'activité d'aires primaires, on comprend pourquoi il est possible de tolérer des dé-synchronisations physiques entre les stimuli. Les traitements tardifs viennent aligner subjectivement les stimuli sur la base de la présomption d'unité. Les voies de traitement descendantes peuvent redescendre relativement profondément vers les aires primitives. Selon Ponton *et al.*

(2009), les traitements phonétiques d'une stimulation visuelle peuvent se mettre en place dès les aires auditives primaires. Nous rejoignons ici, l'idée de facilitation de traitement du stimulus auditif par présentation d'un stimulus visuel (Besle *et al.*, 2004). Avec ces travaux, nous nous éloignons d'une vision hiérarchique des processus d'intégration audiovisuelle. Les processus de traitements phonétiques (tardifs) peuvent prendre place dans des aires primaires et influencer notre évaluation perceptive de la cohérence. Ainsi, Mesgarani *et al.* (2008) ont montré que des processus de classification phonémique pouvaient se dérouler dans les aires auditives primaires. Cette modélisation d'interactions à plusieurs niveaux semble donc tout à fait pertinente.

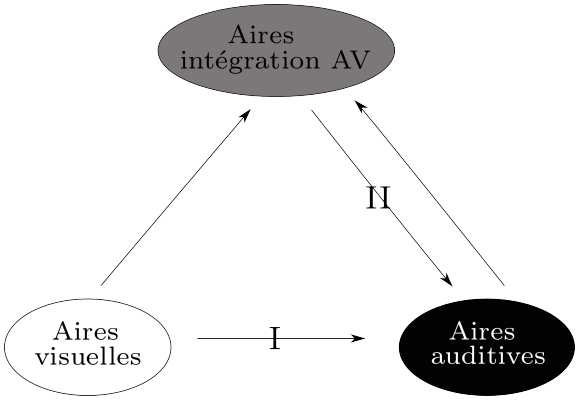


FIGURE 1.25 – Modèle d'intégration audiovisuelle proposé par Arnal *et al.* (2009)

1.3.5 Appariement audiovisuel

Comme nous l'avons mentionné ci-dessus, pour évaluer la cohérence audiovisuelle, nous devons être capable de lier les stimuli auditif et visuel. Si nous échouons dans cette opération de liage, alors nous pourrons émettre l'hypothèse que les stimuli auditifs et visuels sont issus de sources différentes. Plutôt que de parler de liage qui pourrait reflêter un processus passif dans lequel la salience perceptive fait que nous lions les événements entre eux, nous allons introduire dans ce qui suit la notion d'appariement.

La troisième étude que nous avons réalisé a été consacrée à l'exploration des mécanismes du liage pour des signaux de parole (Chapitre 4). Nous avons proposé de décomposer le stimulus visuel de parole en éléments visuels plus simple, à la manière de Treisman et Gelade. Au lieu de présenter des lèvres, nous avons présenté des figures géométriques simples (disque, barres) qui variaient, soit en mouvement, soit en contraste. Cette simplification des indices visuels était alors supposée réduire la cohérence forte de la parole et ainsi réduire la force du liage audiovisuel sous-jacent. Les participants engagés dans cette étude devaient alors apparier ces indices visuels avec une partie du flux auditif. La réussite à cette tâche était alors supposée reflêter la force du liage audiovisuel.

Chapitre 2

Effet de la lecture labiale sur la segregation auditive primitive

Dans cette première étude, nous avons évalué l'effet de la lecture labiale sur les mécanismes de ségrégation irrépressible. Pour cela, nous proposons des séquences de voyelles dont la fréquence alterne entre deux valeurs à la manière de Miller et Heise (1950). Simultanément, des mouvements de lèvres articulant une voyelle sur deux sont présentés. Deux tâches comportementales ont été réalisées. Dans une première expérience, les participants devaient rappeler l'ordre de présentation des voyelles de la séquence. Dans une seconde expérience, les participants devaient détecter une variation dans le rythme de présentation de la séquence. Ces deux tâches sont réputées être plus difficile lorsqu'il y a ségrégation. Les résultats obtenus sont consistant avec l'hypothèse suivant laquelle les mouvements de lèvres viennent influencer la ségrégation auditive irrépressible. L'effet que nous rapportons ici est différent d'un effet de type lecture labiale car les méthodes expérimentales permettent d'étudier des mécanismes de ségrégation précoces. De plus, cette étude souligne l'importance de la cohérence audiovisuelle pour observer cette interaction. En effet, c'est dans les conditions où les signaux auditifs et visuels étaient les plus cohérents que nous sommes parvenus à mettre en évidence cette interaction. Ce point sera plus largement développé dans le chapitre 5.1. *Cet article a été accepté sous réserve de corrections mineures pour pub-*

lication dans JASA *le 2 octobre 2010.*

The effect of lip-reading on primary stream segregation

Aymeric Devergie and Nicolas Grimault [a)]

Laboratoire de Neurosciences Sensorielles Comportement et Cognition,

CNRS UMR 5020 Université Lyon 1 69366 Lyon Cedex 07 FRANCE

Etienne Gaudrain

MRC Cognition and Brain Sciences Unit Cambridge UK

Eric W. Healy

Speech Psychoacoustics Laboratory,

Department of Speech and Hearing Science The Ohio State University Columbus USA

Frédéric Berthommier

GIPSA-Lab CNRS UMR 5216,

Domaine universitaire 38402 Saint Martin d'Hères FRANCE

(Dated: October 20, 2010)

[a)] author to whom correspondence should be addressed. Electronic mail: nicolas.grimault@olfac.univ-lyon1.fr

Abstract

Lip-reading has been shown to improve the intelligibility of speech in multi-talker situations, where auditory stream segregation naturally takes place. This study investigated whether the benefit of lip-reading is a result of a primary audiovisual interaction that enhances the obligatory streaming mechanism. Two behavioral experiments were conducted involving sequences of French vowels that alternated in fundamental frequency. In Experiment 1, subjects attempted to identify the order of items in a sequence. In Experiment 2, subjects attempted to detect a disruption to temporal isochrony across alternate items. Both tasks are disrupted by streaming, thus providing a measure of primary or obligatory streaming. Visual lip gestures articulating alternate vowels were synchronized with the auditory sequence. Overall, the results were consistent with the hypothesis that visual lip gestures enhance segregation by affecting primary auditory streaming. Moreover, increases in the naturalness of visual lip gestures and auditory vowels, and corresponding increases in audiovisual congruence lead to increases in the effect of visual lip gestures on streaming.

PACS numbers: 43.66.Mk, 43.71.Rt

Keywords: auditory scene analysis, primary auditory streaming, lip-reading, audiovisual speech, audiovisual congruence

I. INTRODUCTION

Previous research has explored the segregation mechanisms that are most likely to be employed in competitive listening situations, such as the perception of concurrent speech (Bregman, 1990). Van Noorden (1975) described a helpful experimental paradigm to study the contribution of acoustic cues to auditory segregation and, specifically, sequential segregation mechanisms. This streaming paradigm uses the sound sequence, ABA-ABA-..., composed of two tones, A and B, that differ by some acoustic attribute. Moore and Gockel (2002) found that any salient acoustic difference can help listeners to segregate A tones from B tones and to group them into two distinct auditory streams. In his work, Van Noorden (1975) observed two types of streaming mechanisms depending on the task given to the subject. Obligatory, automatic or primary streaming is observed when subjects try to fuse the sequence into a single stream (but fail to do so), whereas voluntary or schema-based streaming is observed when subjects try to segregate the sequence into two streams (and succeed in doing so). In addition, using a similar paradigm, Bregman (1978) reported that segregation requires about two or three seconds of build-up time to take place.

More recently, electro-physiological studies have been at determining the level of processing at which primary stream segregation takes place. The method used in these studies was to record neural firing rate during the presentation of ABA sequences. Because segregation was initially absent due to build-up, the ABA sequence paradigm enabled researchers to compare different segregation states (i.e., items integrated versus segregated). Using this method, Micheyl *et al.* (2005) recorded single unit responses in the primary auditory cortex of awake rhesus monkeys. Pressnitzer *et al.* (2008) employed the same recording method in the cochlear nucleus of anesthetized guinea pigs. In these two studies, the authors reported two different firing-rate patterns before and after the build-up period: the units responded to all the A and B tones at the beginning of the sequence, but responded selectively to the A tones after 10 s of build-up. These studies therefore indicate that segregation can take place in the primary sub-cortical and cortical structures of the auditory pathway.

In natural environments, another fundamental contribution to the perception of concurrent speech is lip-reading (speechreading). Lip-reading can improve the intelligibility of speech presented in a noisy environment by up to 40% (Sumby and Pollack, 1954). This benefit is likely sustained by multiple levels of interaction in the integration of audiovisual speech. Massaro and Cohen (1983) and Brancazio and Brancazio (2004) found evidence for high level interactions. Additionally, behavioral and neurophysiological studies have suggested that audiovisual interactions can also occur at lower levels of processing. Along with previous studies (Grant et Walden, 1996; Grant and Seitz, 2000; Grant, 2001; Grant *et al.*, 2004), Bernstein *et al.* (2004) found that speech detection in a noisy environment could be enhanced by visual cues that were synchronized with the sound intervals. Moreover, this effect was larger for lip-reading cues that were highly congruent with the auditory input than for other less congruent visual displays.

Bernstein *et al.* suggested that the benefit reported in this two-interval forced choice detection task could rely on an audiovisual interaction that might have occurred at a relatively primary level of processing. Using fMRI, several studies have also reported that the presentation of visual articulatory gestures with sounds (Pekkola *et al.*, 2005; Kayser *et al.*, 2008) can activate primary auditory cortical structures. Other studies involving electrophysiological recordings (e.g. Besle *et al.*, 2008; Van Wassenhove *et al.*, 2005) report similar findings. Altogether, these studies suggest that visual and auditory inputs might interact in primitive neural structures. Since these primary structures seem to support both the primary segregation mechanism and the audiovisual interactions, it can be hypothesized that primary auditory segregation could be modulated by audiovisual interactions.

To further explore the mechanisms of segregation, Rahne and his colleagues (Rahne *et al.*, 2007, 2008) and Rahne and Böckmann-Barthel (2009) built sequences of pure tones designed to induce different perceptual organizations. The frequency of the tones alternated between low and high. Whereas the high-frequency tones appeared in random order, the low-frequency tones together formed a sequence composed of a repeated pattern of three tones rising in pitch, sometime replaced by a deviant pattern of three tones falling in pitch.

In addition, every third tone in the overall sequence was more intense by 15 dB. Perceptual organization could then be based on either a frequency difference (grouping the lower tones in one stream and the higher tones in another stream) or an intensity difference (grouping the louder tones in one stream and the softer tones in an other stream). A visual cue (squares or circles of different sizes) synchronized either with the frequency or with the intensity pattern was added to influence perceptual organization. In one condition, the visual cues promote segregation based on frequency, signaling the deviant pattern whatever the intensity variations. In the other condition, the visual cue promoted segregation based on intensity whatever the frequency variation.

In two electroencephalography studies, Rahne *et al.* (2007) and Rahne and Böckmann-Barthel (2009) reported mismatch negativity (MMN) when the visual cue promoted a perceptual organization based on a pitch difference, indicating that the participants were indeed perceiving the low-frequency tone sequence segregated from the high-frequency tone sequence, and the resulting deviant pattern. They concluded that a visual cue can alter the perceptual organization of an ambiguous tone sequence. However, as acknowledged by the authors, this design allowed detection of only the pitch based segregated percept, which also correspond to the intensity based integrated percept. It is then unclear from this result whether a visual cue can promote integration across an intensity difference, a pitch difference, or both. Rahne *et al.* (2008) extended these results using the same materials and a behavioral design. Participants were instructed to attend to the visual stimuli and to indicate the currently prevailing sound organization (grouped based on frequency or grouped based on intensity) by pressing one of two buttons on a keypad and change buttons if the impression changed. The time during which the lower and the higher tones were grouped together was then measured. Their results indicated that the visual cue aimed to promote segregation on the basis of pitch did not affect the perceptual organization. In contrast, the visual cue synchronized with intensity variations succeeded in reducing frequency based segregation in experimental conditions with a large frequency difference. So, although these authors demonstrated a clear influence of an arbitrary visual cue on the perceptual orga-

nization of pure tone sequences, it remains unclear how this influence operates, and it is difficult to extend these conclusions to auditory processing in more ecological situations.

In the current study, sequences of French vowels, as a first approximation of continuous speech, with alternating fundamental frequency (F0) were presented to listeners who perceived either a single stream or two streams, depending on the F0 difference between alternate vowels. In the first experiment, participants had to recall the order of the vowels presented in the sequence (order-naming task). In the second experiment, participants had to detect a change in the rhythm of presentation of the sequence (isochrony detection task). The participants can only succeed in these two tasks if they can integrate all the vowels in a single auditory stream (Micheyl et Oxenham, 2010). Therefore, poor performance in these tasks indicates that streaming has occurred despite the effort of the participant to prevent it, thus providing a measure of obligatory (i.e., primary) streaming.

To evaluate the contribution of lip-reading to primary streaming, lip gestures articulating alternate vowels of the sequence were presented to participants in both experiments. We hypothesized that performance would be degraded if the visual and auditory inputs interacted at a low level of processing, thus indicating that primary auditory segregation is enhanced by lip-reading.

II. EXPERIMENT 1

Experiment 1 was designed to test for an effect of lip-reading on primary auditory segregation. Segregation was estimated by assessing participants ability to correctly report the order of presentation of sequences of vowels alternating in pitch. Good performance on this task would suggest that participants integrated all the vowels of the sequence into a single stream (Gaudrain *et al.*, 2007, 2008). To test for a lip-reading effect, visual lip gestures articulating alternate vowels were simultaneously displayed with three different degrees of audiovisual congruence (C_0, C_1 and C_2). In the first condition (C_0), the lips were displayed throughout the vowel sequence without moving. Thus, there was no audiovisual

congruence. In the second condition (C_1), the visual lip gestures only provided rhythmic information by opening and closing the mouth in synchrony with the auditory vowels, causing the audiovisual congruence to be limited solely to temporal aspects. In the last condition (C_2), auditory vowels and visual lip gestures were rhythmically and phonetically congruent. If lip-reading can promote primary segregation, the presence of a congruent visual cue should result in poorer performance.

A. Participants

Ten participants aged between 18 and 24 years (mean=20.8, SD=1.8) took part in the experiment. All of the participants were native French speakers and had pure tone audiometric thresholds below 15 dB HL at octave frequencies between 250 and 4000 Hz (American National Standards Institute, 2004). Participants signed an informed consent form and were reimbursed for their time. This study was formally approved by a local ethics committee (CPP Sud-Est II No. 06035).

B. Stimuli

Sequences of six French vowels (/a/, /e/, /i/, /o/, /y/, /u/) with alternating high and low fundamental frequencies were constructed (Fig. 1). The lowest fundamental frequency F0(1) was equal to 100 Hz. The alternate fundamental frequencies F0(2) were set to one of 10 values between 100 and 238 Hz. Each vowel was 166 ms long, including a 10 ms raised-cosine onset and offset ramp, and was adjusted to an RMS value of 85 dB SPL. These vowels were generated using the Klatt algorithm (Klatt, 1980) and had the same formant values used by Gaudrain *et al.* (2007). A sequence of visual lip gestures pronouncing alternate vowels was presented simultaneously with the audio sequence. These visual lip gestures were synthesized using video-recorded frames as in Berthommier (2003). One hundred video-recorded frames entered into an XY diagram. The X dimension reflected the horizontal extension of the lips and Y reflected the vertical extension. The starting point (closed lips) and ending point

(target position of the lips articulating one particular vowel) were selected manually. An algorithm by Berthommier (2003) was used to estimate the trajectory between these two points and select the frames closest to it. Thus, the video frames displaying the lips were real but the trajectories were artificial. Such synthetic visual vowels were used previously in Berthommier 's study, and the decision was made to use these established materials to test the effect of lip-reading on primary streaming. Auditory and visual dimensions of the stimuli were built separately and then combined.

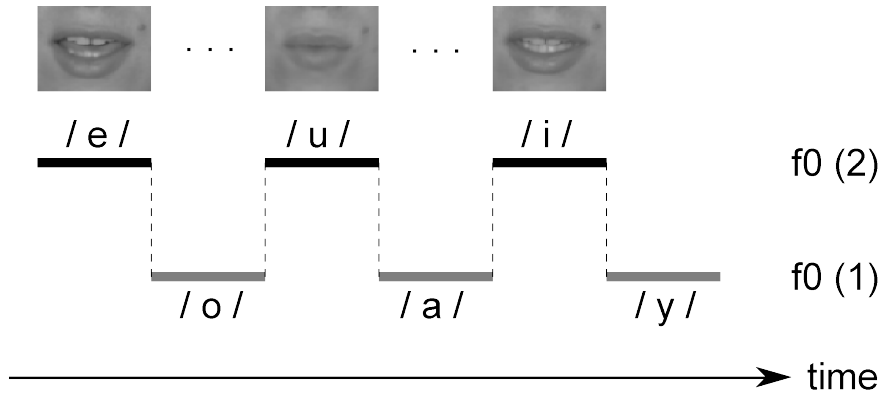


FIG. 1. Schematic representation of an audiovisual sequence. Lip gestures were presented to participants that articulated either the three high-pitch vowels or the three low-pitch vowels, selected randomly across trials (except that in some cases, F0(1)=F0(2)).

Three different types of lip gestures were presented. The first condition (no congruence, C_0) consisted of lips that remained closed during the whole sequence. The second condition (temporal congruence, C_1) consisted of an open-close lips gesture. The lip gesture for the /a/ vowel, which is a neutral open-close gesture, was used for alternate vowels. This visual condition provided a rhythmic cue to either the F0(1) or F0(2) alternate vowels but no phonetic cue. The last condition (temporal and phonetic congruence, C_2) consisted of lip gestures pronouncing the corresponding F0(1) or F0(2) auditory vowel. In addition to the rhythmic cue provided in C_1, this visual condition also provided some phonetic information

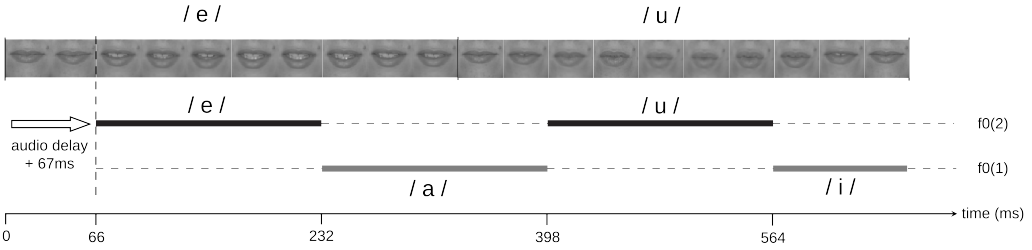


FIG. 2. Synchronization of audio and visual streams. The figure shows lip movements congruent with the F0(2) vowels. The maximum opening of the mouth is centered on the cued audio vowel, but begins slightly before and remains somewhat after each cued auditory vowel. Each picture of the lips was extracted from the lip gesture movie that started 66 ms (i.e., 2 frames) before the corresponding auditory vowel and 99 ms after the offset of the auditory vowel (i.e., 3 frames).

about the vowel. The six different lip gestures corresponding to the six different vowels used are plotted in the Appendix (Fig. 6). The relative timing of the audio and video material is detailed in Fig. 2. The lip gestures started 67 ms (i.e., 2 frames at a rate of 30 frames per second) before the corresponding audio vowel. The closing gesture of the lips occurred during the auditory vowel that immediately followed (see Fig. 2). The choice of the temporal offset was made arbitrarily, but was sufficient to preserve a good subjective synchronization between the auditory and visual dimensions of the signals. Also, it has been well established that delays of up to 170 ms for the auditory signal relative to the video can be employed without affecting the binding of the two signals (Grant *et al.*, 2003).

All sequences were created and stored using a C program prior to conducting the experiments. The stimuli were played diotically using a Digigram VxPocket440 sound card connected to a Sennheiser HD 250 Linear II headphone. Visual stimuli were displayed on a monitor with a visual angle of approximately 6°. The listeners were comfortably seated in a double-walled, sound attenuated booth. The output level was calibrated to 85 dB SPL [Larson Davis AEC101 and 824; American National Standards Institute (1995)].

C. Procedure

The participants began the experiment with an identification test to ensure that the vowels were correctly identified. Each vowel sound was played separately, and the participants selected the corresponding vowel among six choices displayed on a computer screen using orthographic representations ("a, é, i, o, u, ou"). Each of the six vowels was played five times, at random F0s among 100, 147 and 238 Hz. All the participants correctly identified the vowels 100% of the time, with the exception of one participant who correctly identified the vowels 87.5% of the time. Next, the participants engaged in the order naming task in which they were briefly trained before the start of testing. To free the participants from having to memorize the sequence, each audiovisual sequence was loop repeated for ten seconds. Two seconds after the beginning of each sequence, which is approximately the time required for the build-up of auditory streaming (Bregman, 1978), participants were instructed to report the order in which vowels appeared by clicking on a graphical user interface displayed on the computer screen, a few centimeters below the video of the lips. After participants confirmed their response or at the end of the ten-second period during which the sequence was presented, there was silence for a period of seven seconds, followed by the next sequence.

In one block, the thirty combinations of the three conditions of audiovisual congruence and the ten F0s were randomly repeated five times. Each participant ran eight blocks of trials. Overall, each combination was repeated 40 times (i.e., 8 blocks $\times$ 5 repetitions). The experiment was divided into three sessions of two hours each.

D. Results

Figure 3 shows percent-correct as a function of fundamental frequency and audiovisual (AV) congruence, averaged across participants. Responses were considered correct when the six vowels were reported in the correct order. A two-way repeated measures ANOVA was performed with fundamental frequency and AV congruence as repeated factors. Performance significantly decreased as the fundamental frequency increased [$F(9,81)=22.13$; $p<0.0001$],

and AV congruence (visual condition) tended to reduce performance [$F(2,18)=3.47$; $p=0.05$] [1]. There was no interaction between these factors [$F(18,162)=0.87$; $p=0.61$] [2]. To clarify the effect of the AV congruence, a Bonferroni-corrected post-hoc test was conducted. This test revealed that the performance in C_0 was significantly better than in C_2 [$p=0.05$]. There was no difference between C_1 and C_0 [$p=0.37$] or between C_1 and C_2 [$p>0.99$]. Note that the displayed standard error was across subject for each condition, while in the repeated measure ANOVA and in the post-hoc tests, the error term was based on the difference between condition within subjects. Finally, it is worth to note the same ANOVA analysis performed on the responses with 4 vowels in in correct order (instead of six) strengthened the audiovisual effect (Effect of CV : $F(2,18)=10.75$, $p<0.001$; Effect of F0 : $F(9,81)=30$, $p<0.0000$; Interaction : $F(18,162)=1.35$, $p=0.16$).

E. Discussion

The data from this experiment suggest that a phonetically congruent lip movement can enhance obligatory auditory streaming. In contrast, streaming was not significantly enhanced by lip-reading when the lip gesture consisted of a simple open-close gesture. In this latter condition, the audiovisual interaction was likely weakened by the lack of phonetic congruence between the audio and visual inputs. This finding is consistent with the results reported by Rahne *et al.* (2007) and Rahne and Böckmann-Barthel (2009). These authors used basic geometric shapes of various sizes in synchrony with tones, which is similar to our condition C_1 that provided a visual rhythmic cue. Both the current results and those obtained by Rahne *et al.* (2008) suggest that high level of congruence between audio and visual signals may be required to observe an influence of audiovisual input on streaming.

