Émergence des représentations perceptives de la parole : Des transformations verbales sensorielles à des éléments de modélisation computationnelle

Anahita Basirat

UNIVERSITE DE GRENOBLE
INSTITUT POLYTECHNIQUE DE GRENOBLE

N° attribué par la bibliothèque
|__|__|__|__|__|__|__|__|__|__|

T H E S E

pour obtenir le grade de

DOCTEUR DE L'Université de Grenoble
délivré par l'Institut polytechnique de Grenoble

Spécialité : Signal, Image, Parole, Télécoms

préparée au laboratoire **Gipsa-lab, département Parole et Cognition**

dans le cadre de **l'Ecole Doctorale**

Electronique, Electrotechnique, Automatique et Traitement du Signal

présentée et soutenue publiquement

par

Anahita BASIRAT

le 09 Septembre 2010

Émergence des représentations perceptives de la parole : Des transformations verbales sensorielles à des éléments de modélisation computationnelle

Directeur de thèse : M. Jean-Luc SCHWARTZ

JURY

M. Christian JUTTEN	Président
M. Noël NGUYEN	Rapporteur
M. Jean VROOMEN	Rapporteur
M. Jean-Luc SCHWARTZ	Directeur de thèse
M. Daniel PRESSNITZER	Examinateur
M. Marc SATO	Examinateur

À mes parents,
à Amaël,
à Arman

به پدر و مادرم،
به اَمَئِل،
به آرمان

Remerciements

Je tiens tout d'abord à remercier Jean-Luc Schwartz, mon directeur de thèse, pour sa disponibilité, ses conseils, son écoute et son soutien permanent. Il sait laisser les rêves s'envoler tout en les orientant. Jean-Luc, j'ai beaucoup appris de toi durant ces années, tant sur le plan scientifique qu'humain. Pour tout cela, je te remercie très sincèrement.

Mes remerciements vont également à Marc Sato avec qui j'ai eu la chance et le grand plaisir de collaborer. Bien qu'il ne soit pas officiellement associé à cette thèse, il a contribué de très près à ces travaux. Merci Marc pour ta générosité, ta disponibilité, tes encouragements et ton soutien.

Je tiens à remercier les membres du jury pour le temps qu'ils ont consacré à la lecture de ce manuscrit. Je suis très reconnaissante à Noël Nguyen et Jean Vroomen d'avoir accepté d'évaluer et de rapporter ce travail. J'adresse mes plus vifs remerciements à Daniel Pressnitzer pour les discussions et pour sa participation au jury en tant qu'examinateur. Je remercie Marc Sato qui a bien voulu examiner cette thèse. Je tiens à remercier également Christian Jutten d'avoir accepté de présider ce jury, j'en suis très honorée.

J'ai eu la formidable opportunité de collaborer avec Philippe Kahane et Jean-Philippe Lachaux pendant ma thèse, je leur adresse mes sincères remerciements.

Je voudrais également remercier tous les membres du département Parole et Cognition du Gipsa-lab (et de l'ICP). Merci aux jeunes chercheurs (et aux moins jeunes) du labo pour les bons moments au coin café et autres. Un grand merci également au personnel administratif et technique dont l'aide et l'implication ont permis le bon déroulement de mes travaux.

Mes remerciements vont aussi à celles et ceux qui ont participé, avec beaucoup de patience, aux expériences réalisées pendant cette thèse.

Je remercie Bahram, ma belle famille et mes amis pour leur soutien et leurs encouragements.

Les mots ne sont pas assez forts pour exprimer ma reconnaissance envers ma famille : je remercie mes parents, Amaël et Arman pour leur confiance, leur présence et leur patience. Merci Amaël de m'avoir encouragée dans les moments de doute et d'avoir relu et corrigé ce manuscrit.

بابا، مامان، آرمان ازتون ممنونم ...