The data also showed that there was no statistical interaction between the effect of the visual cue and the effect of the fundamental frequency on streaming. This finding suggests that these two factors contribute independently to primary streaming.

However, these conclusions need to be taken with caution because the task did not

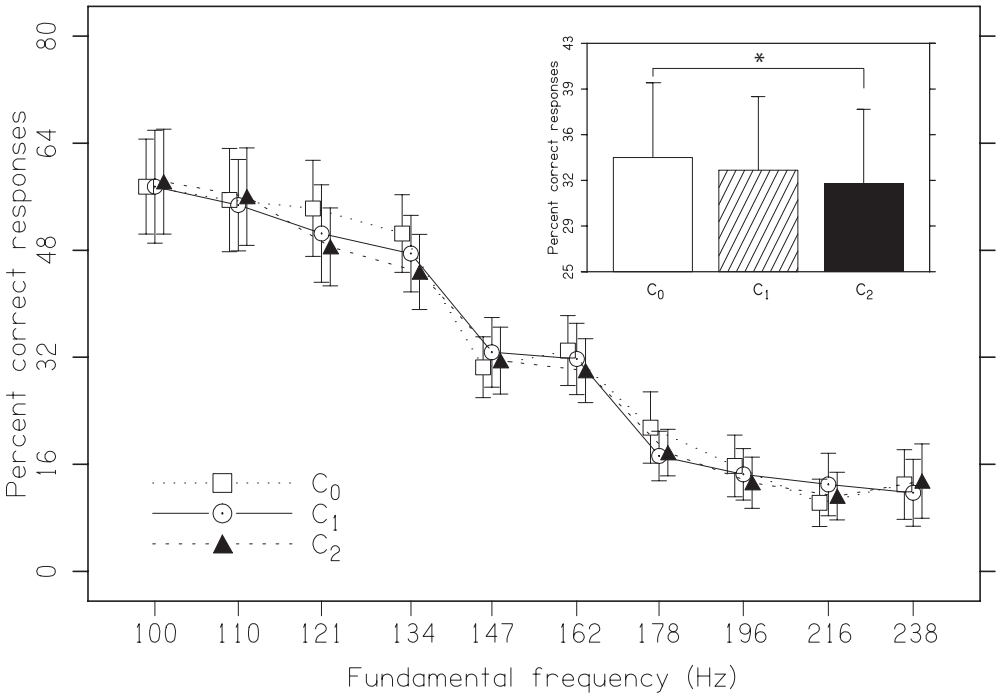


FIG. 3. Results from Experiment 1. Percent correct naming of the order of presentation of the six vowels is shown as a function of the F0 difference between alternate vowels and the visual condition. Bars represent the standard errors.

demonstrate as much sensitivity as expected and the reported effects are small. There are at least two interpretations for these small effects. First, the audio and visual parts of the signal were generated independently, and the degree of audiovisual congruence varied from null (C_0) to moderate (C_2) but remained much lower than in natural lip-reading situations. To account for this concerns, a second experiment was designed using natural audiovisual vowels and a synchrony detection task.

III. EXPERIMENT 2

The results from the first experiment support the hypothesis that primary segregation is enhanced by lip-reading. However, the degree of audiovisual congruence was relatively low and might have accounted for the small (but significant) effect. To enhance the audiovisual coherence in the second experiment, all audiovisual vowels were presented from audio/video recordings of the same male speaker. This enhanced the synchrony between auditory and visual inputs, and provided more natural auditory and visual stimuli.

The task in this experiment was to detect a change in the presentation rate of vowel sequences alternating in pitch similar to those employed in the first experiment. Previous studies have shown that it is difficult to judge the relative timing of sounds that are perceived in different streams (Cusack and Roberts, 2000; Roberts *et al.*, 2002; Stainsby *et al.*, 2004). As in Experiment 1, conditions of differing different congruence were created. In one condition, there was no AV congruence (C_0). In the other condition (C_3), alternate vowels were presented with an AV congruence stronger than that used in Experiment 1. As in the first experiment, poor performance in this task indicates that primary stream segregation has occurred. We predicted that the effect of lip-reading on segregation would be revealed by a lower performance in C_3 than in C_0.

A. Participants

Ten participants aged 18 to 24 (mean=21.5, SD=2.9) took part in Experiment 2. None had participated in Experiment 1. The audiometric threshold criteria were the same as in the first experiment.

B. Stimuli

As in Experiment 1, sequences of French vowels alternating in F0 were created. The same six French vowels were recorded (44.1 kHz, 16 bits) using simultaneous audio and video recording. The fundamental frequencies and durations of the natural productions were

adjusted using STRAIGHT (Kawahara *et al.*, 1999) to reach a low and a high fundamental frequency and a fixed duration equal to 166ms, including a 10ms raised-cosine onset and offset ramp. The low fundamental frequency F0(1) was equal to 100 Hz. The second fundamental frequency F0(2) consisted of two possible values: 100 Hz or 224 Hz. The stimuli were presented in intervals of twelve pairs of alternating vowels. The vowels were randomly chosen for each interval. The sequences started with either a high-pitch vowel or with a low-pitch vowel.

Two different lip gestures were used. In the first condition (C_0) of null AV congruence, the lips remained closed during the entire sequence. In the second condition (C_3) of strong AV congruence, natural lip gestures pronounced the odd-numbered (so alternate) auditory vowels in the sequence. The lip gestures for the six vowel sounds are displayed in Appendix A (Fig. 7). The lip gestures started 51 ms (i.e., 2 frames at 39 frames per second) before the corresponding vowel sound and finished 154 ms after (6 frames at 39 frames per second). Figure 4 shows a schematic representation of a complete C_3 sequence. All the sequences were created on-line using a Python program. The stimuli were played diotically using a Sigmatel sound card connected to a Sennheiser HD 250 Linear II headphone. Visual stimuli were displayed on a monitor with a visual angle of approximately 6°. Listeners were comfortably seated in a double-walled, attenuated sound booth. The output level of the vowels was calibrated to 70 dB SPL (Larson Davis AEC101 and 824; American National Standards Institute (1995)).

C. Procedure

The method used in the current experiment was similar to that used by Cusack and Roberts (2000) and Roberts *et al.* (2002). A two-interval forced-choice method with a three-down, one-up decision rule (Levitt, 1971) was used to measure the smallest detectable temporal shift of the even numbered vowels (audio only) relative to the odd numbered vowels (audiovisual vowels). In each trial, the participants were required to identify the interval

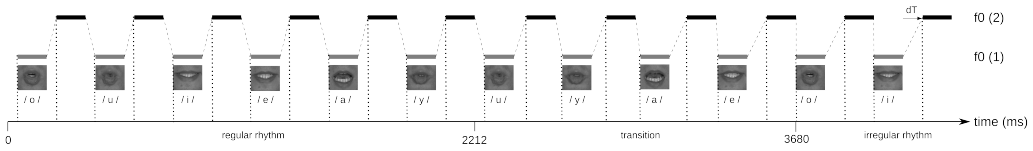


FIG. 4. Schematic representation of an audiovisual sequence. The sequence is first regular and then the audio-only stream (in black) is gradually delayed to reach the final value of dT that the listener has to detect. The audiovisual stream (in grey) is always regular. Vertical dotted lines exhibit the transition from a regular to an irregular rhythm of the audio-only stream.

containing the temporal shift, which was randomly assigned to one of the two intervals. The interval containing the control sequence was an isochronous sequence of vowels, each separated by a constant 40 ms inter-stimulus interval (ISI). In the interval containing the target sequence, the first six pairs of vowels had a constant ISI of 40 ms to allow sufficient time for streaming to build up (Bregman, 1978). In the subsequent four pairs, the ISI between the vowels was progressively increased by an additional delay (rhythm deviation, dT) that ranged from 0 ms to a maximum value of 40 ms to avoid temporal overlap between two successive vowels. The dT reached after this transition phase was maintained during the last two pairs (see Fig. 4). The total duration of each sequence was 5 s. The silence between the two intervals was 7 s. It is worth noting that, because the audio-only vowels were delayed, the rhythm of the audiovisual vowel presentation remained constant in both intervals.

In the adaptive procedure, the initial value of dT was set to 20 ms. The dT was then adjusted on a logarithmic scale. Specifically, the value of dT after each incorrect response was multiplied by 1.414, or divided by 1.414 after three successive correct responses was . Each measurement continued until six reversals were reached. For the last four reversals, the step size was reduced to 1.189. A measurement was considered as saturated when ten successive incorrect responses were provided with a dT value equal to 40 ms. The saturated measurements were assigned a threshold of 40 ms. If more than 50% of the measurements

were saturated, the participant's data were not included in the analyses. In contrast, if the measurement was completed, the geometric mean of the dT values for the last four reversals was used as the threshold estimate. Four such estimates of the threshold were made for each AV congruence and F0 condition, and the geometric mean of these estimates was used as the final value in the analyses. Each block consisted of four measurements (2 AV congruence $\times$ 2 F0) completed in a random order. An initial two blocks were considered training. A final four blocks were used to compute the individual thresholds for each condition.

D. Results

Two participants whose measures were saturated 75% and 62% of the time were not included in the analyses. Figure 5 shows the geometric mean of the dT thresholds across the eight remaining participants, as a function of the fundamental frequency and the AV congruence condition. As the measure is based on a detection threshold, the smaller the threshold, the better is the performance. A two-way repeated measures ANOVA was performed with fundamental frequency and AV congruence condition as factors. The fundamental frequency had a detrimental effect on performance, which was consistent with the findings of Experiment 1 [$F(1,7)=21.68$; $p<0.01$]. Increased AV congruence also had a significant detrimental effect on performance, as indicated by larger dT thresholds [$F(1,7)=5.85$; $p<0.05$]. This detrimental effect of AV cues is consistently reported for 7 out of 8 listeners. The interaction between the two factors was not significant [$F(1,7)=0.16$; $p=0.71$].

E. Discussion

The results from Experiment 2 confirmed the hypothesis that congruent lip gestures enhance primary streaming. This is consistent with the results of Experiment 1 in which a small effect of AV input on streaming was observed. The detection task used in the current experiment was likely more sensitive than that used in Experiment 1. In addition, the use of natural audiovisual speech in the current study likely increased the audiovisual congruence

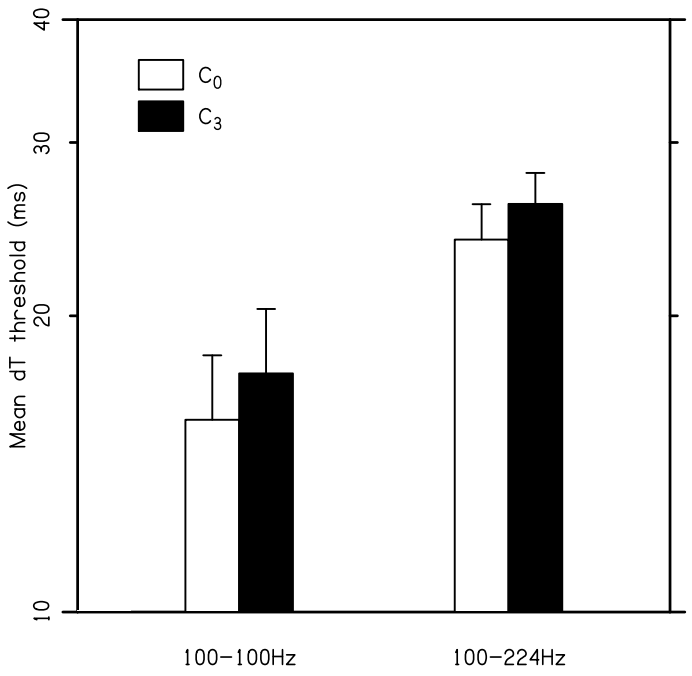


FIG. 5. Results of Experiment 2. The threshold for detecting a temporal offset between streams is plotted for each F0 difference and each visual condition across the eight participants. Errors bars represent one standard error.

and thereby strengthened the effect of the visual cue on auditory segregation. However, it should be noted that the AV stimuli in the current experiment were also altered. The natural AV desynchrony was reduced and the natural vowel durations were reduced to reach inter onset intervals values that ensure that streaming would occur Van Noorden (1975). As a consequence, the modest size of the effect in the current experiment could be related to a coherence that was still lower than in natural situations. Finally, as in Experiment 1, the interaction between the visual cues and the F0 was not significant: the effect of adding a visual cue when the F0 condition induced streaming was not significantly different from that when the F0 condition did not induce streaming. This suggests that fundamental frequency

difference and AV congruence act independently and may both contribute to the automatic segregation of competing speech.

IV. GENERAL DISCUSSION

Effect of lip gestures on obligatory streaming

The current findings suggest that part of the benefit of lip-reading in concurrent speech perception is a result of an interaction between primary segregation and visual input. In fact, visual lip gestures that are phonetically congruent with the auditory speech signal are capable of inducing obligatory streaming. As suggested by Arnal *et al.* (2009); Miller et D'Esposito (2005), who reported that two levels of interaction occur during the perception of AV speech, the interaction between auditory and visual input found in the current data could result from both low-level and high-level interactions.

Audiovisual congruence

The small effect of an audiovisual signal on segregation suggested by Rahne *et al.* (2007) and Rahne and Böckmann-Barthel (2009) was observed in both Experiments 1 and 2. Previous studies showed that a one dB improvement in speech to noise ratio could correspond to a 5-10% increase in intelligibility (Miller et Heise, 1950; Grant et Braida, 1991). Moreover, Grant and Seitz (2000) argued that a 2 dB AV effect on detection could be interpreted as a release from masking and could then lead to large increase in intelligibility. The small AV effect on streaming reported in Experiments 1 and 2 could then also potentially lead to a substantial effect on intelligibility in more realistic situations.

Because the effect was larger in Experiment 2, where AV congruence was greater, it is suggested that the audiovisual congruence of the stimuli may be important in eliciting the effect of visual cues on primary auditory segregation. The congruence of the stimuli may depend on two factors. First, the temporal congruence between the amplitude envelope of the auditory input and the visual input appears to be important. One difference between our

two experiments was the level of synchrony between the amplitude envelope of the auditory vowel input and the visual lip gestures. In Experiment 1, the auditory envelope remained at a constant level during the presentation of the lip gestures. In Experiment 2, the audio and visual parts of the signal were recorded simultaneously, and as such the audio temporal envelope was consistent with the lip gestures.

Second, the phonetic coherence between the auditory and visual inputs may have played a role in the congruence of the signals. In Rahne *et al.* (2007, 2008); Rahne and Böckmann-Barthel (2009), the auditory signals were sequences of pure tones and the visual signals were geometric shapes. Thus, the signals did not have phonetic content or phonetic congruence. This may have been one reason why Rahne *et al.* observed no visual effect when streaming resulted from a clear acoustic cue. In Experiment 1, vowels were generated with the Klatt algorithm and the visual lip gestures were synthesized from a limited number of video frames. In Experiment 2, the vowels and the lip gestures were simultaneously recorded, and the audio stimuli were modified using STRAIGHT. Consequently, both the phonetic content and the phonetic coherence were greater in Experiment 2 than in Experiment 1. The role of phonetic congruence in the effect of visual cues is in accord with Devergie *et al.* (2009), in which greater AV binding was found between auditory vowels and geometric shapes when the shapes changed in accordance with the natural lip gestures. Although it is worth noting that the effect of visual cues might simply be task-dependant and was more easily elicited in Experiment 2 than in Experiment 1.

Neurophysiological correlates

The effect of visual lip gestures on primary streaming observed here is supported by neurophysiological research. Studies have reported that primary auditory cortical areas are sensitive to visual stimuli. For example, changes in a speech signal in the visual modality can be processed in the primary auditory cortex (Moettoenen *et al.*, 2002). This area of the brain can also be activated by a purely visual speech signal (Calvert *et al.*, 1997; Pekkola

et al., 2005). In summary, the current data provide support for a multi-sensory contribution (i.e., lip-reading) to low-level auditory processing (i.e., primary streaming) as reported in a recent review of neurophysiological work (Schroeder *et al.*, 2008).

Concluding remarks and perspectives

The results from these experiments confirmed the hypothesis that congruent lip gestures enhance primary streaming. The effect is consistent across listeners and experiments but remains small. Moreover, audiovisual coherence seems critical to report any effect. In order to strengthen the effect of visual cues on primary streaming, futures studies could introduce more naturalness (natural synchrony, natural duration) and try to use VCV with more lips kinematics to enhance coherence, focus on F0 values leading to bistable percept and compare the duration of the build-up in various audiovisual coherence conditions either psychophysically or with EEG recording.

Acknowledgments

This work was supported by grants from the Région Rhones-Alpes Auvergne Cluster HVN 2007, the Agence Nationale de Recherche (ANR-08-BLAN-0167-01), and the National Institute on Deafness and Other Communication Disorders (DC08594). We are grateful to the participants of this study. We also thank Christophe Savariaux for his help in recording audiovisual material.

APPENDIX A: VIDEOFRAMES OF THE LIP GESTURES

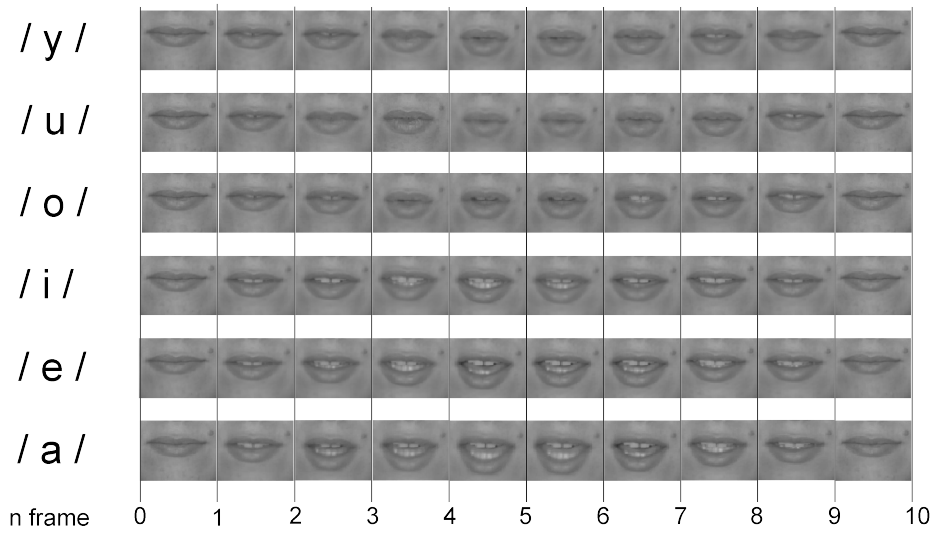


FIG. 6. Visual lip gestures in the Experiment 1.

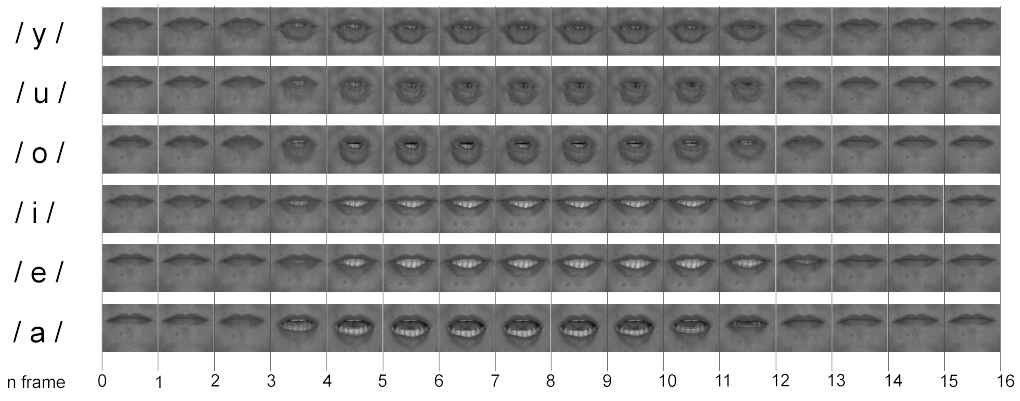


FIG. 7. Visual lip gestures in the Experiment 2.

Endnotes

1. An ANOVA focusing on only the central F0 range (between 121 and 196Hz) which can be assumed to elicit bistable auditory organization, revealed a larger effect of AV congruence [$F(2,18)=8.55$, $p< 0.005$]. 2. A Geisser Greenhouse correction was also performed to compensate for the lack of sphericity and provided the same pattern of results [F0: $\varepsilon(1.89, 17.02)=0.21$, $p<0.0001$, AV congruence: $\varepsilon(1.32, 11.86)=0.65$, $p=0.08$, interaction $\varepsilon(5.26,47.39)=0.29$, $p=0.51$].

REFERENCES

American National Standards Institute (**1995**). *ANSI S3.7-R2003: Methods for Coupler Calibration of Earphones*, American National Standards Institute, New York.

American National Standards Institute (**2004**). *ANSI S3.21-2004: Methods for Manual Pure-Tone Threshold Audiometry*, American National Standards Institute, New York.

Arnal, L. H., Morillon, B., Kell, C. A., and Giraud, A.-L. (**2009**). "Dual neural routing of visual facilitation in speech processing.", J. Neurosci. **29**, 13445–13453.

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

Berthommier, F. (**2003**). "A phonetically neutral model of the low-level audiovisual interaction", in *Proceedings of the International Conference on Audio-Visual Speech Processing*, 89–94 (St Jorioz, France).

Besle, J., Fischer, C., Bidet-Caulet, A., Lecaignard, F., Bertrand, O., et Giard, M.-H. (**2008**). "Visual activation and audiovisual interactions in the auditory cortex during speech perception: intracranial recordings in humans.", J Neurosci **28**, 14301–14310.

Brancazio, L. et Brancazio, L. (**2004**). "Lexical influences in audiovisual speech perception.", J Exp Psychol Hum Percept Perform **30**, 445–463.

Bregman, A. (**1990**). *Auditory Scene Analysis: The Perceptual Organization of Sounds* (MIT Press, Massachussets, USA).

Bregman, A. S. (**1978**). "Auditory streaming is cumulative.", J. Exp. Psychol. Hum. Percept. Perform. **4**, 380–387.

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

Cusack and Roberts (**2000**). "Effects of differences in timbre on sequential grouping", Percept. Psychophys. **62**, 1112–1120.

Devergie, A., Berthommier, F., and Grimault, N. (**2009**). "Pairing audio speech and various visual displays: binding or not binding ?", in *Proceedings of the International Conference on Audio-Visual Speech Processing*, 140–144 (Norwich, UK).

Gaudrain, E., Grimault, N., Healy, E. W., and Béra, J.-C. (**2007**). "Effect of spectral smearing on the perceptual segregation of vowel sequences.", Hear. Res. **231**, 32–41.

Gaudrain, E., Grimault, N., Healy, E. W., and Béra, J.-C. (**2008**). "Streaming of vowel sequences based on fundamental frequency in a cochlear-implant simulation.", J. Acoust. Soc. Am. **124**, 3076–3087.