Table des matières

Table des figures

Liste des tableaux

Problématique

Cette thèse s'inscrit dans une question générale de l'étude de la perception humaine, celle du liage perceptif (*binding*). Elle focalise cette question sur le domaine riche et complexe de l'étude de la parole, ce qui conduit à centrer le questionnement sur « l'objet parole » et, par là, à tenter de mieux cerner la nature de cet objet et sa construction cognitive. Le paradigme expérimental que nous avons choisi pour cette étude est celui de la multistabilité perceptive, paradigme classique d'étude des objets visuels, et dont les travaux récents de Marc Sato à l'ICP (actuellement département de Parole et Cognition au Gipsa-lab) ont montré qu'il peut, à travers le phénomène des transformations verbales, fournir un outil puissant d'étude de la perception de la parole (Sato, 2004). Liage perceptif, objet parole et multistabilité perceptive sont donc les trois composantes qui fondent les travaux expérimentaux et computationnels de cette thèse. Nous allons les situer plus précisément dans cette introduction posant la problématique, avant de fournir notre plan de travail.

Liage perceptif, base de la formation de l'objet

Lorsque nous regardons un objet, notre système visuel traite séparément (au moins en partie) différentes caractéristiques (*features*) de l'objet, telles que sa couleur, sa forme et sa position. Cependant, notre perception de cet objet ne correspond pas à ses caractéristiques d'une manière séparée mais à l'objet en question dans sa totalité. Ce phénomène a été nommé par Kant « l'unité transcendantale de l'aperception » et étudié maintenant sous le nom de problème de liage. Différentes définitions du problème de liage ont été proposées dans la littérature. Par exemple, selon Zeki (1992), le problème du liage concerne les mécanismes neuronaux qui permettent la dissociation des différentes caractéristiques de deux objets différents et l'intégration de celles du même objet :

> There is, next, what is commonly referred to as the binding problem, a critical problem for visual physiology. The problem is that of determining that it is the same (or a different) stimulus which is activating different cells in a given visual area or in different visual areas. (Zeki, 1992, p. 321)

Une autre définition a été proposée par Damasio (1989), dans laquelle il fait référence à la perception et l'expérience de réalité sous le nom de problème du liage :

> Current knowledge from neuroanatomy and neurophysiology of the primate nervous system indicates unequivocally that any entity or event that we normally perceive through multiple sensory modalities must engage geographically separate sensory modality structures of the central nervous system. ... The experience of reality, however, both in ongoing perception as well as in recall, is not parcellated at all. The normal experience we have of entities and events is coherent and "in-register", both

> spatially and temporally. Features are bound in entities, and entities are bound in events. How the brain achieves such a remarkable integration starting with the fragments that it has to work with is a critical question. I call it the binding problem. (Damasio, 1989, p. 29)

Ces deux définitions représentent deux aspects différents du problème du liage, liage phénoménologique et liage fonctionnel, comme proposé par Smythies :

> There are two quite different binding problems, which we can call BP1 and BP2. BP1 asks "How is the representation of information built up in the neural networks that there is one single object out there and not a mere collection of separate shapes, colours and movements ?"... This presupposes an underlying mechanism that locates the right colour in the right shape and keeps both moving together... There do not seem to be any difculties about this question (Smythies, 1994b, p. 54). BP2 asks "How do the brain mechanisms actually construct the phenomenal object ?", which is another matter altogether. (Smythies, 1994a, p. 321)

Spécificité de l'objet parole

La question du liage s'inscrit en relation constante avec une autre question centrale de l'étude de la perception, qui est celle de la nature de l'objet de la perception. Lier les différents aspects de la représentation d'un objet suppose que l'on sait ce qu'est un objet pour la perception. L'objet perceptif définit, et est défini par, la cohérence de ses représentations perceptives. Cette cohérence découle de sa nature propre, ce qui renvoie à des débats majeurs comme ceux de l'idéalisme contre le réalisme : cohérence construite par le cerveau à partir des propriétés sensorielles (idéalisme), ou imposée par le monde et ses propriétés physiques à la perception sous forme « d'affordances » (réalisme).

Or, ce n'est pas d'un objet quelconque que nous traiterons dans cette thèse, mais de « l'objet parole » : dans une scène perceptive, auditive, visuelle ou multisensorielle, contenant des signaux émis par le conduit vocal produisant de la parole, qu'est-ce qu'un flux de parole et de quels objets perceptifs est-il constitué ?