Grant, K. W. et Braida, L. D. (**1991**). "Evaluating the articulation index for auditory-visual input.", J Acoust Soc Am **89**, 2952–2960.

Grant, K. W. et Walden, B. E. (**1996**). "Spectral distribution of prosodic information.", J Speech Hear Res **39**, 228–238.

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

Grant, K. W. (**2001**). "The effect of speechreading on masked detection thresholds for filtered speech.", J Acoust Soc Am **109**, 2272–2275.

Grant, K. W., van Wassenshove, V., and Poeppel, D. (**2003**). "Discrimination of auditory-visual synchrony", in *Proceedings of the International Conference on Audio-Visual Speech Processing*, 31–35 (St jorioz, France).

Grant, K. W., Wassenhove, V., and Poeppel, D. (**2004**). "Detection of auditory (cross-spectral) and auditory-visual (cross-modal) synchrony", Speech Communication **44**, 43–53.

Kawahara, H., Masuda-Katsuse, I., and de Cheveigné, A. (**1999**). "Restructuring speech representations using a pitch-adaptive time-frequency-smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds", Speech Commun. **27**, 187–207.

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

Klatt, D. (**1980**). "Software for a cascade/parallel formant synthesizer", J. Acoust. Soc. Am. **67**, 971–995.

Levitt, H. (**1971**). "Transformed up-down methods in psychoacoustics.", J. Acoust. Soc. Am. **49**, 467–477.

Massaro, D. W. and Cohen, M. M. (**1983**). "Evaluation and integration of visual and auditory information in speech perception.", J. Exp. Psychol. Hum. Percept. Perform. **9**, 753–771.

Micheyl, C. et Oxenham, A. J. (**2010**). "Objective and subjective psychophysical measures of auditory stream integration and segregation.", J Assoc Res Otolaryngol

Micheyl, C., Tian, B., Carlyon, R. P., and Rauschecker, J. P. (**2005**). "Perceptual organization of tone sequences in the auditory cortex of awake macaques.", Neuron **48**, 139–148.

Miller, G. A. et Heise, G. A. (**1950**). "The thrill threshold", J Acoust Soc Am **22**, 637–638.

Miller, L. M. et D'Esposito, M. (**2005**). "Perceptual fusion and stimulus coincidence in the cross-modal integration of speech.", J Neurosci **25**, 5884–5893.

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

Moore, B. C. J. and Gockel, H. (**2002**). "Factors influencing sequential stream segregation", Acta Acustica **88**, 320–333.

Pekkola, J., Ojanen, V., Autti, T., Jääskeläinen, I. P., Möttönen, R., Tarkiainen, A., and Sams, M. (**2005**). "Primary auditory cortex activation by visual speech: an fmri study at 3 t.", Neuroreport **16**, 125–128.

Pressnitzer, D., Sayles, M., Micheyl, C., and Winter, I. M. (**2008**). "Perceptual organization of sound begins in the auditory periphery.", Curr. Biol. **18**, 1124–1128.

LIST OF FIGURES

Chapitre 3

Effet de l'attention rythmique sur la ségrégation de mélodies intercalées

Dans cette seconde étude, nous avons souhaité tester la théorie proposée par Alain et Bernstein (2008). Nous avons proposé un paradigme d'identification de mélodies intercalées. Les participants devaient identifier des mélodies familères intercalées avec des mélodies inconnues. Deux conditions expérimentales ont été proposées. Dans une condition, la mélodie familière et la mélodie inconnue partageaient les mêmes plages de fréquences et de timbres. La séquence intercalée résultante était présentée sur un rythme aléatoire. Dans une seconde condition, le rythme de la mélodie inconnue était rendu régulié, introduisant ainsi une différence perceptive entre la mélodie familière et la mélodie inconnue. Les résultats de cette étude sont cohérents avec la théorie de Alain et Bernstein (2008). Lorsqu'il n'y avait aucune différence perceptive entre la mélodie familière et la mélodie inconnue, les auditeurs étaient tout de même capable d'identifier au dessus du hasard la mélodie familière présente, utilisant leur connaissance de la mélodie. Lorsqu'en plus de cela, il y avait une différence de rythme, ils étaient alors capables de supprimer la mélodie inconnue et d'identifier plus efficacement la mélodie familière par rapport à la condition sans différence perceptive.

Cet article a été publié dans JASA Express Letter *le 24 juin 2010.*

Effect of rhythmic attention on the segregation of interleaved melodies

Aymeric Devergie, Nicolas Grimault,[a] and Barbara Tillmann
Laboratoire de Neurosciences Sensorielles, Comportement et Cognition, CNRS UMR 5020, Université Lyon 1, 69366 Lyon Cedex 07, France
aymeric.devergie@olfac.univ-lyon1.fr; nicolas.grimault@olfac.univ-lyon1.fr; barbara.tillmann@olfac.univ-lyon1.fr

Frédéric Berthommier
GIPSA-Lab, CNRS UMR 5216, Domaine universitaire, 38402 Saint Martin d'Héres, France
frederic.berthommier@gipsa-lab.grenoble-inp.fr

Abstract: As previously suggested, attention may increase segregation via enhancement and suppression sensory mechanisms. To test this hypothesis, we proposed an interleaved melody paradigm with two rhythm conditions applied to familiar target melodies and unfamiliar distractor melodies sharing pitch and timbre properties. When rhythms of both target and distractor were irregular, target melodies were identified above chance level. A sensory enhancement mechanism guided by listeners' knowledge may have helped to extract targets from the interleaved sequence. When the distractor was rhythmically regular, performance was increased, suggesting that the distractor may have been suppressed by a sensory suppression mechanism.

PACS numbers: 43.66.Mk, 43.66.Ba [QJF]
Date Received: March 12, 2010 **Date Accepted:** May 4, 2010

1. Introduction

In everyday listening, sound events rarely appear in isolation. Usually, several acoustic streams issued from various sources compete with each other. According to Bregman (1990), segregation mechanisms based on listeners' knowledge and schemata stored in long-term memory can be used to extract a well-known target from an acoustic mixture. It is generally acknowledged that these extraction mechanisms are strongly related to attentional processes (Hafter *et al.*, 2003). Our present study is aimed at further elucidating the attentional involvement in these extraction mechanisms, referred to fission mechanisms (Moore and Gockel, 2002), for sequential stream segregation.

Numerous studies have implicated differences in pitch, loudness or timbre as mediating the low-level processing of segregation. In contrast, few studies have investigated schema-based processing *per se*. Focusing on fission mechanisms, Dowling *et al.* (1987) proposed an interleaved melody paradigm to test whether listeners are able to identify familiar melodies (targets) interleaved with unknown melodies (distractors). In one condition in which target and distractor melodies shared the same pitch range and timbre, listeners were still able to extract the target melody from the interleaved sequences. The authors concluded that listeners performed the fission task by using their prior knowledge of the target melody. In fact, in this study, participants knew which melody to listen to and to extract from the interleaved sequences. In this situation, the involved attentional processes are guided by familiar melodic schemata stored in memory, and these schemata likely help to extract the relevant information from an auditory mixture.

[a]Author to whom correspondence should be addressed.

In the study by Dowling *et al.* (1987), the interleaved sequences (i.e., target +distractor) started either with a tone from the target melody (on-beat condition) or a tone from the distractor melody (off-beat condition). Target identification was found to be 10% better in the on-beat condition than in the off-beat condition. This result is consistent with Jones' theory of rhythmic attention, which is described as a dynamic process that builds up temporal windows of expectation (Jones, 1976; Jones *et al.*, 1981, 2002). For example, Jones *et al.* (2002) showed that comparing the pitch of two tones was easier when the second tone occurred within an expected time window and was primed by a rhythmic eight-tone sequence. Furthermore, Dowling *et al.* (1987) showed that attentional processes were primed by starting the sequence with a tone of the target melody.

Rhythmic attention may have also contributed to performance in streaming studies testing the effect of the rhythm of presentation on sequential segregation. Van Noorden (1975) and George and Bregman (1989) addressed this idea using an experimental setup that biased listeners toward perceiving a single stream. In the Van Noorden (1975) study, listeners were asked to adjust the pitch difference of tones from two streams in order to perceive all tones integrated into a single stream. Listeners' adjustments were not found to be influenced by the regular or random nature of the tone rhythm. In the George and Bregman (1989) study, listeners were able to integrate two tone sequences into one single stream regardless of the tone rhythms. These findings suggest that the ability to integrate all events into one stream, reputed to be an automatic (bottom-up) process (Bregman, 1990), is independent of rhythmic attention. In contrast, the only study that investigated the effect of rhythm on the ability to selectively listen to part of an acoustic mixture (fission), reported a strong effect of rhythm (Jones *et al.*, 1981). This finding is consistent with Van Noorden (1975) who reported that fission can be influenced by top-down processes.

Using a similar experimental setup as Bregman and Rudnicky (1975), Jones *et al.* (1981) asked participants to judge the temporal order of two tones that were embedded in a sequence of captor tones with lower pitch. They found that performance was weaker when all tones were played with an isochronous rhythm at the same speed than when the probe tones and the captor tones were played at different rates of speed. Furthermore, the segregation of the probe tones induced by the pitch difference was enhanced by the tempo difference. The authors interpreted these results in terms of rhythmic attention.

In addition, for primitive, stimulus-driven segregation (Bregman, 1990), the effect of attention remains a matter of debate. Presuming that attention might be an all or none process, some authors have argued that primitive segregation can be influenced by attention (Carlyon *et al.*, 2001), while others have argued that it is purely pre-attentive (Sussman *et al.*, 2007). Two recent studies suggested that primitive segregation can be decomposed into pre-attentive and attentive mechanisms (Snyder and Alain, 2007; Cusack *et al.*, 2004). These attentional mechanisms, if they exist, would interact with stimulus-driven segregation and would be related to acoustic cues. For schema-based segregation, the effect of attention has generally been acknowledged (Bregman, 1990). However, while top-down attention could focus on a limited range of acoustic features (Hafter *et al.*, 2003), there are only few experimental data showing the effect of attention on schema-based segregation (Dowling *et al.*, 1987).

Despite the lack of experimental data, two theoretical frameworks on attention and segregation have recently been proposed (Fritz *et al.*, 2007; Alain and Bernstein, 2008). Fritz *et al.* (2007) suggested that two attentional mechanisms might be involved in auditory segregation, a bottom-up 'pop-out' process and top-down mechanism. In the bottom-up 'pop-out' process, acoustic features (i.e., pitch, timbre or rhythmic regularity) catch listeners' attention and enhance the processing of the relevant acoustic signals. The top-down mechanism is based on the development of expectancies derived from listeners' knowledge. Alain and Bernstein (2008) suggested a complementary theoretical background in which attention increases segregation via two mechanisms. The first mechanism enhances the processing of task-relevant material, and the second mechanism, a suppression mechanism, attenuates the processing of task-irrelevant material.

Our experiment was conducted to provide some new data to test hypothesis derived from the frameworks of Fritz *et al.* (2007); Alain and Bernstein (2008). Listeners were required

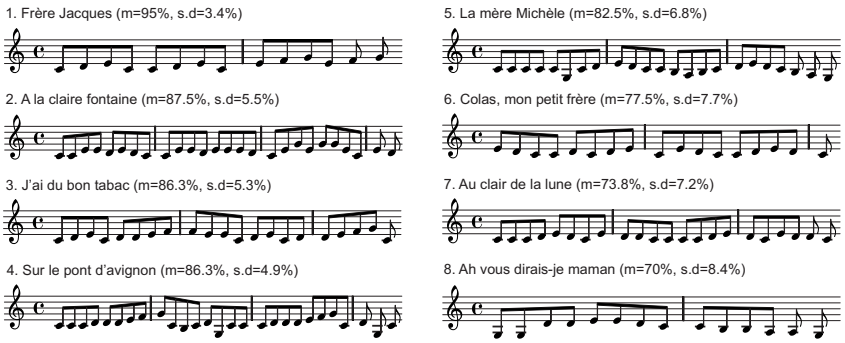


Fig. 1. Musical scores of the eight familiar French melodies. The results from the identification task where the melodies were presented alone are indicated in parentheses (mean percent of correct identification, standard deviation).

to extract a relevant target interleaved with a distractor. The only available cues for segregation were the knowledge about the target and, in some condition, the regularity of the rhythm of part of the signal, which needed to be ignored to test for suppression.

2. Experiment

2.1 Rationale

In the current study, an interleaved melody task was designed in which listeners were instructed to identify familiar target melodies embedded in distractor melodies sharing the same pitch and timbre ranges. Thus, neither pitch cues (Dowling *et al.*, 1987) nor timbre cues (Bey, 1999) would be useful for segregation in this task. Two conditions were utilized. In condition 1, the rhythms of target and distractor melodies were irregular. In condition 2, the rhythm of the target melody was irregular, while the rhythm of the distractor melody was regular. The rationale for using two conditions was twofold. First, in condition 1, identification performance at above chance levels would indicate an enhancement mechanisms based on knowledge. Second, in condition 2, any increase of performance relative to condition 1 would indicate a suppression mechanism that is strengthened by the regular rhythm of the distractor melody.

2.2 Apparatus

2.2.1 Participants

Twenty participants aged 18–30 years (mean=22.7, s.d=2.1) participated in the experiment. All participants were native French speakers and had pure tone audiometric thresholds below 15 dB HL at octave frequencies between 250 and 4000 Hz (American National Standards Institute, 2004). All participants were paid an hourly wage for their participation and signed an informed consent form. This study was formally approved by a local ethics committee (CPP Sud-Est II No. 06035).

2.2.2 Stimuli

Eight familiar French *target* melodies (displayed in Fig. 1) were selected and rendered isochronous. In addition, corresponding *control* melodies matched to each familiar target melody were constructed to analyze listeners' potential use of pitch range cues (see below). The control melodies were constructed by randomly selecting one temporal order of the notes of the familiar target melody among all order possibilities avoiding note repetition. In the interleaved melody task, each of the target and control melodies was mixed with a corresponding *distractor* melody (Fig. 2). The pitches of the distractor notes were chosen to be within the pitch range corresponding to the maximal pitch range across the eight familiar melodies (i.e., between 196 and 392 Hz, g2 and g3 in the musical scale). The target and control melodies were interleaved note-by-note

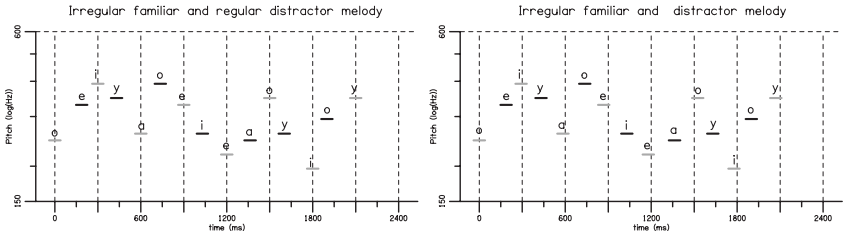


Fig. 2. Schematic representation of an interleaved sequence in the two rhythm conditions. An excerpt from the familiar melody 'Sur le pont d'Avignon' is represented by the black lines and 1 of the 32 possible distractor melodies is represented by the gray lines.

with distractor melodies. Pitches of the distractor notes were randomly chosen without repetition of successive notes.

All notes of all melodies lasted 80 ms, including 10 ms rising and falling ramps. Each of these notes was created with five French vowels /a/, /i/, /e/, /o/, /y/ with various pitches (Fig. 1). The vowels were generated using the KLATT algorithm (Klatt, 1980), and the specific spectral content of each vowel (i.e., formant positions) introduced some timbre variations across vowels within a sequence. As indicated by Singh and Bregman (1997), such timbre variation reduces the global perceptual coherence of the sequence and may contribute to segregation. Vowels of the interleaved sequences were randomly chosen without direct repetition of vowels for successive notes.

Because each vowel was used for target, control and distractor melodies, timbre was not a reliable cue for the segregation and identification task. Within each sequence, the pitch range of the target melody fell within the pitch range of the distractor melody. Moreover, each target melody shared exactly the same pitch range with the corresponding control melodies. To test for the listeners' use of pitch range cues in the segregation task, which would predict identification of a control melody as the associated target melody, we directly compared performance for control and target melodies.

Two different rhythm conditions were defined (Fig. 2). The right panel of Fig. 2 represents condition 1, in which notes of target or control melodies and notes of distractor melodies had random onsets. In this condition, the inter-onset-interval (IOI) separating two successive notes (first-order interval) was randomly chosen between 80 and 140 ms. The left panel of Fig. 2 represents condition 2, in which notes from the target or control melodies had random onsets, and notes from the distractor melodies had regular onsets. Temporal regularity, in the form of isochrony, was introduced by setting all time intervals separating the onsets of two successive distractor notes (second-order intervals) equal to 300 ms. As in condition 1, IOIs between a target note (or a control note) and a distractor note were randomly chosen between 80 and 140 ms. A chi-square analysis performed on the IOI distributions revealed that they were not significantly different ($\chi^2 = 19.81$, $p = 0.997$) for these two conditions. The target melodies were the same for the two rhythm conditions.

Each of the eight familiar melodies was repeated four times in association with a different distractor melody. Overall, 32 different distractors were generated, and the same distractor melodies were combined with the control and target melodies. For example, target melody 1 was associated to the same four distractors as control melody 1. The 64 resulting combinations were then duplicated for the 2 rhythm conditions, yielding 128 experimental trials. Thus, the same interleaved sequences were used in the two rhythm conditions. All interleaved sequences started with a distractor note, and not a target note (leading to the target melody being an offbeat melody, as in Dowling *et al.*, 1987). This was done to avoid that the first tone of the target melody might attract attention and help listeners to perform the task,

The only difference between the two conditions was the rhythm of the distractor stimuli played using a SIGMATEL internal sound card connected to a Sennheiser HD 250

Linear II headphone. Listeners were comfortably seated in a double-walled attenuated sound booth. Output level was calibrated to 70 dB SPL [Larson Davis AEC101 and 824; American National Standards Institute (1995)] with rms value adjustment between all vowels.

2.2.3 Procedure

Before the main experimental task (i.e., identification in the interleaved melodies), listeners performed first a familiarization task, then an identification task on the target melodies. In these two preliminary tasks, the target melodies were generated alone with an isochronous rhythm (IOI of 384 ms). In the familiarization task, the titles of the eight melodies were displayed and the listeners was instructed to listen to each melody as many times they want by selecting the corresponding title. In the initial identification task, the participants listened to the melodies in a random order and were instructed to identify the target melody as fast as possible. The titles of the eight melodies were displayed. An additional title 'unknown melody' was also displayed. Two participants did not reach 50% correct identification and were excluded from further testing. The identification performance averaged across the remaining participants is indicated for each melody in Fig. 1. A high percent of correct target melody identification was reached despite rhythms being isochronous. This finding is consistent with Hébert and Peretz (1997), who reported that pitch contour is a dominant feature used for identification of familiar melodies. This also suggests that identification should remain high in the interleaved melody task even if the rhythm of the target melodies was rendered irregular. In the interleaved melody task, all sequences were played in random order for each participant. Participants had to identify the target melody present in the sequence.

2.3 Results

Responses for control melodies were averaged for each participant. The results showed that 65% (s.d=5.2) of the control melodies were categorized as unknown melodies, 5% (s.d=2.4) were identified as the associated target melody and 4% (s.d=2) were identified as a different target melody than the associated target. A t-test applied to the two latter identification scores did not reveal a significant bias toward the associated target melody [$t(62)=1.29$; $p=0.2$]. These findings indicate that the pitch range of the familiar melody was not used as a reliable cue allowing melody identification. Therefore, neither pitch range nor timbre range aided in segregating target melodies from distractor melodies in our study.

Responses for target melodies were considered correct when the target melody was correctly identified. Figure 3 shows a plot of identification performance for each rhythm condition averaged over all participants. A t-test revealed that performance was significantly better than chance in condition 1 [$t(17)=2.87$; $p=0.01$] and in condition 2 [$t(17)=2.62$; $p=0.018$]. In addition, performance was significantly better in condition 2 than in condition 1 [$t(17)=2.28$; $p=0.036$], indicating an effect of rhythm presentation.

In condition 1, the only cue available for participants was their prior knowledge of the familiar melodies. To test whether listeners' prior knowledge could explain performance in the interleaved melody task, we evaluated the correlation between identification scores of individual melodies in the familiarization phase and identification scores in the interleaved melodies task. We found that better identification of the familiar melody presented alone tended to be correlated with better identification of the target melody in the interleaved sequences ($R^2=0.4655$, $p=0.0623$). Due to the small number of data points available, an additional *Monte-Carlo* simulation analysis was also applied to the correlation data. This analysis yielded a similar significance value ($p=0.055$).

3. Discussion

Familiar melody identification in the interleaved melody task was above chance regardless of the rhythm of the distractor and was better when the rhythm of the distractor was regular. In retrospect, these performance levels suggest that both the stimuli and the task were appropriate to test our hypotheses.

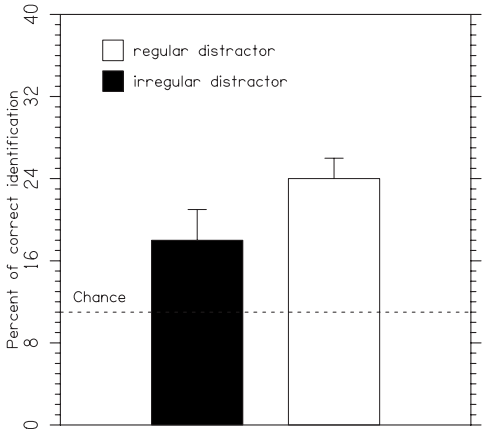


Fig. 3. Percent of correctly identified target melodies interleaved with irregular distractors (black bar) or regular distractors (white bar). Chance level is equal to 11% (1/9 possibilities). Error bars represent standard deviation.

Condition 1 was designed to test the theory proposed by Alain and Bernstein (2008) of the existence of an enhancement mechanism based on listeners' knowledge, involved in auditory fission. All previous studies measuring fission boundary used stimuli with either acoustic cues (Jones *et al.*, 1981) (for a review, see Fritz *et al.*, 2007) or rhythm cues (Dowling *et al.*, 1987). The better than chance performance in condition 1 provides the first evidence for a pure attentional fission mechanism based only on prior knowledge of a pitch sequence (decoupled from its original rhythm). In fact, listeners' knowledge was the only reliable cue that enabled identification of target melodies. Thus, segregation can be influenced by knowledge when the schemata are stored in long term memory [Dowling *et al.* (1987), this study]. In contrast, by using unfamiliar melodies presented just before the test to produce schemata stored only in short term memory, Bey and McAdams (2002) reported performances at chance levels when acoustical cues were lacking. These apparently divergent findings may be reconciled by the strength of knowledge being important for segregation, as previously hypothesized by Bey (1999). Our results are consistent with this hypothesis (Bey, 1999) as indicated by the positive correlation tendency between identification scores in the familiarization phase and identification in condition 1. The melodies were reputed to be familiar to native French speakers and, thus, more likely to be stored in long term memory.

In addition to the influence of knowledge stored in long term memory, Dowling *et al.* (1987) showed the influence of the rhythmic position of the target melody (i.e., on- or offbeat). The improved performance during the on-beat condition was consistent with rhythmic attention described by Jones (1976), as discussed in the Introduction. Although our target melodies were off-beat, our findings are consistent with the rhythmic attention theory. Jones *et al.* (2002) provided information regarding the size of the attentional window. Based on their data (Fig. 4 from Jones *et al.*, 2002), the size of the expectancy window, defined by a performance decrease of 10%, was a few tens of milliseconds. In our experiment, the rhythm of the target melodies was always irregular, but at least part of each target note fell within the expectancy window. Despite randomized IOIs and the hypothesis that expectancy windows could vary with contextual irregularity, the results of the current experiment were still consistent with the rhythmic attention theory.

Comparison of conditions 1 and 2 allowed testing for the effect of rhythmic attention on the suppression mechanism. Our data revealed that identification increased significantly when the rhythm of the distractor was regular. This finding is consistent with those of Demany and Semal (2002), who showed that second-order temporal regularity could be beneficial for

perception. In our experiment, this second-order regularity presumably helped to build a rhythm attention cycle (Jones *et al.*, 2002) synchronized to the notes of the distractor melody.

In sum, our current study provides the first evidence of pure attentional segregation based on knowledge that can be strengthened by rhythm regularity of the part of the signal that needs to be suppressed. These results are consistent with the rhythmic attention theory of Jones and Boltz (1989) and demonstrate the relevance of rhythmic attention for auditory scene analysis.

Acknowledgments

This work was supported by grants from the Région Rhones-Alpes Auvergne 'Cluster HVN 2007' and the Agence Nationale de Recherche (Grant No. ANR-08-BLAN-0167-01). Special thanks to Mary Riess Jones for very helpful comments on a previous version of the manuscript, to Jay Dowling for interesting suggestions and to Charlotte Dépalle for managing participants.

References and links

Alain, C., and Bernstein, L. J. (**2008**). "From sounds to meaning: The role of attention during auditory scene analysis," Curr. Opin. Otolaryngol. Head Neck Surg. **16**, 485–489.
American National Standards Institute (**1995**). "Ansi s3.7-r2003: Methods for coupler calibration of earphones."
American National Standards Institute (**2004**). "Ansi s3.21-2004: Methods for manual pure-tone threshold audiometry."
Bey, C. (**1999**). "Recognition of interleaved melodies and formation of auditory streams: Functional analysis and neuropsychological exploration," Ph.D. thesis, EHESS, Paris, France.
Bey, C., and McAdams, S. (**2002**). "Schema-based processing in auditory scene analysis," Percept. Psychophys. **64**, 844–854.
Bregman, A. (**1990**). *Auditory Scene Analysis: The Perceptual Organization of Sounds* (MIT, Cambridge, MA).
Bregman, A. S., and Rudnicky, A. I. (**1975**). "Auditory segregation: Stream or streams?," J. Exp. Psychol. Hum. Percept. Perform. **1**, 263–267.
Carlyon, R. P., Cusack, R., Foxton, J. M., and Robertson, I. H. (**2001**). "Effects of attention and unilateral neglect on auditory stream segregation," J. Exp. Psychol. Hum. Percept. Perform. **27**, 115–127.
Cusack, R., Deeks, J., Aikman, G., and Carlyon, R. P. (**2004**). "Effects of location, frequency region, and time course of selective attention on auditory scene analysis," J. Exp. Psychol. Hum. Percept. Perform. **30**, 643–656.
Demany, L., and Semal, C. (**2002**). "Limits of rhythm perception," Q. J. Exp. Psychol. A **55**, 643–657.
Dowling, W. J., Lung, K. M., and Herrbold, S. (**1987**). "Aiming attention in pitch and time in the perception of interleaved melodies," Percept. Psychophys. **41**, 642–656.
Fritz, J. B., Elhilali, M., David, S. V., and Shamma, S. A. (**2007**). "Does attention play a role in dynamic receptive field adaptation to changing acoustic salience in A1?," Hear. Res. **229**, 186–203.
George, M. F.-S., and Bregman, A. S. (**1989**). "Role of predictability of sequence in auditory stream segregation," Percept. Psychophys. **46**, 384–386.
Hafter, E. R., Sarampalis, A., and Psyche, L. (**2003**). *Springer Handbook of Auditory Research: Auditory Perception of Sound Sources* (Springer, New York), Vol. **5**, pp. 115–142.
Hébert, S., and Peretz, I. (**1997**). "Recognition of music in long-term memory: Are melodic and temporal patterns equal partners?," Mem. Cognit. **25**, 518–533.
Jones, M., Moynihan, H., MacKenzie, N., and Puente, J. (**2002**). "Temporal aspects of stimulus-driven attending in dynamic arrays," Psychol. Sci. **13**, 313–319.
Jones, M. R. (**1976**). "Time, our lost dimension—Toward a new theory of perception, attention and memory," Psychol. Rev. **83**, 323–355.
Jones, M. R., and Boltz, M. (**1989**). "Dynamic attending and responses to time," Psychol. Rev. **96**, 459–491.
Jones, M. R., Kidd, G., and Wetzel, R. (**1981**). "Evidence for rhythmic attention," J. Exp. Psychol. Hum. Percept. Perform. **7**, 1059–1073.
Klatt, D. (**1980**). "Software for a cascade/parallel formant synthesizer," J. Acoust. Soc. Am. **67**, 971–995.
Moore, B. C. J., and Gockel, H. (**2002**). "Factors influencing sequential stream segregation," Acta Acust. **88**, 320–333.
Singh, P. G., and Bregman, A. S. (**1997**). "The influence of different timbre attributes on the perceptual segregation of complex-tone sequences," J. Acoust. Soc. Am. **102**, 1943–1952.
Snyder, J. S., and Alain, C. (**2007**). "Toward a neurophysiological theory of auditory stream segregation," Psychol. Bull. **133**, 780–799.
Sussman, E. S., Horvath, J., Winkler, I., and Orr, M. (**2007**). "The role of attention in the formation of auditory streams," Percept. Psychophys. **69**, 136–152.
Van Noorden, L. P. A. S. (**1975**). "Temporal coherence in the perception of tone sequences," Ph.D. thesis, Eindhovern University of Technology, Eindhoven, The Netherland.

Chapitre 4

Spécificité du liage audiovisuel pour la parole : appariement d'indices visuels avec des séquences de voyelles audio

Dans cette troisième étude, nous avons proposé un paradigme expérimental original permettant d'évaluer le liage audiovisuel de voyelles auditives avec des indices visuels variés. Des séquences de six voyelles alternées en fréquence fondamentale ont été proposées. Simultanément, un indice visuel synchronisé avec une voyelle sur deux était présenté. Différents indices visuels élémentaires similaires à ceux proposés par Schwartz *et al.* (2003) ont été testés. Les participants devaient apparier une partie du flux sonore avec le flux visuel. Les performances d'appariement étaient supposées refléter la capacité de liage audiovisuel. Cinq expériences ont été menées. Lorsque nous avons introduit un indice phonétique, à savoir la dynamique d'ouverture des lèvres, les performances d'appariement étaient élevées pour tous les participants. En revanche, quand les indices visuels étaient de plus bas niveau, l'appariement était plus faible. De plus, les performances variaient selon les participants et le contexte de chaque expérience. L'appariement audiovisuel impliquant une sélection d'une partie du flux sonore semble donc

être dépendant de la cohérence audiovisuelle. Celle-ci apparaît comme étant particulièrement forte quand il s'agit d'un appariement phonétique. Cependant, cette étude suggère qu'il reste possible de lier un indice visuel non langagier avec des voyelles auditives. Il faut malgré tout être prudent concernant cette interprétation au vue de la grande variabilité rapportée. *Article en préparation.*

Article in preparation

On the specificity of audiovisual binding in speech

Pairing of visual displays with auditory sequences of vowels

Aymeric Devergie · **Frédéric Berthommier** · **Nicolas Grimault**

Abstract In the present study, we investigated audiovisual binding using a novel pairing paradigm. This paradigm consisted of selecting one of two interleaved audio triplets of vowels that was perceived to be in synchrony with a triplet of visual displays that was composed of 2 bars or of a disk. The triplets were presented for 10 s. Five experimental sessions were conducted. One session introduced a phonetic cue in the visual display: the mouth opening dynamic. The other sessions presented non-vowel-specific variations of contrast and movement; therefore, the pairing decision was only based on low level visual features. We also varied the envelope shape and the delta-f0-dependent streaming between the two vowel triplets. The level of performance of pairing was high for all of the participants in the experimental group that received a phonetic cue. In contrast, the performance was weak, subject-dependent, task-sensitive, and based on different features in each experimental session. These results revealed a clear, but unstable, capacity for pairing audio speech and non-speech specific visual displays, which depended on the overall temporal coherence of the low level cues.

Keywords Audiovisual binding · Speech specificity · Pairing · Temporal coherence · Auditory streaming · Audiovisual coherence

F. Author and T. Author
Laboratoire de Neuroscience Sensorielle, Comportement et Cognition, CNRS UMR 5020, Université Lyon 1, Lyon, France
Tel.: +33-4-37-28-74-91
Fax: +33-4-37-28-76-01
E-mail: aymeric.devergie@olfac.univ-lyon1.fr

S. Author
GIPSA-Lab, CNRS UMR 5216, Université Stendhal, Grenoble, France

1 Introduction

Until the last decade, the results of studies that investigated the binding of auditory and visual events supported the idea that speech was special. On the one hand, speech has been shown to elicit tolerance to incoherence, especially for spatial and synchrony attributes. Spatial and temporal incoherence between auditory and visual speech has been shown to lead to classical ventriloquism effects (Radeau and Bertelson, 1977; Bertelson and Aschersleben, 2003). Others studies have reported that people are not sensitive to onset asynchronies for audio and visual speech events if these events occur within the temporal window of integration (Massaro et al, 1996; Grant et al, 2003) (a few hundred milliseconds). On the other hand, contextual cues of binding such as gender coherence (Green et al, 1991) or emotional contents of audio (voice) and visual (face) speech are also specific for speech perception (de Gelder et al, 2002).

Although most of the results from the literature about audiovisual speech perception have not been interpreted in terms of binding, studies that have introduced lip-reading to better identify (Sumby and Pollack, 1954; MacLeod and Summerfield, 1987) or to better detect (Kim and Davis, 2003; Bernstein et al, 2004) speech in noise may have been related to binding as a supplementary condition that was established before speech identification, occurring somewhere in low levels of processing (Berthommier, 2004).

In the literature, binding has been explicitly addressed in studies involving basic stimuli. For example, binding was the focus of studies that employed very elementary signals, such as beeps and flashes, to manipulate physical binding with more accuracy (Bertelson and Aschersleben, 2003; Recanzone, 2003). To fill the gap between the audiovisual speech perception studies and the studies dedicated to binding, Schwartz et al (2003) proposed to investigate the visual benefit reported in lip-reading studies that used elementary visual features that mimicked lip movements. The audio syllables were embedded in several conditions of signal to noise ratios and were dubbed with a visual rectangle that had a vertical extension that varied in synchrony with the auditory envelope of the speech signal. Under such experimental conditions, only the detection of the syllables remained enhanced by visual cues. Binding also occurred as a condition for enhancing detection. Last, the nature of the binding, in terms of the cues of the binding, remained largely unclear.

In the current study, we investigated the binding of auditory tokens (French vowels) with various visual displays that were related to lip movements, e.g., two moving bars or a disk. The cues of the binding were assessed by introducing contrast variations and movements with several degrees of temporal coherence with the audio part of the signal. In only one experimental session, the visual stimuli carried vowel-specific information, whereas in the other sessions, the physical variations were not vowel-specific. Another important aspect of binding may be the relationship between the perceived auditory object that results from a sound analysis (Bregman, 1990) and the perceived visual object. Keetels et al (2007) estimated the potential effect of auditory fission (van Noorden, 1975) on audiovisual binding by evaluating the effect of temporal ventriloquism. In the visual modality, two flashes appeared consecutively with a constant stimulus onset asynchrony (SOA). In the auditory modality, two pure tones with the same frequency were played. In the experimental control context, the flashes and beeps were time-aligned. In the context of temporal asynchrony, the first beep was played 100 ms before the first flash, and the second beep was played 100 ms after the second flash. This configuration induced a temporal ventriloquism effect and led to a perceptual increase of the SOA between the flashes. In this previous experiment, the two beeps were either integrated in a flanking sequence of tones with the same frequency or were segregated from the flanking sequence (and the tones were of different frequencies).

The authors reported a temporal ventriloquism effect only when the two tones were segregated from the flanking sequence. They concluded that auditory fission was a requirement for audiovisual binding. In other words, these results might suggest that binding is enhanced when the auditory events to be bound with the visual events are clearly segregated from the background.

Cross-modal interaction investigations have shown that the unity assumption is fundamental in the binding process (Welch, 1972; Vatakis and Spence, 2007). According to this theory, the auditory and visual events were thought to be bound together when the physical properties (such as the spatial or synchronal attributes) and the cognitive aspects (such as the gender and affective content) were congruent (Schutz and Kubovy, 2009). Binding may also be influenced by low level features, such as the temporal correlation between the sound envelope and the contrast variation, and by the experimental context of the presentation of the stimuli.

To further understand the binding of speech, we proposed a study with a novel behavioral task that involved the pairing of a visual cue with an auditory target (French vowels) that was embedded in a sequence. The performance of the correct pairing between visual and auditory inputs was hypothesized to reflect the strength of the binding that occurred. The sequences of the vowels that were repeated in a loop were generated with two alternating fundamental frequencies to control the auditory fission. Depending on the fundamental frequency difference, the sequences of vowels were either perceived as a single integrated stream or as two segregated streams. These auditory sequences were dubbed with visual elementary shapes that varied along movement or contrast dimensions in synchrony with the presentation rate of the auditory vowels. The physical congruity between the audio and visual parts of the signal was also manipulated in an attempt to modulate binding. The decision process associated to the pairing task was also important in the experiment.. When the subject heard the sequences that were running in a loop for approximately 10 s, he or she could easily switch from one percept to the other and decide which one was the most coherent. The pairing paradigm in the present study was a new proposal that included some differences from the paradigms that were involved in grounding the unity assumption (namely the TOJ and the SJ) because the target stimulus was not presented in isolation. Moreover, in the experimental sessions in which the visual stimuli did not carry phoneme-specific information, the available information was the synchrony between the physical characteristics of the audio and video signals; therefore, the two alternatives were equally supported by the unity assumption, except for their synchronization (one was in phase, whereas the other was out of phase). The task was ambiguous, the decision process was difficult, and the expected scores were not above 50%. This method was used to exploit this perceptual ambiguity to measure the strength of the binding.

2 Rationale

2.1 Participants

Seventy participants, aged 18 to 30 years, were included in these experiments. All of the participants were native French speakers and had pure tone audiometric thresholds below 15 dB HL at octave frequencies between 250 and 4000 Hz (American National Standards Institute, 2004). All of the participants were paid an hourly wage for their participation and signed an informed consent. This study has been formally approved by a local ethical committee (CPP Sud-Est II No. 06035).

2.2 Stimuli

2.2.1 Sequences of vowels

The sequences of six different vowels were built using a *Matlab* routine (example of one sequence: /a e i o y u/). The vowels were synthesized using a Klatt algorithm (Klatt, 1980). The duration of the presentation of each vowel was 184 ms, including a 10-ms raise-cosine rising and falling ramp. The sequences were repeated in a loop, leading to a total duration of 10 s for each run. In addition, a linear ramp of 1104 ms was added at the beginning of the sequence to prevent the participants from detecting the first vowels of the sequence and biasing their answer toward them. The first auditory pattern, defined as aF0 (alternated F0), proposed sequences with an alternating pitch between 100 and 134 Hz. This large difference in pitch led to a clear perception of the two streams of the three vowels without any effort. The second auditory pattern, defined as cF0 (constant F0), led to the perception of a continuous sequence. However, the rhythm was regular and fast and was easy to split in two streams by choosing one vowel out of two. This process is known as rhythmic selection or rhythmic fission.

2.2.2 Visual displays

The frames were composed of a black background and three different white shapes (a disk, vertical bars, and horizontal bars). These three shapes were used to build movies at a rate of 60 frames/s. Two different types of visual displays were proposed. In the contrast condition, the color of the shape alternated between black and white. In the movement condition, the radius of the disk varied (leading to the perception of a looming disk), and the spacing between the bars varied. Figure 1 represents the two different visual displays with the three different shapes. The three frames for each shape represent the three different snapshots of the movies. These frames were later named targets. Thus, the targets were open movement, closed movement, or white contrast. Each group was instructed to detect different specific targets.

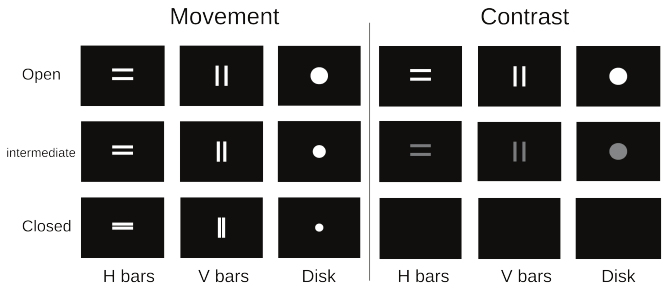


Fig. 1 Visual displays grouped by type (Movement or Contrast) and shape (horizontal bars, vertical bars, and disk). The movement and contrast varied between several targets. The participants were asked to detect these targets

2.2.3 Audiovisual synchronization

Figure 2 represents the audiovisual synchronization between the auditory vowels and the visual displays. One cycle of variation of the visual display corresponds to the amount of time that separated the two open targets or the two closed targets. One visual cycle covered two audio vowels. The center of one audio vowel was synchronous with the open target, and the center of the following vowel was synchronous with the closed target. The duration of a vowel was 184 ms; therefore, two consecutive audio vowels could be perceived in synchrony with the open target. For speech, the audiovisual integration window is reputed to be quite large, or approximately 350 ms (Massaro et al, 1996), which is similar to the duration of the visual period. Moreover, for vowels that are pronounced in isolation, the timing of the natural vocalic cores was poorly related to the lip gestures. This poor relationship allowed us to build a pairing paradigm that was based on this perceptual ambiguity.

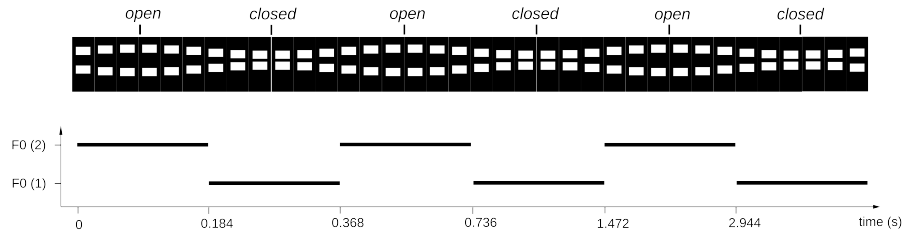


Fig. 2 Representation of an audiovisual sequence timeline. On the upper part of the panel, the visual display is the target of the movement with the horizontal bars shape (Hbar). On the lower part of the panel, the auditory sequence of vowels is condition aF0, for which F0(1) is equal to 100 Hz, and F0(2) is equal to 134 Hz. The closed and open targets were centered on the middle of the corresponding auditory vowel.

2.2.4 Method

Task The participants were instructed to pair the auditory vowels that appeared in synchrony with one particular target of the visual display that was open or closed. The sequence of audio vowels contained six vowels, and this sequence was presented in a loop without a silent period. The use of this presentation led to the possibility of pairing only two different vowel triplets (one in synchrony with one target, and the other in synchrony with the other target). These two triplets were displayed at the end of the presentation. For each visual display, we asked the participants to watch a specific target that was open or closed. These targets varied across the groups of participants. After the presentation of a sequence for 10 s, the participants were instructed to pick one of the two possible triplets that they believed to be synchronized with the target. During the 10 s that the sequence was presented, the participants were able to perceptively switch between the two alternatives to ascertain their degree of coherence. The participants were divided in five different groups. Each group was tested with different sets of conditions.

3 Group 1

3.1 Features

3.1.1 Visual displays

The aim of the present study was to determine which visual display would help the participants to pair the visual and auditory inputs. We choose to decompose the visual display into two types. Because previous neural studies (e.g. Vaina, 1994) have outlined that contrast and movement may be processed differently, we proposed that these two types of low level visual variations exist. For each type, there were three shapes, i.e., horizontal bars, vertical bars, and disks (as represented in Fig 1). The two horizontal bars moved relative to each other and were suggestive of the shape of lips and of the movement of lips, and the disk varied in size. We also proposed a third condition in which the bars were displayed vertically to test whether this orientation produced any effect in pairing. The participants that were shown the open-closed movement displays were asked to watch the closed target. The participants that were shown the black-white contrast displays were asked to watch the white target. We made the assumption that the detection of the closed state was easier than the detection of the open state because the perception of the bars touching was more salient. The prediction of the timing of the closing of the bars was also easier than the opening because the closed state presented an event that is more predictable. In other words, the temporal representation of the movement was more precise and sharper for the closed target.

3.1.2 Fundamental frequency differences

The pairing task involved the process of auditory fission so that one could appreciate the audiovisual temporal coherence for each stream separately. Because a difference in the fundamental frequency between the auditory events produces pre-attentive auditory streaming, the levels of performance were expected to be enhanced compared to the other condition, in which each stream is built rhythmically by picking one vowel out of two. In the constant F0 condition (cF0), all of the vowels had the same fundamental frequency of 100 Hz. In the alternated F0 condition (aF0), the vowels alternated in fundamental frequency between 100 and 134 Hz. This alteration led to the perception of a segregated sequence of two streams of three vowels with one unique F0 each.

3.1.3 Procedure

For each group, the participants first took part in a presentation session. During this session, all of the different combinations of the visual displays and F0s were presented in a random order. Next, the participants took part in a training session. Each combination of the visual displays and F0s was repeated 4 times and were randomized over the course of the session. Lastly, the participants took part in the test session. The test session contained 20 repetitions of each visual display and F0 (and modulation for groups concerned) combination that were distributed in 4 blocks with equal probability. The runs were randomized in each block, and the block order was randomly varied across the participants.

3.2 Results

The panels in Figure 3 show the performance for each participant as a function of the target and F0. In each panel, one feature was kept constant. For all of the panels, the subject performance averages of correct pairing were ranked from the best to the worst. In the upper left panel, the F0 was constant and was equal to 100 Hz. In the upper right panel, the F0 alternated between 100 and 134 Hz. In the lower left panel, the visual feature was the contrast display. In the lower right panel, the visual feature was the movement display. Two-way repeated measures analyses of variance (ANOVA) were performed with the fundamental frequency conditions (cF0 and aF0) and the type of visual display (movement and contrast) as the parameters. The performances were averaged across the visual shapes because no statistical differences were found between them. This analysis revealed a statistical effect of the type of visual display because the performances were better for the movement displays than for the contrast displays [$F(1,19)=21.46$, $p<0.01$]. The F0 separation had no effect on pairing [$F(1,19)=3.59$, $p=0.07$]. No interaction was found between the type of the display and the F0 [$F(1,19)=1.15$, $p=0.29$].