Cette question est elle-même une question majeure de la perception de la parole, autour du débat entre théories auditives et motrices (reprenant, sous un format propre, le débat entre idéalisme et réalisme). Une originalité de cette thèse est qu'elle aborde cette question, classique, de la nature de l'objet parole, sous l'angle, original, du liage, en se demandant comment se construit dans le cerveau la cohérence perceptive - multisensorielle, nous le verrons - de l'objet parole.

Multistabilité, révélateur de la nature des objets

Les travaux de cette thèse portent donc sur le problème du liage en parole. Les principales questions traitées dans le cadre de cette thèse concernent les mécanismes

de groupement de flux de parole, de l'émergence et la stabilisation des représentations de parole. Nous abordons ces questions avec une approche cognitive, à la fois expérimentale et computationnelle. Comme outil d'investigation, nous avons choisi le paradigme de la multistabilité perceptive, qui est maintenant devenu un paradigme classique dans les études sur le liage perceptif et la conscience. La multistabilité perceptive désigne des changements de perception qui ne sont associés à aucun changement physique dans la stimulation. Ce paradigme permet d'étudier plus directement les questions posées par le problème du liage en séparant les mécanismes et les activités neuronaux dus à la stimulation sensorielle et ceux de la perception consciente. En vision, la multistabilité perceptive peut se produire lorsqu'on observe des images « réversibles » comme le cube de Necker et le vase de Rubin. La figure I, à gauche, illustre le cube de Necker dont la perception est compatible avec les deux formes présentées au milieu de la figure. La perception du vase de Rubin (I, à droite) est également compatible avec deux formes, soit deux visages en noir soit un vase en blanc. Le paradigme de rivalité binoculaire entraîne également une perception visuelle multistable. Ce paradigme consiste à présenter deux images différentes à chaque œil, ce qui entraîne des bascules perceptive entre ces deux images (e.g. Tong *et al.*, 2006). Certains stimuli visuels dynamiques tels que les plaids dynamiques conduisent aussi à une perception multistable (e.g. Rubin et Hupé, 2005). Les plaids dynamiques consistent en deux grilles superposées qui se déplacent vers deux directions différentes dont le mouvement peut être compatible avec celui d'une seule surface (percept cohérent) ou celui de deux surfaces ayant deux directions différentes (percepts transparents) (voir figure 3.10).

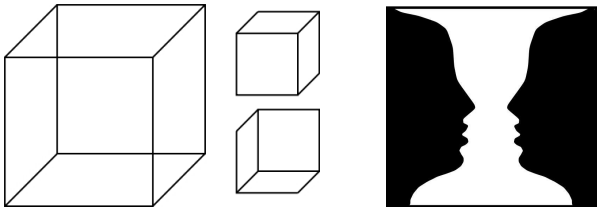


FIGURE I : Exemples d'images « réversibles » : à gauche, le cube de Necker (Necker, 1832) et à droite, le vase de Rubin (Rubin, 1958).

La plupart des études sur la multistabilité perceptive concernent la modalité visuelle mais ce phénomène existe également en modalité auditive. En audition, la suite de deux sons purs AB (A : fréquence basse, B : fréquence haute) peut également entraîner deux percepts différents (Bregman et Campbell, 1971). Si la différence de fréquence (Δf) entre A et B est faible et le débit de répétition est lent, on entend un seul flux auditif ABAB$\cdots$. En revanche, si Δf est grande et le débit est rapide, deux flux distincts, A$-$A$\cdots$ et B$-$B$\cdots$, sont perçus. La multistabilité perceptive peut avoir lieu lorsque Δf a une valeur intermédiaire. Dans ce cas, la perception bascule d'une forme (un flux auditif) à l'autre (deux flux auditifs) et vice versa (figure II).

En parole, la multistabilité perceptive est connue sous le nom de l'effet de transformation verbale (Warren et Gregory, 1958; Warren, 1961b) : en écoutant une répétition du mot anglais *life*, notre perception bascule du mot *life* au mot *fly* et ensuite de *fly* à *life* et ainsi de suite. Ce phénomène se produit également en