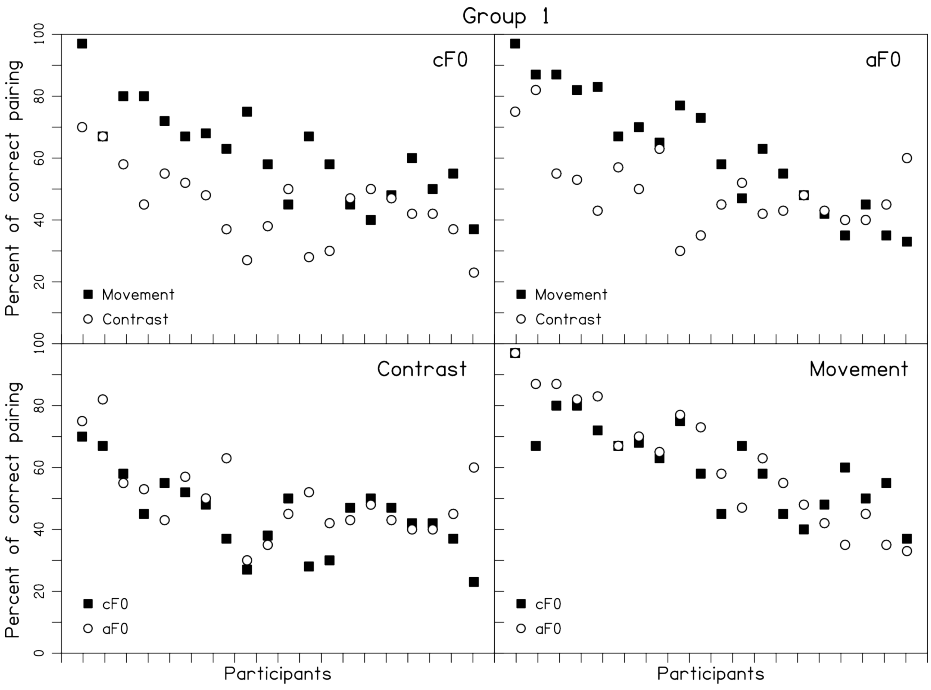


Fig. 3 Group 1. Performance of pairing for each participant averaged across repetitions for each feature. The participants were ranked based on their performance averages across all of the conditions and repetitions. The results for the two types of visual targets for each F0 condition are represented in the two upper plots. The results for the two F0 conditions for each type of visual target are represented on the two lower plots.

3.3 Discussion

First, pairing seemed to be easier for the visual target of a closing movement than for an opening movement. The pairing task may be interpreted in terms of a task of time alignment between the auditory and visual streams. Therefore, the participants were more accurate in pairing the two streams when they were presented with a visual display that was highly predictable. Second, the F0 separation did not enhance the performance of pairing compared to the performance without a difference in the F0. No significant effect of the type of display was found, and no difference was observed between the vertical bars and the others (horizontal bars and the disk). The similarity of the visual image to the shape of lips was not a strict condition, but we cannot rule out a role of this resemblance. Remarkably, the difference in the fundamental frequency had no significant effect on performance

4 Group 2

To test whether the enhancement in pairing of the movement cue was independent from the direction of movement (i.e., closing vs. opening), the second experiment involved the same features (i.e. visual cues and F0 cues), but the participants were instructed to pair the auditory stream with the visual target that exhibited an open movement.

4.1 Features

4.1.1 Visual displays

The visual features were the same as for group 1. For the open-closed movement displays, the participants of this group were asked to watch the open target. For the black-white contrast displays, the participants were asked to watch the white target.

4.2 Results

Figure 4 plots the performance for each participant as a function of the visual target and F0. The analysis was similar to experiment 1. This analysis revealed a statistical effect of the F0 separation for which the performances were better for the alternated F0s than for the constant F0s [$F(1,19)=13.70$, $p<0.01$]. In contrast to group 1, the visual display had no effect on the pairing [$F(1,19)=2.22$, $p=0.15$]. No interaction was found between the type of the display and the F0 [$F(1,19)=0.06$, $p=0.80$].

4.3 Discussion

The only difference between group 1 and group 2 was the instructions that were given to the participants. In group 2, we instructed the participants to pair the open target with the auditory stream. The poor representation of the timing of the opening that was discussed in a previous section may explain the loss of the enhancement in pairing that was observed with the movement cue. The participants may have had difficulty determining the synchronization points between the open targets and the corresponding triplet of vowels. Moreover, the F0

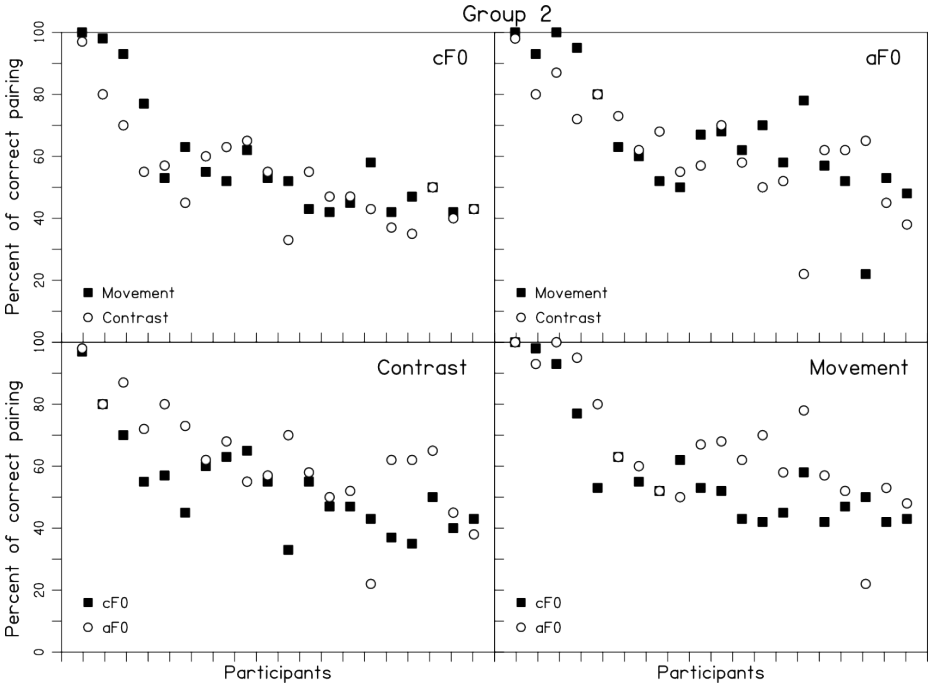


Fig. 4 Group 2. Performance of pairing for each participant averaged across repetitions for each feature. The participants were ranked based on their performance average across all of the conditions and repetitions. The results for the two types of visual targets for each F0 condition are represented on the two upper plots. The results for the two F0 conditions for each type of visual target are represented on the two lower plots.

separation was found to aid in pairing in group 2. The participants likely took advantage of the only available cue to perform the pairing task. Interestingly, the poor video temporal information could not be aligned with the audio when the streaming was not automatic. One hypothesis is that the rhythmic fission did not provide the temporal information necessary for binding on the audio side, whereas the streaming did provide this information.

5 Intermediate conclusion

Presented with ambiguous audio and visual streams that need to be bound to correctly perform the task, the participants exhibited a large amount of variability, and several subjects exhibited levels of performance that were not significantly different from chance. Depending on the task, the available cue was used adaptively. The results were globally significant, and we concluded that low level cues allowed participants to pair an auditory stream of vowels with a visual stream of geometric shapes that varied in time. A small amount of low level binding occurred and depended on the temporal information that was carried by the elementary features.

6 Group 3

The task of pairing could be considered to be a task of the alignment of anchor points or of the alignment of the temporal information. In the first case, only the points in time, audio and visual, were paired, whereas in the second case, the format of the temporal representation also played a role. This idea is supported by the results of experiments 1 and 2. We were interested in manipulating the envelope modulation of the auditory stream of vowels. In the two following experiments, we proposed different manipulations of the auditory envelope. Pairing may also be influenced by temporal co-modulations of the auditory and visual inputs. The more the inputs were perceived as synchronous, the greater the amount of the binding, and as a consequence, the better the pairing was. To test for the role of auditory envelope manipulation on the pairing performance, we used different conditions of envelope modulation.

6.1 Features

Similar to groups 1 and 2, the two conditions of the F0 separation were used (i.e., cF0 and aF0).

6.1.1 Visual displays

The visual displays used for group 3 consisted of horizontal bars for the movement display and a disk for the contrast display. For the movement display, the participants were asked to watch the closed target. For the contrast display, the participants were asked to watch the white target.

6.1.2 Modulation

Two conditions of modulation of the auditory envelope were used in the present experiment. The first one consisted of a flat envelope (flat): no modulation was applied. The second condition was a modulated condition (mod): a triangular window was applied to each vowel of 184 ms long. Figure 5 represents the variation of the visual display from open to closed and from white to black (in the same direction as the changes) and the modulation of the envelope of the audio stream below. For the flat and mod conditions of modulation, the envelope was applied to each 184 ms-long vowel.

6.2 Results

Figure 6 plots the performance for each participant as a function of the target and F0. In the upper left panel, the F0 was constant and was equal to 100 Hz. In the upper right panel, the F0 was alternated between 100 and 134 Hz. In the lower left panel, the modulation cue was the flat cue. In the lower right panel, the modulation cue was the modulated cue. Two-way repeated measures analyses of variance (ANOVA) were performed with the fundamental frequency condition (cF0 and aF0), the type of visual display (movement and contrast), and the modulation type (flat and mod) as the parameters. This analysis revealed a significant effect of the type of the visual display for which the performances were better for the movement than for the contrast conditions [$F(1,9)=11.6$, $p<0.01$]. Neither the F0 [$F(1,9)=2.33$,

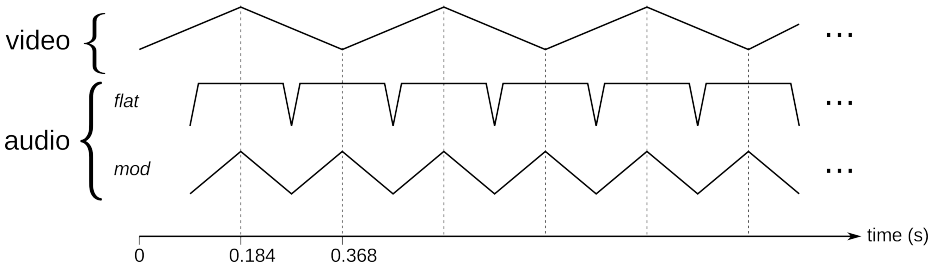


Fig. 5 Top: Variation of the visual display from open to closed or from white to black (in the same direction as the changes). Bottom: Modulation of the envelope of the audio stream.

p=0.16] nor the modulation [F(1,9)=0.267, p=0.627] had an effect on pairing. No interaction was significant, except for that between the type of the visual target and the modulation [F(1,9)=7.14, p<0.05].

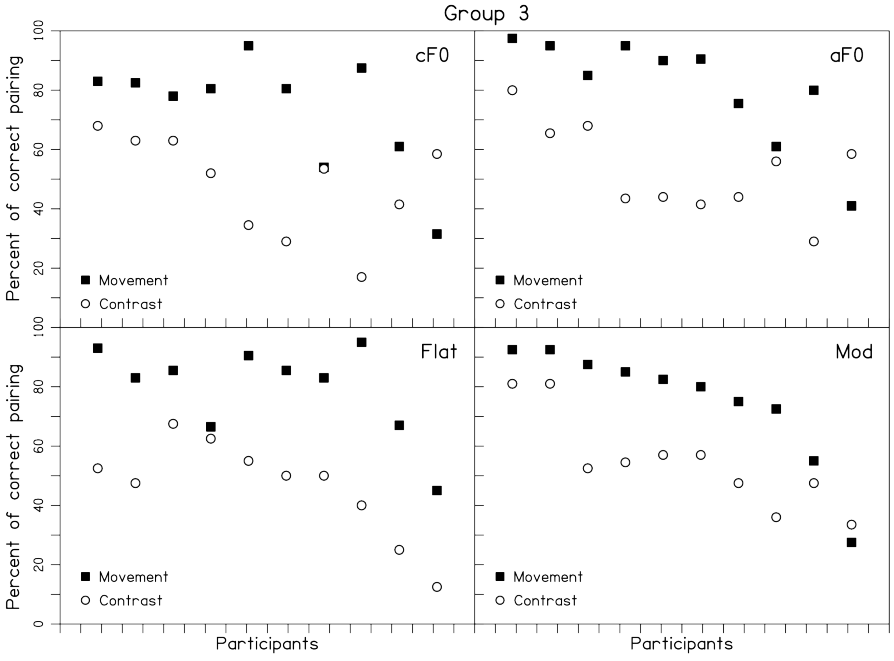


Fig. 6 Group 3. Performance of pairing for each participant averaged across repetitions for each feature. The participants were ranked based on their performance averages across all of the conditions and repetitions. The results for the two types of visual targets for each F0 condition are represented in the two upper plots. The results for the two modulations for each type of visual target are represented in the two lower plots.

6.3 Discussion

The data were consistent with the data from group 1, and we found a significant effect of the movement of the bars. These data did not show a pure effect of modulation on pairing but did show an interaction between the modulation and the visual factor. In Fig 5 (lower panel), we show that modulation (mod condition) negatively affected pairing with the movement cue compared to the flat condition. This result might be due to the temporal course of the modulation. In the mod condition, the audio envelope was not well correlated with the visual display, but the maximum pairing value indicated that the synchronization point was a better anchor point set on the audio side than that in the flat condition. A negative interaction between the modulation and the visual cue was observed, and as a result, we conclude that the maximum value was not the only cue used to perform the alignment. The entire envelope may have also had a role.

7 Group 4

The following experiment was proposed to test the effect of co-modulation between the auditory and visual streams on pairing. We also aimed to build auditory sequences with vowels that had audio envelopes with a closer temporal correlation to the video variation.

7.1 Features

7.1.1 Enhancing the temporal correlation between audio and video

Because the duration of the video period was twice that of the initial vowel duration, the successive vowels were overlapped by half of their duration. The variation was also slower and easier to follow. To overlap the vowels, the durations of the vowels were doubled to 368 ms. A triangular envelope on a log scale was applied to these long vowels, leading to a high degree of overlap between the vowels (modulation condition: high). A sharpened envelope was also used, leading to a low degree of overlap (modulation condition: low) (Figure 7).

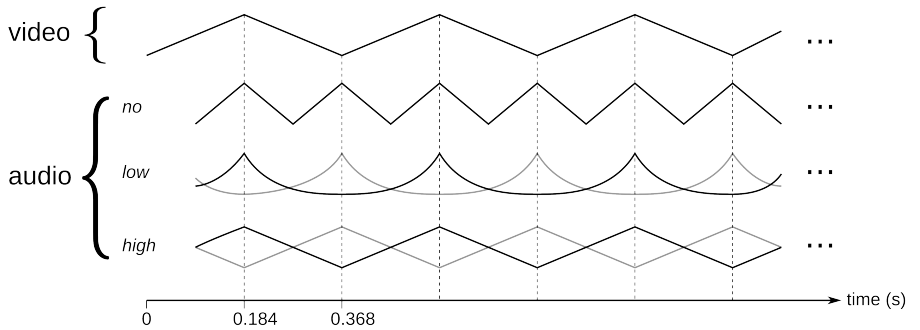


Fig. 7 Top: Variation of the visual display from open to closed or from white to black (in the same direction as the changes). Bottom: Modulation of the envelope of the audio stream.

7.2 Results

Figure 8 plots the performance for each participant as a function of the target and the F0. Two-way repeated measures analyses of variance (ANOVA) were performed with the type of visual display (movement and contrast) and modulation (mod, low, and high) as parameters. This analysis revealed a statistical effect of the type of the visual display in which the performances of pairing were better for the contrast displays than for the movement displays [$F(1,9)=9.76$, $p<0.05$]. Neither the F0 [$F(1,9)=0.03$, $p=0.85$] nor the modulation [$F(1,9)=1.29$, $p=0.29$] had an effect on pairing. No interactions were found between any of the parameters.

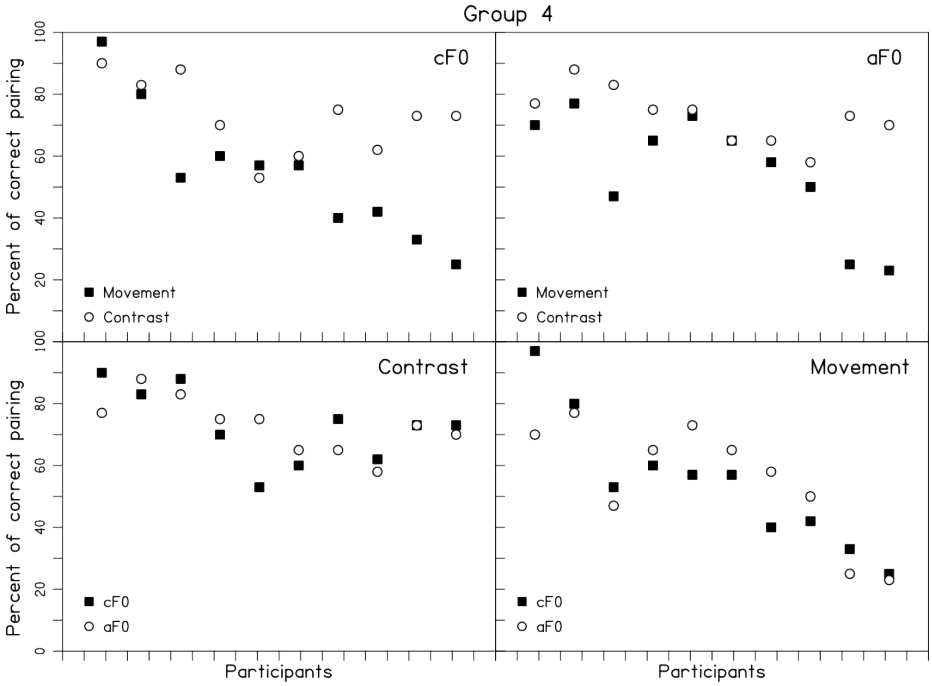


Fig. 8 Group 4. Performance of pairing for each participant averaged across repetitions for each feature. The participants were ranked based on their performance averages across all of the conditions and repetitions. The results for the two types of the visual targets for each F0 condition are represented in the two upper plots. The results for the two F0 conditions for each type of visual target are represented in the two lower plots.

7.3 Discussion

In contrast to group 1 and group 3, the movement cue was no longer used. Instead, the contrast display elicited the best performance of pairing. We conclude that the temporal coherence of the audio envelope with the contrast variation was well exploited by the participants. The performances may also be attributed to the slowing of the audio envelope variation. However, there was no difference between the three envelope conditions, including the modulated condition, which was present in group 3. This condition included only

one-third of the set of stimuli. However, if the pairing task was stimuli driven, we would have observed the same effect as in group 3: a significant level of pairing in the close movement condition. We conclude that participants globally adapt to the task and select the cue that was the most efficient to perform the pairing task. This conclusion is in agreement with the results of experiments 1 and 2 that focused on the task dependence and is also in agreement with the participants poor performances. The pairing task was not a classical stimulus-response paradigm, and the participant had to deliberate for a long time (10 s) before choosing the triplet that was more coherent with the visual display. Because the pairing task was difficult, and because two different conditions were present in the same experiment, one of the participants strategies was to cut off the half of the stimuli that were perceived to be difficult to concentrate his or her effort on the other stimuli.

8 Group 5

Previous experiments have focused on physical features such as the contrast, movement, or auditory cues, such as the F0 or the envelope modulation. In the current experiment, we introduced vowel-specific information in the visual display that was limited to the temporal course of the lips movements and the degree of the aperture of the lips. In this experiment, the goal was to build a continuum of stimuli from the least similar stimulus of the lips or a looming disk, to the most similar stimulus, the horizontal bars that exhibited the dynamic of the aperture of the lips. In this experiment, videos of a male speaker articulating the six vowels were recorded, and the real positions of the upper and lower lips were extracted for the temporal course of the vertical aperture parameter. The degree of aperture was also preserved so that a limited amount of information was displayed, allowing for identification of the vowel. It was clear that pairing would be achieved if we also displayed the horizontal aperture because this aperture was based on complete identification. Using only the aperture parameter will force the participant to combine several partial vowel identifications and/or to perform a low level binding, as in the other experiments. To test these two alternatives, we introduced a continuum of 3 stimuli (Figure 9) and two conditions of the audio envelope (flat and mod). Three phonetic movement displays were proposed (Figure 9): a disk with a varying radius that varied in a similar manner to the spacing between the upper and lower lips (disk), horizontal bars centered on the screen that reproduced the relative movements of the lips (1Dsym), and horizontal bars that reproduced the absolute movement of the lips with a fixed upper lip position (1D). The participants were instructed to watch the open target for all of the phonetic displays.

8.1 Results

Figure 10 plots the performance for each participant as a function of target and modulation. Two-way repeated measures analyses of variance (ANOVA) were performed with the type of the visual display (Disk, 1DSym, and 1D) and the modulation (flat and mod) as the parameters. This analysis revealed no statistical effects of the type of the visual display on performance [$F(2,18)=1.94$, $p=0.18$]. No effect of modulation on performance was observed [$F(1,9)=0.005$, $p=0.94$]. No interaction was found between the type of the display and the modulation [$F(2,18)=0.012$, $p=0.98$].

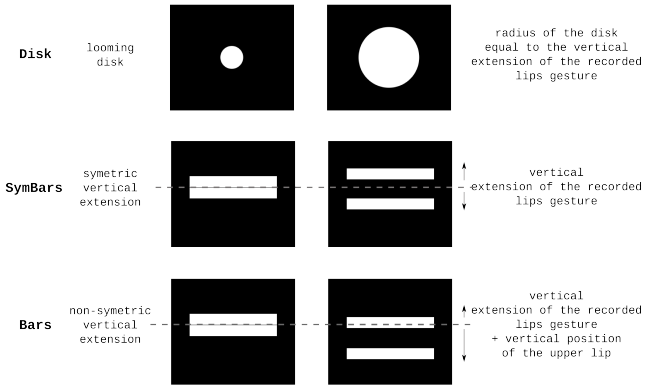


Fig. 9 Visual displays with phonetic content

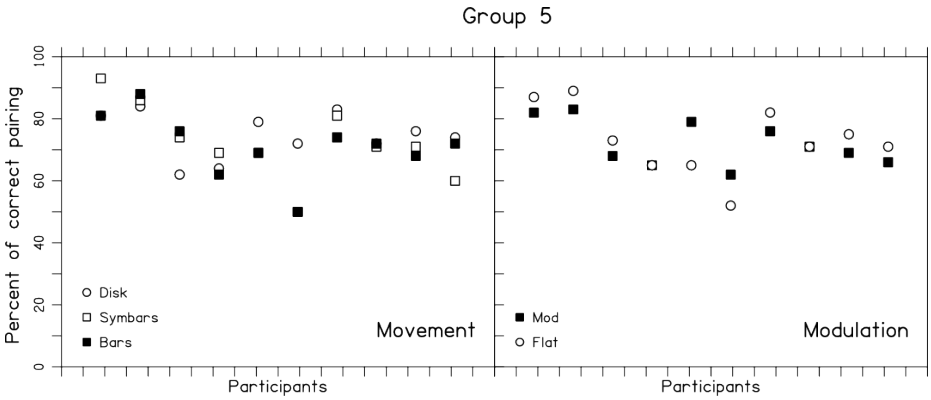


Fig. 10 Group 5. Performance of pairing for each participant averaged across repetitions for each feature. The participants were ranked based on their performance averages across all of the conditions and repetitions. The results for each visual target that were averaged across the modulation conditions are represented in the left panel. The results for each modulation that were averaged across the visual targets are represented in the right panel.

8.2 Discussion

Introducing a variation in the vertical extent of the visual display relative to the phonetic content of the corresponding auditory vowel enhanced pairing performance. We observed an increase in performance of approximately 20 percent compared to the other groups. The phonetic relationship between the auditory and visual parts of the stimuli may have enhanced the coherence between the streams, leading to a better binding of inputs. Pairing, which reflects binding, was therefore facilitated when the phonetic cues were introduced. In the present experiment, all of the subjects performed the pairing significantly better than chance, and we observed an absence of an effect of the low level cues; therefore, we conclude that the pairing task was mainly achieved at the phonetic level. In addition, we conclude that all of the other effects cannot be attributed to this mechanism. This result confirms that low level binding was engaged in the other experiments (groups 1 to 4)

9 Conclusion

1. The pairing paradigm. The pairing paradigm was successful because it allowed for contrastive results between conditions. For each set of conditions, the observed aptitude of pairing in synchrony was assumed to be a consequence of the binding process and a measure of the strength of the binding. The audio and video were bound together for at least one condition in each experiment, with a large amount of variability between the participants. All five experiments produced statistically significant results. However, the role played by the different factors had to be carefully interpreted across the different experiments because the same features (visual cues and auditory cues) may have had different effects.
2. The participant's variability. For each experiment, the scores were ranked. This ranking revealed that a small proportion of participants did not perform the task and that the other portion exhibited weak performances. This variability was large.
3. Sensitivity to the task. The only difference between experiment 1 and experiment 2 was that we asked the participants to pair the mouth closing or the mouth opening. The experimental design and the stimuli were exactly the same for these two experiments. In the first experiment, the movement was the determinant cue, whereas the delta-f0 in the second experiment enhanced the pairing. One possible explanation is the visual salience of the closing movement. The visual salience could be inherent to an internal visual representation of the movement, which was sharper for the closing movement. When the opening movement was the target, the participants used the only available cue to assist in the pairing decision, which was the delta-f0, but the effect was globally weaker. The streaming of vowel triplets allowed for a better contrast between the two possible percepts, and the appreciation of their coherence was better. This result suggests that the decision process was guided by some overall estimation about the cues that were currently available to perform the task.
4. Sensitivity to the envelope shape. In experiment 3, there was an interaction between the envelope shape and the movement/contrast factor.
5. Overall estimation process. In experiments 1 and 3, the non-specific movement was the most relevant visual feature, allowing pairing of the auditory speech with the non-speech visual cues In experiment 4, the contrast was the best cue. In this experiment, the audio envelope was manipulated to be better correlated with the video variation for two-thirds of the stimuli. For the other one-third, the same flat envelope used in experiments 1 and 2. We observed that contrast-dependent binding appeared for all of the stimuli. Therefore, the binding was not strictly related to the physical characteristics of the stimuli. We conclude that an overall estimation mechanism operated to select the contrast cue as the best one, and this theory explains the transfer to the flat envelope and the lack of pairing for the movement cue.
6. Phonetic cues. Last, the effect of a visual phonetic cue was relevant for the pairing task because it did not allow for the identification of each vowel separately. We introduced the dynamic of the opening movement parameter, which was captured from a recording of a man reading the vowel sounds. All of the participants exhibited a significant level of performance. The pairing performance was similar for all of the displays, and the performance was not sensitive to the audio envelope shape. The previous low level binding mechanism can be overruled by phonetic processing, which used the partial identification of each vowel. The robust pairing that was observed after introducing the mouth dynamic was different from the weak and unstable binding we observed in the other

experiments. This weak and unstable binding was based on only the temporal coherence between the low level features.

Acknowledgements This work was supported by Grants from the Région Rhones-Alpes Auvergne 'Cluster HVN 2007' and the Agence Nationale de Recherche (ANR-08-BLAN-0167-01). Special thanks to participants who took part in these experiments.

References

American National Standards Institute (2004) Ansi s3.21-2004: Methods for manual pure-tone threshold audiometry

Bernstein LE, Auer ETJ, Takayanagi S (2004) Auditory speech detection in noise enhanced by lipreading. Speech Communication 44:5–18

Bertelson P, Aschersleben G (2003) Temporal ventriloquism: crossmodal interaction on the time dimension. 1. evidence from auditory-visual temporal order judgment. Int J Psychophysiol 50(1-2):147–155

Berthommier F (2004) A phonetically neutral model of the low-level audio-visual interaction. Speech Communication 44(1-4):31–41

Bregman A (1990) Auditory Scene Analysis: The Perceptual Organization of Sounds. MIT Press

de Gelder B, Pourtois G, Weiskrantz L (2002) Fear recognition in the voice is modulated by unconsciously recognized facial expressions but not by unconsciously recognized affective pictures. Proc Natl Acad Sci U S A 99(6):4121–4126

Grant KW, v Wassenshove V, Poeppel D (2003) Discrimination of auditory-visual synchrony. In: AudioVisual Speech Perception

Green KP, Kuhl PK, Meltzoff AN, Stevens EB (1991) Integrating speech information across talkers, gender, and sensory modality: Female faces and males voices in the mcgurk effect. Percept Psychophys 50:524–536

Keetels M, Stekelenburg J, Vroomen J (2007) Auditory grouping occurs prior to intersensory pairing: evidence from temporal ventriloquism. Exp Brain Res 180(3):449–456

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

MacLeod A, Summerfield Q (1987) Quantifying the contribution of vision to speech perception in noise. Br J Audiol 21(2):131–141

Massaro DW, Cohen MM, Smeele PM (1996) Perception of asynchronous and conflicting visual and auditory speech. J Acoust Soc Am 100(3):1777–1786

van Noorden L (1975) Temporal coherence in the perception of tone sequences. Unpublished doctoral dissertation, Technische Hogeschool Eindhovern, Eindhoven, The Netherlands

Radeau M, Bertelson P (1977) Adaptation to auditory-visual discordance and ventriloquism in semirealistic situations. Percept Psychophys 22:137–146

Recanzone GH (2003) Auditory influences on visual temporal rate perception. J Neurophysiol 89(2):1078–1093

Schutz M, Kubovy M (2009) Causality and cross-modal integration. J Exp Psychol Hum Percept Perform 35(6):1791–1810

Schwartz, Berthommier, Savariaux (2003) Auditory syllabic identification enhanced by non-informative visible speech. In: Audio Visual Speech Perception

Sumby WH, Pollack I (1954) Visual contribution to speech intelligibility in noise. Journal of th Acoustical Society of America 26:212–215

Vaina LM (1994) Functional segregation of color and motion processing in the human visual cortex: clinical evidence. Cereb Cortex 4(5):555–572

Vatakis A, Spence C (2007) Crossmodal binding: evaluating the ”unity assumption” using audiovisual speech stimuli. Percept Psychophys 69(5):744–756

Welch RB (1972) The effect of experienced limb identity upon adaptation to simulated displacement of the visual field. Percept Psychophys 12:453–456

Chapitre 5

Discussion, Perspectives et Conclusions

Les travaux présentés dans ce manuscrit ont apporté des éléments de discussion et soulevés plusieurs perspectives concernant le rôle d'un indice visuel pour la ségrégation auditive. Nous allons dans un premier temps rappeler succintement les principaux résultats rapportés dans chacune des études. Ensuite, nous mettrons en perspective les différents travaux entre eux et avec la littérature. Nous proposerons dans cette discussion des perspectives à court et moyen terme à donner à ces travaux.

5.1 Discussion

5.1.1 Principaux résultats obtenus

Interactions audiovisuelles et ségrégation irrépressible

La première étude était consacrée aux interactions entre un stimulus visuel de parole (mouvement de lèvres) et les mécanismes de ségrégation irrépressible. Les tâches proposées aux participants ont permis de mettre en évidence un effet de la présentation de mouvement de lèvres articulant une partie du flux auditif sur la ségrégation irrépressible. Cette influence a été observée uniquement dans le cas où la cohérence entre le mouvement des lèvres

et les voyelles auditives prononcées était importante. La cohérence était à la fois phonétique et temporelle. Selon Vatakis et Spence (2006); Vatakis *et al.* (2007, 2008a), la cohérence phonétique est un facteur important pour établir la présomption d'unité. Dans notre étude, différents degrés de cohérence ont été manipulés et l'effet visuel sur la ségrégation irrépressible n'a été rapporté que lorsque nous avons utilisé des mouvements de lèvres réelles doublés avec des voyelles auditives enregistrées par le même locuteur. La présomption d'unité (ou l'estimation de la cohérence) a produit une fusion audiovisuelle probablement forte. Le second résultat, intéressant dans cette étude, a été que nous n'avons pas observé d'interaction entre le facteur visuel et le facteur acoustique (la différence de fréquence fondamentale entre les flux) sur la ségrégation irrépressible. Les deux facteurs ont contribué indépendamment à la ségrégation irrépressible. L'indice acoustique n'a pas renforcé l'effet de l'indice visuel. Ceci peut tenir au fait que l'indice visuel était fortement cohérent avec le flux auditif. Ainsi, l'indice acoustique n'ajoutait rien de plus au processus de ségrégation. Notons pour finir que cette première étude met en évidence pour la première fois une influence de l'indice visuel sur la ségrégation auditive irrépressible. De plus, la méthode expérimentale, consistant en la détection d'une déviation du rythme de présentation, n'impliquait pas l'identification des voyelles auditives. Ce type de protocole expérimental semble adapté pour tester les populations ayant des difficultés pour identifier les voyelles comme les personnes malentendantes par exemple.

Ségrégation basée sur les schémas et attention

La seconde étude nous a amené à étudier la contribution des mécanismes attentionnels dans la ségrégation basée sur les schémas. Ce thème de recherche reste animé par un vif débat dans la communauté scientifique. Nous avons proposé une expérience basée sur l'identification de mélodies intercalées ne partageant aucune différence acoustique exceptée une différence de rythme de présentation. Cette étude a permis, elle aussi pour une première fois, de mettre en évidence une contribution des mécanismes attentionnels dans les mécanismes de ségrégation auditive. En l'absence de différence acous-

tique, l'attention guidée par les schémas, à elle seule, permet de réaliser de la ségrégation. L'attention calée sur le rythme de présentation a permis de supprimer la mélodie distractrice, facilitant ainsi l'identification de la mélodie cible à la manière d'une ségrégation de type figure sur fond.

Interactions audiovisuelles et ségrégation basée sur les schémas

Enfin dans la troisième étude, nous avons proposé une tâche originale d'appariement entre des stimuli visuels élémentaires et un flux de voyelles auditives. Le but de cette étude a été d'évaluer le liage audiovisuel et sa spécificité pour la parole. Les séquences de voyelles auditives pouvaient être intégrées en un seul flux ou ségrégées en deux flux en fonction de la différence de fréquence fondamentale introduite. Cette tâche mettait en œuvre un mécanisme de ségrégation tardive. Le premier élément que nous apporte cette expérience tient au fait que dès que l'indice visuel contient un trait phonétique (ici, la dynamique d'ouverture des lèvres), les performances d'appariement sont meilleures que lorsque l'indice visuel ne contient pas de trait phonétique. Le gain en cohérence phonétique a probablement renforcé le liage audiovisuel permettant aux participants d'apparier plus facilement les deux flux. Le second élément est la forte variabilité rapportée entre les participants. Chaque expérience présentée dans cette étude a démontré que la plupart des participants avaient des difficultés à réaliser la tâche. Les indices visuels non phonétiques proposés ont pu conduire une proportion de participants à réaliser un appariement correct mais non une majorité.

Cependant, il semble qu'un indice visuel de mouvement pourrait permettre d'apparier les flux. Le liage audiovisuel serait alors possible avec ce type d'indice. En s'appuyant sur les observations rapportées dans les travaux de Vroomen et Stekelenburg (2010), le facteur déterminant semble être la prédictabilité. En recueillant les corrélats neuronaux, marqueurs de l'intégration audiovisuelle, Besle *et al.* (2004) démontrent que lorsque l'indice visuel permet d'anticiper l'instant ou les stimuli visuels et auditifs sont synchronisés, l'intégration est optimale. Ainsi, dans notre étude, l'indice de mouvement semble être plus prédictible que l'indice de contraste et donc

nous pouvons supposer que le liage audiovisuel est facilité dans ces conditions. Cette notion de prédictabilité peut aussi s'appliquer à plus long terme. Vroomen et Stekelenburg (2010) montrent également qu'en réunissant la même condition expérimentale dans un seul bloc, cela permettait d'observer cet effet de prédictabilité. Cet effet disparaît si les conditions sont mélangées dans chaque bloc. Dans notre expérience, les conditions étaient mélangées dans chaque bloc, réduisant potentiellement cet effet de prédictabilité.

5.1.2 Mise en perspective

Capture attentionnelle

L'étude concernant la ségrégation des mélodies intercalées (chapitre 3) a permis de montrer que l'attention captée par les attributs auditifs ou l'attention basée sur les schémas pouvaient renforcer la ségrégation même en l'absence de différence acoustique entre la mélodie à extraire et la mélodie distractrice. Cette capture attentionnelle, appuyée par la présentation d'un indice visuel synchronisé dans les études 1 et 3, a pu potentiellement contribuer aux effets rapportés. Cette hypothèse peut être soutenue par une étude réalisée par Marozeau *et al.* (2010) impliquant elle aussi des mélodies intercalées (figure 5.1). Dans cette étude les participants écoutaient des mélodies familières intercalées avec des mélodies distractrices. Les deux mélodies partageaient la même plage de fréquence au début de l'écoute. Au fur et à mesure de l'écoute en boucle, la plage fréquentielle couverte par la mélodie distractrice diminuait ou augmentait en fréquence moyenne. Les participants étaient invités à rapporter la difficulté qu'ils ont eu à identifier la mélodie familière au cours du temps. Dans certaines conditions expérimentales, une représentation visuelle illustrant la hauteur de chaque note était affichée. Les auteurs sont parvenus à montrer qu'avec l'indice visuel qui représente symboliquement le flux auditif à extraire, les participants ont des performances d'identification meilleures. Les auteurs ont également testé la différence entre des participants musiciens et non musiciens. Pour les sujets musiciens, les performances sont meilleures. Cette étude montre qu'un indice visuel synchronisé avec le flux auditif améliore la ségrégation basée sur les schémas même pour des

participants non musiciens. Pour cette population, nous pouvons suggérer qu'il s'agit de mécanismes attentionnels basés sur les traits perceptifs. Pour les sujets musiciens, la connaissance de la notation de la musique leur à permis d'engager des schémas stockés en mémoire expliquant les performances meilleures que pour les non musiciens. Cette étude de Marozeau *et al.* est le maillon intermédiaire entre notre étude sur les mélodies intercalées, l'attention et notre étude sur le liage audiovisuel.

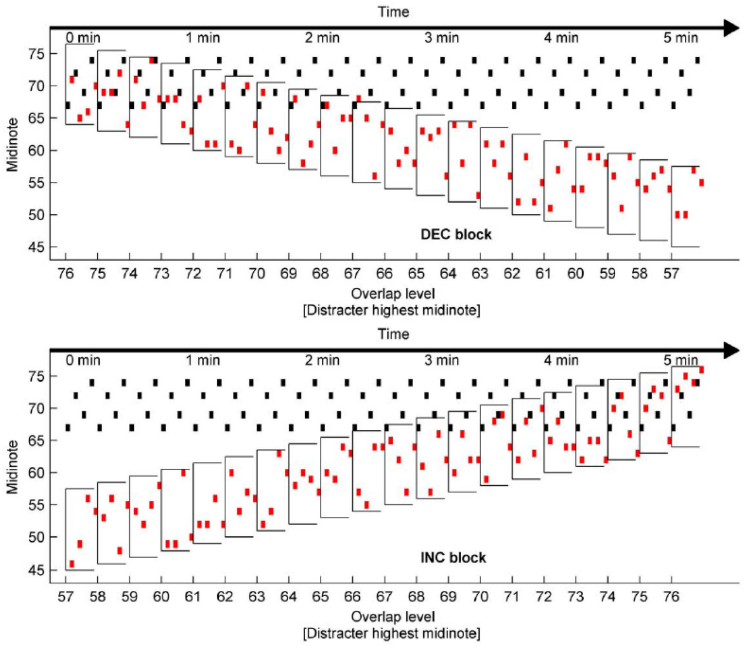


FIGURE 5.1 – Stimuli présentés dans l'étude de Marozeau *et al.* (2010)

Dans notre étude sur le liage audiovisuel (chapitre 4), la tâche consistait en une sélection de certaines voyelles auditives parmi d'autres. L'indice visuel, qui est selon la condition plus ou moins lié, permet de réaliser cet appariement (cette sélection). L'attention calée sur le rythme de présentation des voyelles auditives peut potentiellement avoir été renforcée par le rythme de présentation de l'indice visuel. Nous pourrions donc à la lumière de ces différentes études proposer une tâche de mélodies intercalées avec présentation d'un indice visuel élémentaire comme ceux de notre troisième étude. Dans l'expérience de Marozeau *et al.* (2010), l'attention basée sur les connaissances est engagée fortement pour les participants musiciens et peu,

voire pas du tout, pour les participants non musiciens. Dans notre étude sur les mélodies intercalées, les sons utilisés étaient des voyelles. Nous pourrions ajouter un film qui prononcerait les voyelles correspondantes aux notes de la mélodie cible (ou distractrice). Ceci permettrait de tester un effet de renforcement du mécanisme de sélection (ou de suppression) dû à la cohérence audiovisuelle phonétique. Pour tester un effet de renforcement purement attentionnel bas niveau cette fois-ci, il serait intéressant de présenter des films avec des signaux visuels élémentaires comme des flashs. Cet indice simple pourrait entraîner un cycle attentionnel. Dans cette perspective, nous pourrions présenter avant la séquence intercalée, une succession de flashs uniquement visuel sur le rythme de la mélodie à extraire pour observer l'effet de cette amorce sur la tâche d'identification de la mélodie intercalée.

Enfin, notre étude concernant l'influence visuelle sur les mécanismes de ségrégation auditive irrépressible (chapitre 2) amène à poser la question de l'attention pour expliquer les effets observés. Les trois conditions visuelles que nous avons proposé permettent d'écarter dans un premier temps un effet purement attentionnel. Même si l'attention peut avoir participé, elle ne peut rendre compte à elle seule de l'effet observé. En suivant le dessin expérimental de la seconde expérience (i.e. tâche de détection de déviation de rythme), il serait possible de présenter deux conditions visuelles engageant la même perturbation visuelle qui pourrait détourner de manière équivalente l'attention. Deux conditions visuelles avec le même contenu phonétique devront être proposées. Dans une condition cohérente, les voyelles visuelles prononceraient les voyelles auditives correspondantes. Dans la condition incohérente, les voyelles visuelles prononceraient une voyelle qui ne correspond jamais à la voyelle auditive et cela de manière aléatoire. La seule différence entre les deux conditions serait alors la cohérence phonétique entre les flux visuels et auditifs. Si dans ces conditions, nous observions un effet similaire de l'indice visuel cohérent sur la tâche de détection, nous pourrions alors avancer la même conclusion que celle que nous avons tiré de notre étude 1 avec probablement plus de certitude. Cependant, si le détournement de l'attention vers la cible visuelle est la conséquence d'une fusion audiovisuelle forte (phonétique), la conclusion resterait la même à savoir qu'un indice vi-

suel cohérent peut influencer la ségrégation auditive irrépressible. Quelque soit le dessin expérimental, il sera toujours extrêment complexe d'évacuer avec certitude la composante attentionnelle.

Ségrégation irrépressible et ségrégation tardive

Nos deux études impliquant des stimuli audiovisuels (chapitres 2 et 4) mettent en avant une contribution différente du facteur de fréquence fondamentale. Dans l'étude concernant la ségrégation auditive irrépressible, le facteur acoustique n'interagit pas avec l'indice visuel pour renforcer la ségrégation auditive. Dans l'expérience concernant le liage audiovisuel et la ségrégation auditive tardive, certains contextes permettent de mettre en avant la contribution de la différence de fréquence fondamentale. La nature des mécanismes de ségrégation auditive mis en œuvre pourrait expliquer pourquoi l'effet d'une différence de fréquence fondamentale n'est pas le même dans les deux études. Cependant, il faut garder à l'esprit que les deux tâches sont différentes. Nous avons par ailleurs démontré que les participants étaient très sensibles au contexte dans l'étude sur le liage audiovisuel. L'élément que nous pouvons examiner de plus près et qui nous amènera à proposer notre modèle est la notion de cohérence audiovisuelle.

Cohérence audiovisuelle

La cohérence audiovisuelle est probablement l'élément déterminant pour observer des interactions entre le flux visuel et les mécanismes de ségrégation auditive. Dans notre première étude (chapitre 2), nous avons proposé des stimuli de natures différentes. Dans la première expérience de cette étude, les voyelles auditives étaient générées avec l'algorithme de Klatt (Klatt, 1980). Les mouvements de lèvres visuels étaient générés à partir de successions d'images. Les trajectoires articulatoires étaient recomposées à partir d'un espace d'images en quantité limitée. Les stimuli auditifs et visuels étaient mixés dans un second temps. L'effet de l'indice visuel sur la ségrégation irrépressible était alors juste significatif. Dans la seconde expérience, les stimuli visuel et auditif étaient enregistrés en même temps, garantissant une cohérence temporelle et

phonétique plus importante. L'effet visuel sur la ségrégation auditive était ainsi plus important. La cohérence audiovisuelle apparaît dans cette première étude comme déterminante.

Dans notre étude concernant le liage audiovisuel (chapitre 4), nous retrouvons des éléments d'interprétation similaires. En effet, le liage audiovisuel entre des stimuli visuels non langagiers et un flux de voyelles auditives est sensible à des effets de variabilité entre participants ou de contexte. Dès l'instant où nous introduisons un paramètre phonétique, les performances sont meilleures pour l'ensemble des participants. La cohérence phonétique assure un liage audiovisuel plus fort que pour des indices non langagiers. Cette observation se positionne entre les travaux de Rahne *et al.* (2007) et la notion de présomption d'unité défendue par Vatakis *et al.* (2007). En effet, Rahne *et al.* ne sont pas parvenus pas à mettre en avant un effet d'un indice visuel élémentaire (comme le disque) lorsque un indice acoustique fort (comme la fréquence fondamentale) permettait d'organiser le flux sonore. La présomption d'unité, conséquence d'une évaluation de la cohérence audiovisuelle, conduit les participants à lier fortement ou non les stimuli visuels et auditifs.

Par ailleurs, la notion de prédictabilité présentée par Vroomen et Stekelenburg (2010) apparaît comme très intéressante. Dans notre étude (chaptitre 4), nous ne sommes pas parvenus à mettre en évidence un effet simple des traits visuels et auditifs. Nous avons en effet mélangé les différents traits dans chaque session expérimentale. En présentant les traits par bloc, il sera éventuellement possible de renforcer la prédictabilité des traits et ainsi d'en faciliter l'utilisation par les participants. Enfin, cette notion de prédictabilité semble prendre part de manière assez importante dans ce qui est notre évaluateur de cohérence. Plus un stimulus est prédictible, plus il sera facile de le traiter et de le lier à un stimulus cohérent dans une autre modalité.