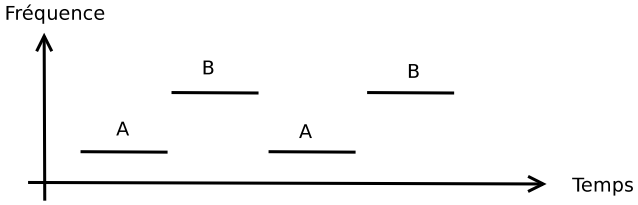


FIGURE II : Bistabilité auditive (Bregman et Campbell, 1971) : une différence fréquentielle intermédiaire entre les sons A et B peut conduire à une perception bistable, soit un seul flux ABA$\cdots$ soit deux flux séparés A$-$A$-\cdots$ et B$-$B$-\cdots$.

répétant soi-même le mot *life*. Pendant une tâche perceptive de transformation verbale, le stimulus peut donc entraîner deux ou plusieurs percepts différents, chacun compatible avec un liage différent entre les éléments du stimulus[1]. Ainsi, l'effet de transformation verbale permet d'isoler et de clarifier les processus fondamentaux qui conduisent à une interprétation appropriée d'un son ambigu (Warren et Warren, 1970). Ces processus fondamentaux associent mécanismes de liage, compétition entre différents liages possibles et prise de décision. C'est l'ensemble de ces processus qui conduisent à la constitution d'un « objet parole » au sens phénoménologique. Pendant cette thèse, nous essayerons de mieux comprendre la nature de cet « objet parole » tel qu'il est constitué dans le phénomène de multistabilité, mis en évidence par l'effet de transformation verbale.

Plan

Ce manuscrit est organisé en quatre parties. La première partie concerne l'état de l'art sur les trois composantes principales de cette thèse : le liage perceptif, la multistabilité perceptive et la parole. La deuxième et la troisième parties regroupent nos études expérimentales sur l'organisation perceptive multisensorielle de la parole, sur les corrélats neuronaux de l'émergence de l'objet parole et sur les mécanismes (possibles) impliqués dans l'organisation perceptive de la parole du point de vue computationnel. Nous conclurons cette thèse dans une quatrième partie en introduisant un cadre général pour expliquer le liage perceptif en parole et en faisant des propositions pour des études futures.

Chaque partie de ce manuscrit est organisée en chapitres, sections et sous-sections numérotés. Tout au long du manuscrit, nous utilisons les abréviations en anglais pour désigner les aires cérébrales. « BA » indique des aires corticales de Brodmann.

[1] Il est à noter que toutes les transformations perçues ne sont pas directement le résultat d'un liage entre les différents éléments du stimulus, voir section 3.3.

Première partie

État de l'art

Chapitre 1

Le liage perceptif, base de la formation des objets

Sommaire

Le liage perceptif, on l'a compris, recouvre des aspects multiples de la question de l'unité des objets perçus. Revonsuo (1999) en propose une définition au sens large : le liage concerne la capacité du cerveau à créer une représentation cohérente et intégrée du monde extérieur et des séquences fonctionnelles de comportements, ceci malgré le fait que les informations d'entrée du monde extérieur sont de nature sensorielle différente et qu'elles sont initialement traitées dans différents circuits cérébraux et siègent d'une manière fragmentée dans le cerveau. La question du liage perceptif peut être posée à trois niveaux : niveau phénoménologique, niveau des mécanismes neuronaux et niveau des mécanismes cognitifs (Revonsuo, 1999). Nous présentons dans la première section de ce chapitre des questions posées par la psychologie de la forme sur la nature phénoménologique des objets visuels. Dans la même section, nous présentons les mécanismes neuronaux et cognitifs qui pourraient être à la base du liage dans le cadre de la perception visuelle, fournissant les pré-requis de l'unité perceptive des objets et de la conscience. Il est à noter que nous n'étudions pas dans cette thèse le liage perceptif du point de vue de la philosophie de l'esprit en lien avec

le problème difficile de la conscience[1] tel qu'il est discuté par exemple par Chalmers (1995). La deuxième section de ce chapitre porte sur la formation des objets sonores et les mécanismes neuronaux impliqués. Une introduction sur l'objet parole et sa spécificité sera présentée dans la section trois.