5.2 Conclusions

5.2.1 Interactions audiovisuelles et analyse de scènes auditives

Pour conclure, nous proposons un modèle permettant d'illustrer la manière dont peuvent être liées les informations visuelles et auditives dans le cas de la parole (figure 5.2).

Un ensemble d'événements visuels et auditifs stimulent notre système perceptif. L'analyse de scènes auditives (ASA) et l'analyse de scènes visuelles (ASV) traitent les stimuli dans chaque modalité. La salience perceptive de certains traits, combinée à la cohérence potentielle de traits perceptifs entre eux, capturent notre focus attentionnel (attention basée sur les stimuli). Une première fusion perceptive peut se produire. Nous parlerons dans ce cas de figure de liage implicite. Les traits perceptifs peuvent être liés sans que nous faisions appel de manière explicite (d'où cette opposition) à nos représentations stockées en mémoire. Le terme *implicite* utilisé ici ne fait aucunement référence à la terminologie utilisée pour caractériser un état de conscience. Pendant ce temps d'exposition, notre système prend le temps d'intégrer les différents traits perceptifs et de créer des représentations dans des niveaux de traitement plus élevés. Le liage tardif se met alors en place. En fonction de la cohérence des représentations du signal, qui peut être considérée comme optimale pour les signaux de parole, notre système crée des objets audiovisuels cohérents. L'utilisation de ces objets et de cette cohérence tardive peut venir impacter la perception de la cohérence à un plus bas niveau. C'est pour cette raison probablement, que nous pouvons tolérer des dé-synchronisations plus importantes entre les stimuli visuels et auditifs de parole que pour d'autres signaux non langagiers. La perception de synchronie subjective (*point of subjective synchrony* : PSS), selon Vatakis *et al.* (2008b), pourrait donc être dépendante de cette cohérence audiovisuelle. De plus, les processus attentionnels mis en œuvre cette fois-ci reposeraient sur le concept d'objet (Santangelo et Spence, 2007). La cohérence audiovisuelle dans cette modélisation est donc le point central.

Retenons donc que les résultats, mis en évidence à travers nos trois études, s'inscrivent dans ce modèle. C'est la cohérence audiovisuelle qui va déterminer la force du liage audiovisuel et qui va permettre à un indice visuel d'influencer les mécanismes de la ségrégation auditive. Si la cohérence audiovisuelle est forte, comme elle a pu l'être dans notre étude 1 (chaptitre 2), le liage entre le flux visuel et le flux auditif est suffisamment fort pour moduler les mécanismes de ségrégation de bas niveau. En revanche, si la cohérence audiovisuelle est plus faible, comme dans la première expérience de notre étude 1 ou dans notre étude 3 (chapitre 4), le liage plus faible a pu moduler la ségrégation basée sur les schémas.

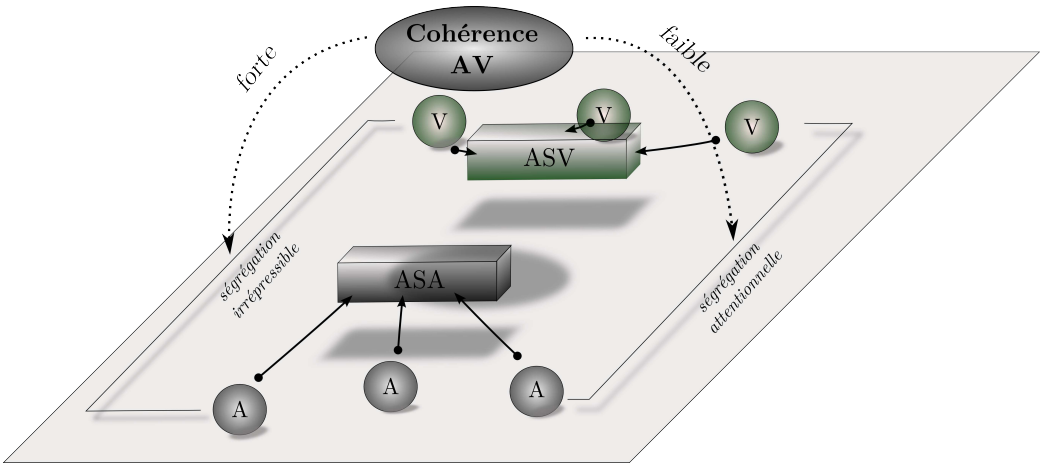


FIGURE 5.2 – Modèle d'intégration audiovisuelle de la parole

5.2.2 Dernière remarque

La perception de la parole repose sur une collaboration ascendante et descendante des aires corticales associatives et primaires et même des structures sous-corticales auditives. Les interactions précoces et tardives tissent un faisceau de connexions entre les différents niveaux de traitement. Notre

système perceptif est alors capable d'exploiter le plus petit indice permettant d'analyser la scène audiovisuelle et de mettre en place des processus d'anticipation et de facilitation afin d'être le plus efficace possible.

Bibliographie

Alain, C. et Arnott, S. R. (**2000**). "Selectively attending to auditory objects.", Front Biosci **5**, D202–D212.

Alain, C. et Bernstein, L. J. (**2008**). "From sounds to meaning : the role of attention during auditory scene analysis.", Curr Opin Otolaryngol Head Neck Surg **16**, 485–489.

Alais, D., Morrone, C., et Burr, D. (**2006**). "Separate attentional resources for vision and audition.", Proc Biol Sci **273**, 1339–1345.

Alpert, G. F., Hein, G., Tsai, N., Naumer, M. J., et Knight, R. T. (**2008**). "Temporal characteristics of audiovisual information processing.", J Neurosci **28**, 5344–5349.

Alsius, A., Navarra, J., Campbell, R., et Soto-Faraco, S. (**2005**). "Audiovisual integration of speech falters under high attention demands.", Curr Biol **15**, 839–843.

Anstis, S. et Saida, S. (**1985**). "Adaptation to auditory streaming of frequency-modulated tones", J Exp Psychol Hum Percept Perform **11**, 257–271.

Arnal, L. H., Morillon, B., Kell, C. A., et Giraud, A.-L. (**2009**). "Dual neural routing of visual facilitation in speech processing.", J Neurosci **29**, 13445–13453.

Assmann, P. (**1994**). "The role of formant transitions in the perception of concurrent vowels", J Acoust Soc Am **97**, 575–584.

Assmann, P. F. et Summerfield, Q. (**1989**). “Modeling the perception of concurrent vowels : Vowels with the same fundamental frequency”, J Acoust Soc Am **85**, 327–338.

Assmann, P. F. et Summerfield, Q. (**1990**). “Modeling the perception of concurrent vowels : vowels with different fundamental frequencies.”, J Acoust Soc Am **88**, 680–697.

Bernstein, L. E., Auer, E. T. J., et Takayanagi, S. (**2004**). “Auditory speech detection in noise enhanced by lipreading”, Speech Commun **44**, 5–18.

Bertelson, P. et Aschersleben, G. (**2003**). “Temporal ventriloquism : cross-modal interaction on the time dimension. 1. evidence from auditory-visual temporal order judgment.”, Int J Psychophysiol **50**, 147–155.

Besle, J., Fort, A., Delpuech, C., et Giard, M.-H. (**2004**). “Bimodal speech : early suppressive visual effects in human auditory cortex.”, Eur J Neurosci **20**, 2225–2234.

Best, V., Ozmeral, E. J., Kopco, N., et Shinn-Cunningham, B. G. (**2008**). “Object continuity enhances selective auditory attention.”, Proc Natl Acad Sci U S A **105**, 13174–13178.

Bey, C. (**1999**). “Reconnaissance de mélodies intercalées et formation des flux auditifs : Analyse fonctionnelle et exploration neuropsychologique”, Unpublished phdthesis.

Bey, C. et McAdams, S. (**2002**). “Schema-based processing in auditory scene analysis.”, Percept Psychophys **64**, 844–854.

Boltz, M. G. (**1993**). “The generation of temporal and melodic expectancies during musical listening.”, Percept Psychophys **53**, 585–600.

Botte, M. C., Drake, C., Brochard, R., et McAdams, S. (**1997**). “Perceptual attenuation of nonfocused auditory streams.”, Percept Psychophys **59**, 419–425.

Bregman, A. (**1990**). *Auditory Scene Analysis : The Perceptual Organzation of Sounds* (MIT Press).

Bregman, A. S. (**1978**). "Auditory streaming is cumulative.", J Exp Psychol Hum Percept Perform **4**, 380–387.

Bregman, A. S., Ahad, P. A., Crum, P. A., et O'Reilly, J. (**2000**). "Effects of time intervals and tone durations on auditory stream segregation.", Percept Psychophys **62**, 626–636.

Bregman, A. S. et Rudnicky, A. I. (**1975**). "Auditory segregation : stream or streams ?", J Exp Psychol Hum Percept Perform **1**, 263–267.

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

Carlyon, R. P., Cusack, R., Foxton, J. M., et Robertson, I. H. (**2001**). "Effects of attention and unilateral neglect on auditory stream segregation.", J Exp Psychol Hum Percept Perform **27**, 115–127.

Chait, M., de Cheveigné, A., Poeppel, D., et Simon, J. Z. (**2010**). "Neural dynamics of attending and ignoring in human auditory cortex.", Neuropsychologia .

Cherry, E. C. (**1953**). "Some experiments on the recognition of speech, with one and with two ears", J Acoust Soc Am **25**, 975–979.

Conrey, B. L. et Pisoni, D. B. (**2003**). "Audiovisual asynchrony detection for speech and nonspeech signals", *in Audio Visual Speech Perception*.

Cusack, R., Deeks, J., Aikman, G., et Carlyon, R. P. (**2004**). "Effects of location, frequency region, and time course of selective attention on auditory scene analysis.", J Exp Psychol Hum Percept Perform **30**, 643–656.

Darwin, C. J. (**1984**). "Perceiving vowels in the presence of another sound : constraints on formant perception.", J Acoust Soc Am **76**, 1636–1647.

de Cheveigné, A. (**1999**). "Vowel-specific effects in concurrent vowel identification.", J Acoust Soc Am **106**, 327–340.

de Cheveigné, A. (**2005**). *Pitch : Neural Coding and Perception*, chapitre Pitch perception models (Springer).

Dowling, W. J., Lung, K. M., et Herrbold, S. (**1987**). "Aiming attention in pitch and time in the perception of interleaved melodies.", Percept Psychophys **41**, 642–656.

Drake, C., Jones, M. R., et Baruch, C. (**2000**). "The development of rhythmic attending in auditory sequences : attunement, referent period, focal attending.", Cognition **77**, 251–288.

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

Efron, R. (**1970**). "The relationship between the duration of a stimulus and the duration of a perception.", Neuropsychologia **8**, 37–55.

Endress, A. D. (**2010**). "Learning melodies from non-adjacent tones.", Acta Psychol (Amst) .

Fairhall, S. L. et Macaluso, E. (**2009**). "Spatial attention can modulate audiovisual integration at multiple cortical and subcortical sites.", Eur J Neurosci **29**, 1247–1257.

Fritz, J. B., Elhilali, M., David, S. V., et Shamma, S. A. (**2007**). "Does attention play a role in dynamic receptive field adaptation to changing acoustic salience in A1 ?", Hear Res **229**, 186–203.

Giraud, A. L. et Truy, E. (**2002**). "The contribution of visual areas to speech comprehension : a pet study in cochlear implants patients and normal-hearing subjects.", Neuropsychologia **40**, 1562–1569.

Grant, K. W. (**2001**). "The effect of speechreading on masked detection thresholds for filtered speech.", J Acoust Soc Am **109**, 2272–2275.

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

Grimault, N. et Gaudrain, E. (**2006**). “The consequences of cochlear damages on auditory scene analysis”, Curr Top Acoust Res **4**, 17–24.

Hasson, U., Skipper, J. I., Nusbaum, H. C., et Small, S. L. (**2007**). “Abstract coding of audiovisual speech : beyond sensory representation.”, Neuron **56**, 1116–1126.

Jones, G. V. (**1984**). “Fragment and schema models for recall.”, Mem Cognit **12**, 250–263.

Jones, M., Moynihan, H., MacKenzie, N., et Puente, J. (**2002**). “Temporal aspects of stimulus-driven attentding in dynamic arrays”, Psychol Sci **13**, 313–319.

Jones, M. R., Kidd, G., et Wetzel, R. (**1981**). “Evidence for rhythmic attention.”, J Exp Psychol Hum Percept Perform **7**, 1059–1073.

Jäncke, L., Mirzazade, S., et Shah, N. J. (**1999**). “Attention modulates activity in the primary and the secondary auditory cortex : a functional magnetic resonance imaging study in human subjects.”, Neurosci Lett **266**, 125–128.

Kanai, R., Sheth, B. R., Verstraten, F. A. J., et Shimojo, S. (**2007**). “Dynamic perceptual changes in audiovisual simultaneity.”, PLoS ONE **2**, e1253.

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

Klatt, D. (**1980**). “Software for a cascade/parallel formant synthesizer”, J Acoust Soc Am **67**, 971–995.

Koehler, W. (**1967**). “Gestalt psychology.”, Psychol Forsch **31**, 18–30.

Koelewijn, T., Bronkhorst, A., et Theeuwes, J. (**2010**). "Attention and the multiple stages of multisensory integration : A review of audiovisual studies.", Acta Psychol (Amst) **134**, 372–384.

Lavie, N. (**1995**). "Perceptual load as a necessary condition for selective attention.", J Exp Psychol Hum Percept Perform **21**, 451–468.

Ludman, C. N., Summerfield, A. Q., Hall, D., Elliott, M., Foster, J., Hykin, J. L., Bowtell, R., et Morris, P. G. (**2000**). "Lip-reading ability and patterns of cortical activation studied using fmri.", Br J Audiol **34**, 225–230.

Macaluso, E., George, N., Dolan, R., Spence, C., et Driver, J. (**2004**). "Spatial and temporal factors during processing of audiovisual speech : a pet study.", Neuroimage **21**, 725–732.

Macken, W. J., Tremblay, S., Houghton, R. J., Nicholls, A. P., et Jones, D. M. (**2003**). "Does auditory streaming require attention ? evidence from attentional selectivity in short-term memory.", J Exp Psychol Hum Percept Perform **29**, 43–51.

Marozeau, J., Innes-Brown, H., Grayden, D. B., Burkitt, A. N., et Blamey, P. J. (**2010**). "The effect of visual cues on auditory stream segregation in musicians and non-musicians.", PLoS One **5**, e11297.

Massaro, D. W. et Cohen, M. M. (**1983**). "Evaluation and integration of visual and auditory information in speech perception.", J Exp Psychol Hum Percept Perform **9**, 753–771.

Massaro, D. W., Cohen, M. M., et Smeele, P. M. (**1996**). "Perception of asynchronous and conflicting visual and auditory speech.", J Acoust Soc Am **100**, 1777–1786.

Massaro, D. W., Kallman, H. J., et Kelly, J. L. (**1980**). "The role of tone height, melodic contour, and tone chroma in melody recognition.", J Exp Psychol Hum Learn **6**, 77–90.

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

Meddis, R. et Hewitt, M. J. (**1992**). “Modeling the identification of concurrent vowels with different fundamental frequencies.”, J Acoust Soc Am **91**, 233–245.

Mesgarani, N., David, S. V., Fritz, J. B., et Shamma, S. A. (**2008**). “Phoneme representation and classification in primary auditory cortex.”, J Acoust Soc Am **123**, 899–909.

Micheyl, C., Carlyon, R. P., Gutschalk, A., Melcher, J. R., Oxenham, A. J., Rauschecker, J. P., Tian, B., et Wilson, E. C. (**2007**). “The role of auditory cortex in the formation of auditory streams.”, Hear Res **229**, 116–131.

Miller, G. A. et Heise, G. A. (**1950**). “The thrill threshold”, J Acoust Soc Am **22**, 637–638.

Moettoenen, R., Krause, C. M., Tiipana, K., et Sams, M. (**2002**). “Processing of changes in visual speech in the human auditory cortex”, Cognitive Brain Research **13**, 417–425.

Mohammed, T., Campbell, R., Macsweeney, M., Barry, F., et Coleman, M. (**2006**). “Speechreading and its association with reading among deaf, hearing and dyslexic individuals.”, Clin Linguist Phon **20**, 621–630.

Moore, B. C. J. et Gockel, H. (**2002**). “Factors influencing sequential stream segregation”, Acta Acustica **88**, 320–333.

Musacchia, G., Sams, M., Nicol, T., et Kraus, N. (**2006**). “Seeing speech affects acoustic information processing in the human brainstem.”, Exp Brain Res **168**, 1–10.

Münte, T. F., Spring, D. K., Szycik, G. R., et Noesselt, T. (**2010**). “Electrophysiological attention effects in a virtual cocktail-party setting.”, Brain Res **1307**, 78–88.

Navarra, J., Alsius, A., Soto-Faraco, S., et Spence, C. (**2010**). “Assessing the role of attention in the audiovisual integration of speech”, Inf fusion **11**, 4–11.

Navarra, J., Vatakis, A., Zampini, M., Soto-Faraco, S., Humphreys, W., et Spence, C. (**2005**). "Exposure to asynchronous audiovisual speech extends the temporal window for audiovisual integration.", Brain Res Cogn Brain Res **25**, 499–507.

Pariyadath, V. et Eagleman, D. (**2007**). "The effect of predictability on subjective duration.", PLoS ONE **2**, e1264.

Pekkola, J., Ojanen, V., Autti, T., Jääskeläinen, I. P., Möttönen, R., Tarkiainen, A., et Sams, M. (**2005**). "Primary auditory cortex activation by visual speech : an fmri study at 3 t.", Neuroreport **16**, 125–128.

Pick, H., Warren, D., et Hay, J. C. (**1969**). "Sensory conflict in judgments of spatial direction", Percept Psychophys **6**, 203–205.

Poghosyan, V. et Ioannides, A. A. (**2008**). "Attention modulates earliest responses in the primary auditory and visual cortices.", Neuron **58**, 802–813.

Ponton, C. W., Bernstein, L. E., et Auer, E. T. (**2009**). "Mismatch negativity with visual-only and audiovisual speech.", Brain Topogr **21**, 207–215.

Pressnitzer, D., Sayles, M., Micheyl, C., et Winter, I. M. (**2008**). "Perceptual organization of sound begins in the auditory periphery.", Curr Biol **18**, 1124–1128.

Rahne, T., Böckmann, M., von Specht, H., et Sussman, E. S. (**2007**). "Visual cues can modulate integration and segregation of objects in auditory scene analysis.", Brain Res **1144**, 127–135.

Rahne, T. et Böckmann-Barthel, M. (**2009**). "Visual cues release the temporal coherence of auditory objects in auditory scene analysis.", Brain Res **1300**, 125–134.

Rahne, T., Deike, S., Selezneva, E., Brosch, M., König, R., Scheich, H., Böckmann, M., et Brechmann, A. (**2008**). "A multilevel and cross-modal approach towards neuronal mechanisms of auditory streaming.", Brain Res **1220**, 118–131.

Rouger, J., Fraysse, B., Deguine, O., et Barone, P. (**2008**). "Mcgurk effects in cochlear-implanted deaf subjects.", Brain Res **1188**, 87–99.

Rouger, J., Lagleyre, S., Fraysse, B., Deneve, S., Deguine, O., et Barone, P. (**2007**). "Evidence that cochlear-implanted deaf patients are better multisensory integrators.", Proc Natl Acad Sci U S A **104**, 7295–7300.

Sams, M., Aulanko, R., Hämäläinen, M., Hari, R., Lounasmaa, O. V., Lu, S. T., et Simola, J. (**1991**). "Seeing speech : visual information from lip movements modifies activity in the human auditory cortex.", Neurosci Lett **127**, 141–145.

Santangelo, V. et Spence, C. (**2007**). "Multisensory cues capture spatial attention regardless of perceptual load.", J Exp Psychol Hum Percept Perform **33**, 1311–1321.

Scheffers, M. (**1983**). "Sifting vowels : Auditory pitch anaysis and sound segregation", Thése de doctorat, University of Groningen, The Netherlands.

Schmuckler, M. A. et Boltz, M. G. (**1994**). "Harmonic and rhythmic influences on musical expectancy.", Percept Psychophys **56**, 313–325.

Scholl, B. J. (**2001**). "Objects and attention : the state of the art.", Cognition **80**, 1–46.

Schul, J. et Sheridan, R. A. (**2006**). "Auditory stream segregation in an insect.", Neuroscience **138**, 1–4.

Schwartz, Berthommier, et Savariaux (**2003**). "Auditory syllabic identification enhanced by non-informative visible speech", *in The International Conference on Audio-Visual Speech Processing*, 19–24 (St Jorioz).

Snyder, J. S., Alain, C., et Picton, T. W. (**2006**). "Effects of attention on neuroelectric correlates of auditory stream segregation.", J Cogn Neurosci **18**, 1–13.

Soto-Faraco, S. et Alsius, A. (**2009**). "Deconstructing the mcgurk-macdonald illusion.", J Exp Psychol Hum Percept Perform **35**, 580–587.

Spence, C., Nicholls, M. E., et Driver, J. (**2001**). "The cost of expecting events in the wrong sensory modality.", Percept Psychophys **63**, 330–336.

Stekelenburg, J. J. et Vroomen, J. (**2007**). "Neural correlates of multisensory integration of ecologically valid audiovisual events.", J Cogn Neurosci **19**, 1964–1973.

Suh, M.-W., Lee, H.-J., Kim, J. S., Chung, C. K., et Oh, S.-H. (**2009**). "Speech experience shapes the speechreading network and subsequent deafness facilitates it.", Brain **132**, 2761–2771.

Sumby, W. H. et Pollack, I. (**1954**). "Visual contribution to speech intelligibility in noise", J Acoust Soc Am **26**, 212–215.

Sussman, E. S., Horvath, J., Winkler, I., et Orr, M. (**2007**). "The role of attention in the formation of auditory streams.", Percept Psychophys **69**, 136–152.

Talsma, D., Doty, T. J., et Woldorff, M. G. (**2007**). "Selective attention and audiovisual integration : is attending to both modalities a prerequisite for early integration ?", Cereb Cortex **17**, 679–690.

Talsma, D., Senkowski, D., Soto-Faraco, S., et Woldorff, M. G. (**2010**). "The multifaceted interplay between attention and multisensory integration.", Trends Cogn Sci **14**, 400–410.

Talsma, D. et Woldorff, M. G. (**2005**). "Selective attention and multisensory integration : multiple phases of effects on the evoked brain activity.", J Cogn Neurosci **17**, 1098–1114.

Tan, M. N., Robertson, D., et Hammond, G. R. (**2008**). "Separate contributions of enhanced and suppressed sensitivity to the auditory attentional filter.", Hear Res **241**, 18–25.

Tiippana, K., Sams, M., et Andersen, T. (**2001**). "Visual attention influences audiovisual speech perception", *in Audiovisual Speech Perception Conference.*

Treisman, A. M. et Gelade, G. (**1980**). “A feature-integration theory of attention.”, Cognit Psychol **12**, 97–136.