1.1 Les objets visuels, de la psychologie de la forme aux modèles neurocognitifs

1.1.1 Les objets de la psychologie de la forme

Dans le champ de la perception visuelle, les psychologues de la « Psychologie de la forme » (*Gestalt Theory*) se sont intéressés dès la fin du XIXème siècle à la question de la nature de la cohérence des objets visuels au sein de scènes complexes. La question était alors essentiellement phénoménologique : qu'est-ce qui, dans une scène visuelle telle que celles de la Figure 1.1, permet d'extraire, spontanément, c'est-à-dire immédiatement, inévitablement et universellement, des objets tels que trois groupes distincts chacun composé de deux cercles sur la figure à gauche ou un triangle blanc en avant-plan sur la figure à droite.

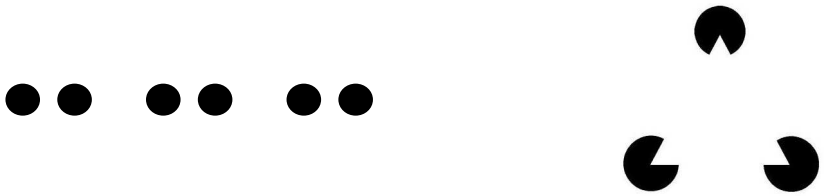


FIGURE 1.1 : Les objets visuels et la psychologie de la forme (*Gestalt theory*). À gauche : notre perception de cette scène visuelle correspond à trois groupes distincts composés de deux cercles. À droite : triangle de Kanizsa. Un triangle blanc est perçu en avant-plan, avec ses trois sommets situés à l'intérieur des trois cercles.

Les Gestaltistes ont proposé quelques principes d'organisation des objets dans des scènes, tels que loi de bonne continuité, loi de proximité, loi de similitude, loi de destin commun (voir figure 1.2). Ainsi, la perception de trois groupes distincts de deux points et celle d'un triangle en avant-plan sur la figure 1.1 peuvent être respectivement expliquées par la loi de proximité et la loi de réification. Selon la loi de réification, le système perceptif est capable de compenser les informations manquantes et le percept contient des informations spatiales plus importantes que le stimulus sensoriel sur lequel il est basé.

Les questions posées par la psychologie de la forme sur l'organisation des objets visuels ont concentré un grand nombre d'études expérimentales ou d'expériences de pensée dans toute la première partie du XXème siècle, en conduisant à des notions, souvent plus qualitatives que quantitatives ; de cohérence perceptive, de forme et

[1] Selon Chalmers (1995), le problème difficile de la conscience désigne le problème d'expliquer les expériences phénoménales qualitatives, autrement dit, l'aspect subjectif des expériences. Par exemple, quand nous voyons, nous avons l'expérience des sensations visuelles telles que la qualité de la rougeur ou de la profondeur dans le champs visuel.

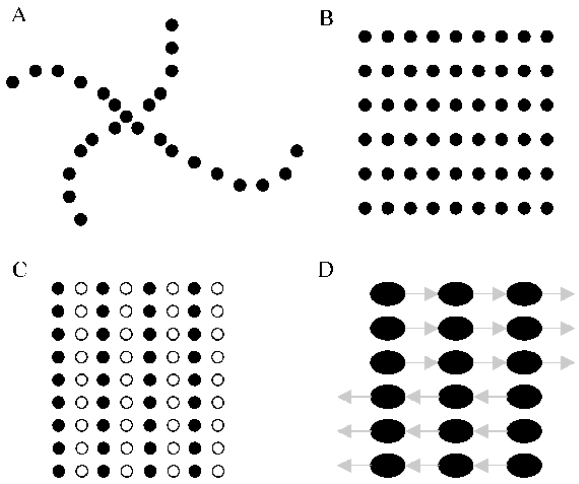


FIGURE 1.2 : Quelques principes à la base de l'organisation des objets proposés par la psychologie de la forme. A : loi de bonne continuité, B : loi de proximité, C : loi de similitude, D : loi de destin commun (le mouvement dans la même direction).

de fond, de contours et de propriétés, toujours autour de la question de la nature phénoménologique des objets.