Töllner, T., Gramann, K., Müller, H. J., et Eimer, M. (**2009**). “The anterior n1 component as an index of modality shifting.”, J Cogn Neurosci **21**, 1653–1669.

van Noorden, L. (**1975**). “Temporal coherence in the perception of tone sequences”, Unpublished doctoral dissertation, Technische Hogeschool Eindhovern, Eindhoven, The Netherlands.

van Wassenhove, V., Grant, K. W., et Poeppel, D. (**2005**). “Visual speech speeds up the neural processing of auditory speech.”, Proc Natl Acad Sci U S A **102**, 1181–1186.

van Wassenhove, V., Grant, K. W., et Poeppel, D. (**2007**). “Temporal window of integration in auditory-visual speech perception.”, Neuropsychologia **45**, 598–607.

Vatakis, A., Ghazanfar, A. A., et Spence, C. (**2008**a). “Facilitation of multisensory integration by the ”unity effect” reveals that speech is special.”, J Vis **8**, 14.1–1411.

Vatakis, A., Navarra, J., Soto-Faraco, S., et Spence, C. (**2007**). “Temporal recalibration during asynchronous audiovisual speech perception.”, Exp Brain Res **181**, 173–181.

Vatakis, A., Navarra, J., Soto-Faraco, S., et Spence, C. (**2008**b). “Audiovisual temporal adaptation of speech : temporal order versus simultaneity judgments.”, Exp Brain Res **185**, 521–529.

Vatakis, A. et Spence, C. (**2006**). “Audiovisual synchrony perception for music, speech, and object actions.”, Brain Res **1111**, 134–142.

Vatakis, A. et Spence, C. (**2007**a). “Crossmodal binding : evaluating the ”unity assumption” using audiovisual speech stimuli.”, Percept Psychophys **69**, 744–756.

Vatakis, A. et Spence, C. (**2007**b). “How ’special’ is the human face ? evidence from an audiovisual temporal order judgment task.”, Neuroreport **18**, 1807–1811.

Vatakis, A. et Spence, C. (**2008**). “Evaluating the influence of the ’unity assumption’ on the temporal perception of realistic audiovisual stimuli.”, Acta Psychol (Amst) **127**, 12–23.

Vroomen, J. et Stekelenburg, J. J. (**2010**). “Visual anticipatory information modulates multisensory interactions of artificial audiovisual stimuli.”, J Cogn Neurosci **22**, 1583–1596.

Watkins, S., Shams, L., Josephs, O., et Rees, G. (**2007**). “Activity in human v1 follows multisensory perception.”, Neuroimage **37**, 572–578.

Zwicker, U. (**1984**). “Auditory recognition of diotic and dichotic vowel pairs”, Speech Commun **3**, 265–277.

Annexe A

Appariement de la parole audio avec des indices visuels variés : liage ou non liage ?

Cette étude consiste en une partie de l'étude présentée dans le chapitre 4. *Cet article a été acceptée pour publication dans le proceeding de la conférence* AVSP 2009, Norwich, Royaume-Uni, *le 23 juin 2009.*

Pairing audio speech and various visual displays: binding or not binding ?

Aymeric Devergie[1]*, Frédéric Berthommier*[2] *and Nicolas Grimault*[1]

[1]Laboratoire Neuroscience Sensorielle, Comportement et Cognition, CNRS UMR 5020, Université Lyon 1, Lyon, France
[2]Gipsa-Lab,CNRS UMR 5216, INPG, UJF, Université Stendhal, Grenoble, France
{adevergi,ngrimault}@olfac.univ-lyon1.fr, berthommier@gipsa-lab.inpg.fr

Abstract

Recent findings demonstrate that audiovisual fusion during speech perception may involve pre-phonetic processing. The aim of the current experiment is to investigate this hypothesis using a pairing task between auditory sequences of vowels and non speech visual cues. The audio sequences are composed of 6 auditory French vowels alternating in pitch (or not) in order to build 2 interleaved streams of 3 vowels each. Various elementary visual displays are mounted in synchrony with one vowel stream out of the two. Our hypothesis is that, in a forced choice pairing task, the AV synchronized vowels will be found more frequently if such a perceptual binding operates. We show that the most efficient visual feature increasing pairing performance is the movement.

Surprisingly, some features we manipulated do not provide the increase in pairing performances. The visual cue of contrast variation is not correctly paired with the synchronized auditory vowels. Moreover, the auditory segregation, based on the pitch difference between the vowels streams, has no additional effect on pairing. In addition, the modulation of the auditory envelop, synchronized with the variation of the visual cue, has also no effect. Finally, when we introduce a phonetic cue in the visual display, pairing increases in comparison with non specific visual cues. The relative contribution of perceptual binding and late phonetic fusion is discussed.

Index Terms: Audiovisual fusion, perceptual binding, multimodal phonetic processing

1 Introduction

In speech, fusion of audio and visual inputs has been widely investigated through intelligibility tasks. It has been assumed that late phonetic fusion occurs during the perception of speech [1]. Only recently, other hypothesis arose, assuming that audio and visual inputs could interact at a pre-phonetic level. Intelligibility tasks are not appropriate to test this hypothesis because it necessarily involves 'lip reading' of the stimuli.

In order to focus on the lower level of processing involved in audiovisual fusion, some studies proposed a detection paradigm. When presenting a visual cue related to the auditory speech, it enhances detection of speech in noise [2] [3]. This observation argues in favour of a fusion at a more pre-phonetic level.

Findings about the ventriloquism effect showed that we are able to pair A and V inputs, even if they are not spatially coherent. This suggests that an underlying binding mechanism based on the temporal coherence is involved. The role of the temporal coherence in speech has been investigated with asynchrony detection tasks. In speech, despite the introduction of an offset asynchrony (e.g. with audio lag) between audio and visual inputs, a multimodal event could be perceived as coherent. This corresponds to a quite large temporal window of integration of about 250ms as described in [4].

Some electrophysiological studies proposed sequences of non speech auditory events grouped in several configurations [5]. Adding an elementary visual cue has been demonstrated to affect the perception of the auditory sequence and facilitated access to particular events in the auditory stream. The facilitation would suggest that audio and visual may be bound at a more prephonetic level.

From that point, the question of audiovisual binding in speech should be addressed. Our current study aimed to focus on prephonetic level of audiovisual fusion. As in [6], the experimental visual material we built consisted in elementary display varying in contrast or in movement. Such basic visual features would prevent phonetic processing to occur. The auditory material consisted in sequences of six French vowels alternating in pitch. The contrast or movement feature of visual display varied in synchrony with the auditory vowels. We defined a 'open-state' and a 'close state' for the two visual displays. The open-state for the contrast cue was the white disk and the close-state was the black disk. Open-state for the movement display corresponded to the large visual display and the close-state to the small display. In every sequences, one group of three vowels was synchronous with the 'open state' and the other group of three vowels was synchronous with the 'close state' (see figure 1 for a representation of the different states). The task consisted in pairing the group of three vowels which was synchronous with a particular state of the visual display. It is important to notice that in our experimental design, the phonetic identification of the auditory vowels would not have helped participants to perform the pairing task because the auditory vowels were not phonetically correlated to the visual displays.

Since it has been demonstrated that speech was tolerant to asynchrony [4], this probably suggests that the underlying binding process is relatively robust. Thus, in order to weaken the strength of binding, we introduced some temporal ambiguity in the experimental material. The vowel's duration was shorter than an open-close cycle of the visual cue. One open-close cycle overlapped parts of two successive vowels. Thus, pairing of A and V inputs was difficult even if variations of the visual cue were synchronous with the center of auditory vowels. We hypothesized that only the strength of binding would help participants to pair the correct audio vowels with the visual display.

Finally, in the current study, we have also investigated the role

of phonetic cues in pairing by introducing some phonetic features in the visual displays. Since, pairing of audio and visual probably involved either perceptual and phonetic aspects, it became interesting to design a kind of continuum between pure non speech and speech visual cues.

2 Movement and contrast features

2.1 Material and Methods

Sequences of six French vowels without overlap and silence were built. Each sequence was repeated in loop. Fundamental frequency of the vowels had two possible values: 100 or 134Hz. Two auditory patterns were proposed: one with constant f0 set to 100Hz and one with an alternating f0 (between 100 and 134Hz). In the latter condition, the f0 difference lead to the clear perception of two separate streams. Figure 1 represents the different visual cues. Three shapes and two visual features were proposed. In the contrast condition, shapes varied from black to white contrast on a black background frame synchronously with 1 out of 2 vowels. In the movement condition, shapes varied from open to close position with the same temporal pattern. Vowels were generated with [Klatt algorithm (1980)]. Stimuli were played using a SIGMATEL internal sound card and Sennheiser HD 250 Linear II headphones. Output level was set to 70dB SPL with RMS-value adjustment. Video display was achieved using a Samsung SyncMaster 540N TFT 17" display with a video frame rate set at 60Hz. Figure 7 shows a schematic view of one sequence represented on the timeline for the two kinds of visual displays.

2.2 Procedure

Twenty participants aged between 18 and 30 years took part in the experiment. They had to choose the triplet of vowels synchronized with a particular state of the visual display: in the contrast condition, they had to identify the group of three vowels synchronized with the open state (white disk) and in the movement condition the group of three vowels synchronized with the close state. This was made in accordance with pilot observations revealing a better detection of the close state for the movement condition. The temporal evolution of these two visual conditions were represented on figure 7.

The test was divided into 3 sessions. Participants were seated comfortably in a double-walled sound booth. The first session was a presentation session. All combinations of visual shapes and visual conditions were presented randomly. Then, the adaptation session began. In this session, they performed the pairing task with a set of 36 different runs. Each combination of shape, visual condition and f0 difference was repeated four times. Each sequence lasted 10 seconds. At the end of each sequence, the two triplets of vowels are displayed in the lower part of the screen. Participants have to choose the triplet synchronized with the target visual display. After this adaptation session, the test session started. It consisted in 240 runs divided into 4 blocks. In each block, all combinations were repeated five times each. The whole experiment lasted 30 minutes.

2.3 Results

On Figure 2, correct pairing of the target vowel triplet synchronized with the target visual state were averaged for each visual condition and f0 difference for all participants. A repeated-measure ANOVA with factors f0 difference and visual condition grouped by visual shape was performed. Visual condition has a significant effect on performances [$F(1,18)=10.86$; $p<0.01$]. F0 difference has no effect [F(1.18)=0.001; p=0.97]. No interaction between these two factors has been found [$F(1.18)=0.057$; $p=0.81$]. Performances for each combination (Visual type and Frequency) were compared to chance level. T-tests were performed and revealed only significant difference to chance level (50% correct) for the movement cue [Movement / Same F0: $t(19)=3.26$; $p<0.01$, Movement / different F0: $t(19)=2.67$; $p=0.015$]. Correct responses were grouped by visual display type in figure 3 for movement cues and in figure 4 for contrast cues. Responses were significantly different from chance level only for the vertical and horizontal bars in the movement condition.

The first experiment revealed that the most salient visual cue allowing perception of the temporal synchrony between auditory speech and non speech visual cue was the movement (Figure 2). Since the performance was at chance level with the Contrast visual display, this cue did not support to perceive the synchrony detection between the inputs.

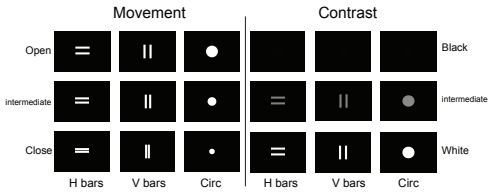


Figure 1: The different visual cues are represented in their key states (open, close and intermediate). The movement display varied from open to close. The contrast display varied from white to black. They also varied along orientation.

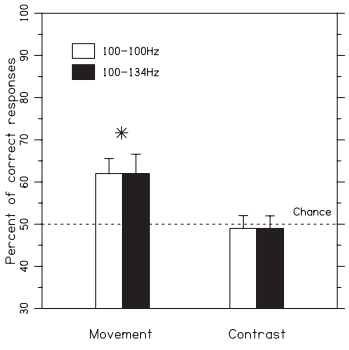


Figure 2: Percent of correct identification of the vowel triplet synchronized with white disk in the Contrast condition and close state in the Movement condition depending on Visual cue type and f0 difference between auditory vowels. Only the movement cue was significantly different from chance. Chance level was equal to 50%

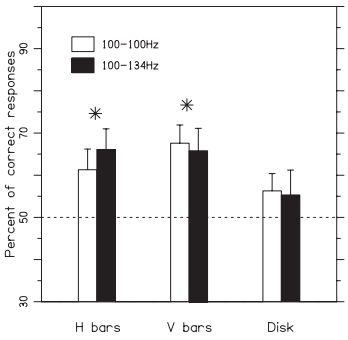


Figure 3: Percent of correct identification of the vowels triplet synchronized with the visual shape for each f0 difference grouped by visual display for the movement condition. Performance for vertical bars and horizontal bars were significantly better from chance.

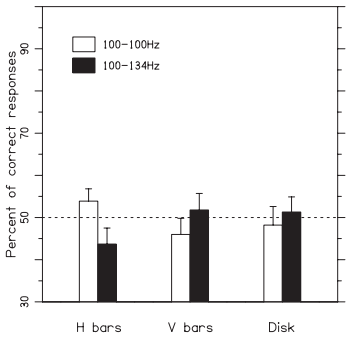


Figure 4: Percent of correct identification of the vowels triplet synchronized with the visual shape for each f0 difference grouped by visual display for the contrast condition. No visual cue elicited identification better than chance level.

3 Auditory envelop modulation

In the second experiment, we hypothesized that perception of AV synchrony would be facilitated if the modulation of the visual parameter is matched with a coherent envelop modulation of the auditory signal. The purpose of the Experiment 2 was to enhance detection of AV synchrony, particularly for the contrast condition.

3.1 Material and Methods

Ten participants took part in this experiment. Stimuli were similar to those used in Experiment 1. The number of visual shapes was reduced to two: the horizontal bars varying in movement and the disk varying in contrast. The two f0 conditions were maintained. We introduced modulation of the auditory envelop. In the 'No modulation' condition, the auditory envelop remained flat. In the 'Modulation condition', the level of each vowel was modulated with a triangular window in synchrony with the variation of the visual parameter. Global RMS-levels of the two modulation conditions were equalized. The experimental design was the same as in Experiment 1.

3.2 Results

A repeated-measure ANOVA with factors, visual condition, frequency difference and envelop modulation was performed. Figure 5 showed percent of correct pairing for all participants averaged for each visual condition, and envelop modulation for the f0 condition '100-100Hz' in the left panel and for f0 condition '100-134Hz' in the right panel. Visual condition had a significant effect on performances [$F(1,9)=12.61$; $p<0.01$]. F0 difference alone had no effect on performances [$F(1,9)=1.94$; $p=0.19$]. Theses results are consistent with the experiment 1. Results showed that envelop modulation had no effect on performances [$F(1,9)=0.11$; $p=0.74$]. Concerning the interactions between the three factors, no interaction was found between f0 difference and visual condition [$F(1,9)=0.15$; $p=0.70$] nor between f0 difference and envelop modulation [$F(1,9)=1.32$; $p=0.28$]. Right panel of the Figure 5 suggested that an interaction between visual condition and envelop modulation occured [$F(1,9)=10.31$; $p=0.011$]. T-tests were performed for each visual condition and did not revealed any significant difference between envelop modulation condition (Movement display: $t(18)=0.54$; $p=0.59$, Contrast display: $t(18)$; $p=0.10$). In addition, the performances in the 'modulation' condition for the contrast display was not significantly different from chance level [$t(9)=1.95$; $p=0.08$]

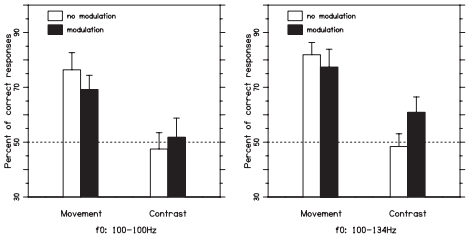


Figure 5: Percent of correct identification of the target triplet for each visual condition (x-axis) grouped by envelop modulation (white bar: no modulation, black bar: modulation). The left panel shows the performances for the f0 condition '100-100Hz'. The right panel shows the performances for the f0 condition '100-134Hz'.

4 Phonetic processing

Altogether, Experiments 1 and 2 demonstrated that the only salient cue allowing pairing between A and V cue is the movement. The experiment 3 included the following features. The auditory envelop modulation was maintained. The movement cue was further investigated by introducing the natural vertical extend of the lips. Using natural lips movement for building the visual display was also expected to introduce some phonetic features.

4.1 Material and Methods

Ten participants took part in this experiment. Design was the same as in Experiment 1 and 2. New visual conditions were introduced. Three visual cues were proposed. The first one consisted in the disk varying in size (as in Experiment 1). Dynamics of the visual parameter defining the size of the shape was extracted from video records of natural lip movements. The vertical extend of natural lips, as referred as 'A parameter' in the literature, controlled the variation of the radius of the disk varying in size. Two visual cues were derived from the horizontal bars cue present in Experiment 1 and 2. A first one, called '1Dsym' (one dimension - symmetric) had its vertical extend defined by the 'A parameter' and had its global movement centered on the vertical axis. A second one, called '1D' (one dimension) had its vertical extend defined by the 'A parameter' and the upper bars had the same vertical coordinate than the video-recorded upper lip. A single f0 difference was used in this design because it had been demonstrated (in Experiment 2) that f0 difference has no effect. Two conditions of modulation were used here: 'modulation' and 'no-modulation'. The experimental procedure was the same as in Experiment 1 and Experiment 2.

4.2 Results

A repeated-measure ANOVA with factors visual condition and envelop modulation was performed. Figure 6 shows percent of correct pairing for all participants averaged for each visual condition, and envelop modulation. Visual condition has no significant effect on performances [$F(2,12)$=1.91; p=0.18]. On overall, envelop modulation has no effect on performances [$F(1,6)$=0.0057; p=0.94]. No interaction is found between envelop modulation and visual condition [$F(2,12)$=0.0124; p=0.98]. Comparisons to similar conditions presented in Experiment 1 and in Experiment 3 reveal differences on averaged performances. Percent of correct identification significantly increases. The only difference introduced between Exp 1 and Exp 3 is the adding of phonetic cues. This significant difference could only be attributed to this adding.

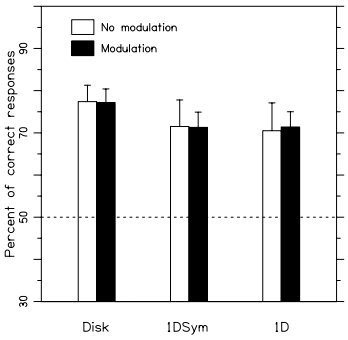


Figure 6: Percent of correct identification across the different visual cues with the two modulation conditions. All condition were significantly different from chance. Compared with similar condition in Experiment 1 (disk varying in size), identification performances increases significantly.

5 Discussion

The results found in the three experiment showed that the ability of pairing could be a consequence of an underlying binding process allowing detection of audiovisual synchrony. First surprisingly, the contrast visual display did not enabled participants to pair correctly the audio and visual stimuli even if physical synchrony was ensured. Second, the movement feature was the most relevant visual feature allowing pairing of auditory speech with non speech visual cue. In the experiment 1, we asked participants to detect synchrony between audio and the close state of the visual display. This was in contradiction with the natural lip movement, which would have been related to the open state of our visual displays. Moreover, the vertical range of the bars remained the same for all the vowels. As a consequence, no account for a speech specific processing could explain the better performance of pairing. When we provided more audiovisual coherence, thanks to the auditory envelop modulation, the detection of synchrony was not enhanced. Moreover the auditory stream organization induced by the difference of fundamental frequency between the two vowel streams did not impact the pairing. In sum, the features influencing the organization of the auditory input did not affect the pairing processing and thus the underlying binding. Finally, the effect of the phonetic discrimination across the different vowels was relevant. The same visual feature (disk varying in size), which contains or not this phonetic cue, provided better pairing performances when the phonetic parameter was present. The underlying perceptual binding could have been overruled by a phonetic processing which could have enhanced pairing.

6 Acknowledgments

This work was supported by Grants from the Region Rhones-Alpes Auvergne 'Cluster HVN 2007' and the Agence Nationale de Recherche (ANR-08-BLAN-0167-01). Special thanks to running participants.

References

[1] D. W. Massaro and M. M. Cohen, "Evaluation and integration of visual and auditory information in speech perception." *J Exp Psychol Hum Percept Perform*, vol. 9, no. 5, pp. 753–771, Oct 1983.

[2] J. Kim and C. Davis, "Hearing foreign voices: does knowing what is said affect visual-masked-speech detection?" *Perception*, vol. 32, no. 1, pp. 111–120, 2003.

[3] L. E. Bernstein, E. T. A. Jr., and S. Takayanagi, "Auditory speech detection in noise enhanced by lipreading," *Speech Communication*, vol. 44, pp. 5–18, 2004.

[4] D. W. Massaro, M. M. Cohen, and P. M. Smeele, "Perception of asynchronous and conflicting visual and auditory speech." *J Acoust Soc Am*, vol. 100, no. 3, pp. 1777–1786, Sep 1996.

[5] T. Rahne, M. Beckmann, H. von Specht, and E. S. Sussman, "Visual cues can modulate integration and segregation of objects in auditory scene analysis." *Brain Res*, vol. 1144, pp. 127–135, May 2007.

[6] Schwartz, Berthommier, and Savariaux, "Auditory syllabic identification enhanced by non-informative visible speech," in *Audio Visual Speech Perception*, 2003.

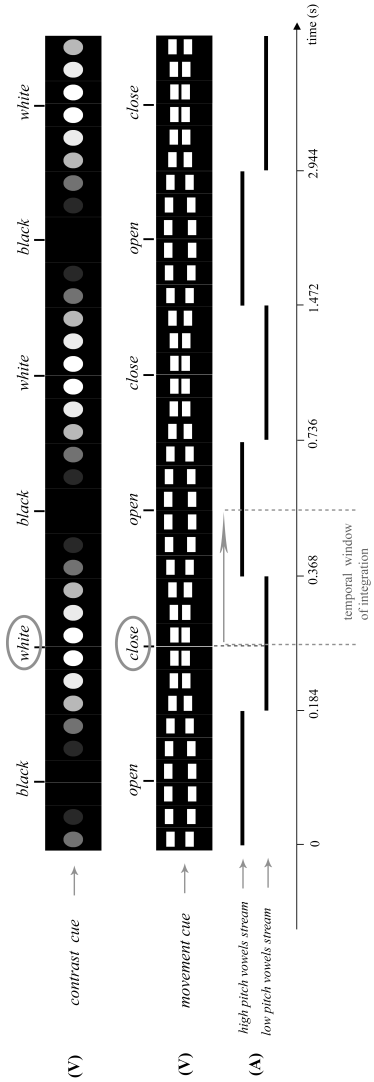


Figure 7: Representation of the stimuli on the temporal axis. The 2 different types of visual cue are represented. Frames shows the variation of the visual cues over time. Each initial and ending state of the cue is tagged. Duration of the temporal window of integration was defined equal to 250ms. Since the duration of the integration window was longer than duration of the vowels, uncertainty in binding could have appeared. For example, the second close state in the movement cue could either have been paired with the vowel starting at 0.184ms or with the vowel starting at 0.368ms because each vowel dropped in the span of the temporal window of integration. This hypothesis could account for the perceptual ambiguity.