1.1.2 Les modèles neuro-corrélationnels

Avec le développement des connaissances sur le cerveau et des mécanismes de traitement neurophysiologique de l'information visuelle, s'est posée une seconde question qui est celle du liage des représentations sensorielles correspondant aux multiples traces d'un objet ou d'une scène dans les nombreuses cartes visuelles décrites dans le cerveau humain. Les années 80-90 ont vu l'émergence de la notion de cohérence temporelle de phénomènes d'oscillations neuronales, sous des formes théoriques multiples, jusqu'à la mise en évidence expérimentale de cette notion, notamment par l'équipe de Wolf Singer (Gray *et al.*, 1989; Engel *et al.*, 1991; Singer et Gray, 1995). Selon ces données expérimentales, l'oscillation neuronale, spécialement dans la bande de fréquence gamma (>30 Hz), pourrait être le mécanisme permettant le liage perceptif entre différentes parties d'un objet visuel (liage perceptif de type bottom-up). Par exemple, Kreiter et Singer (1996) ont observé chez le singe macaque éveillé, dans l'aire MT impliquée dans la perception du mouvement (ou V5), des activités oscillatoires en réponse aux mouvements cohérents d'une barre mais pas lors de l'observation des mouvements indépendants de deux barres différentes.

Dans une expérience EEG (*Electroencephalography*) chez l'humain, Tallon-Baudry *et al.* (1996) ont également observé les oscillations gamma entre 30 et 60 Hz en lien avec la perception cohérente produite par les mécanismes de liage bottom-up. La figure 1.3, en haut, illustre les stimuli utilisés dans cette expérience. Les oscillations gamma ont été observées dans la condition 1 et 2 où le percept est cohérent avec un triangle mais pas dans la condition 3 où les différentes parties de la scène visuelle ne sont pas liées l'une à l'autre (voir figure 1.3, en bas). Il est important de noter

que l'activité gamma observée en lien avec le liage perceptif est de type induit (*induced gamma activity*). Contrairement à l'activité gamma évoquée (*evoked gamma activity*), les oscillations induites ne sont pas verrouillées en phase (*phase-locked*) par rapport à l'onset du stimulus. Autrement dit, le délai entre le début de stimulus et des oscillations induites est différent d'un essai (*trial*) à l'autre (pour une revue, voir Tallon-Baudry et Bertrand, 1999).

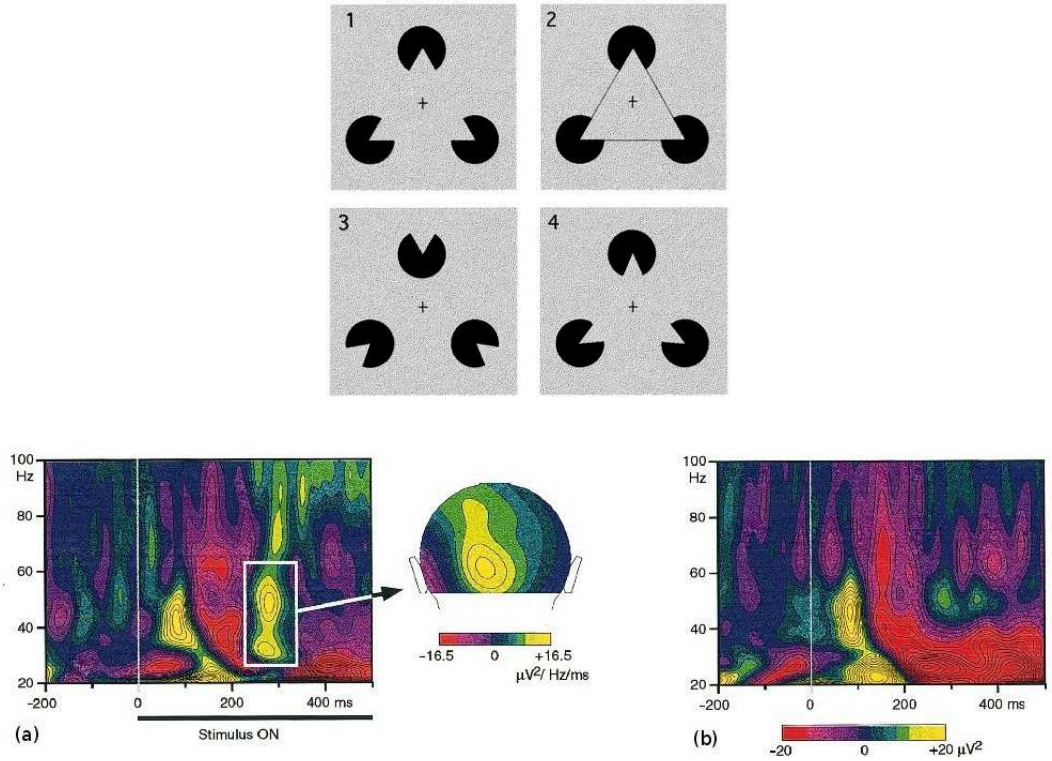


FIGURE 1.3 : En haut : les stimuli utilisés par Tallon-Baudry *et al.* (1996) pour vérifier le rôle des oscillations gamma dans le liage bottom-up ; 1 : percept d'un triangle illusoire (triangle de Kanizsa), 2 : percept d'un triangle réel, 3 : le percept n'est pas cohérent avec un triangle (pas de liage), 4 : le stimulus cible que les sujets devaient compter pendant l'expérience. En bas : la représentation temps-fréquence de l'activité d'une électrode (Cz) lors de la présentation d'un triangle de Kanizsa de l'image 1 (a) et lors de la présentation du stimulus de l'image 3 (b). Figures tirées de Tallon-Baudry *et al.* (1996) et Tallon-Baudry et Bertrand (1999).

Les oscillations de bande gamma joueraient également un rôle dans le liage perceptif de type top-down, par exemple, lors qu'on cherche un objet caché dans une figure. Dans une expérience EEG, Tallon-Baudry *et al.* (1997) ont présenté aux sujets la figure du Chien Dalmatien (figure 1.4). Les sujets naïfs ne percevaient pas le dalmatien et aucune activité induite en bande gamma n'a été observée pour ces sujets. Une fois les sujets entraînés à détecter le dalmatien, les auteurs ont observé une activité induite en bande gamma même s'il n'existait pas de dalmatien sur la figure. Tallon-Baudry *et al.* (1997) ont suggéré que cette activité reflète l'activation

top-down de la représentation de l'objet.

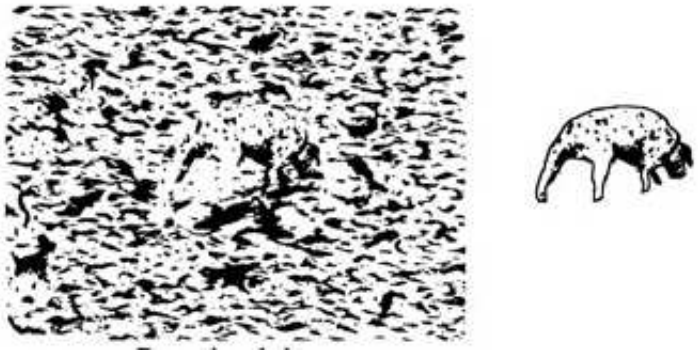


FIGURE 1.4 : Un des stimuli utilisés par Tallon-Baudry *et al.* (1997) mettant en évidence le rôle des oscillations en bande gamma dans l'activation top-down des représentations des objets. Figure tirée de Tallon-Baudry *et al.* (1997).

Bien que les premières études sur le rôle des activités oscillatoires en lien avec le liage perceptif aient été effectuées sur le liage entre différentes parties d'un objet visuel (*feature binding*), ces activités ont été également observées en lien avec le liage temporel utilisant les stimuli auditifs. Dans une expérience MEG (*Magnetoencephalography*), Joliot *et al.* (1994) ont présenté à des sujets deux clics successifs, et ils ont montré que les oscillations autour de 40 Hz fournissaient un bon prédicteur de la perception des sujets. En effet, les sujets percevaient un seul clic lorsque l'intervalle inter-stimulus était court et deux clics pour les intervalles inter-stimulus longs. Or, pour les intervalles courts ($<$12 ms), seule une activité oscillatoire correspondant au premier stimulus a été observée. Pour les intervalles plus longs, une autre activité oscillatoire apparaissait qui correspondait au deuxième stimulus (voir figure 1.5 pour une illustration).

Des activités oscillatoires cohérentes[2] ont été également observées dans des tâches sur l'attention sélective : les oscillations neuronales, surtout dans la bande gamma, sont plus importantes pour des stimuli auxquels les sujets portent attention par rapport aux conditions d'absence d'attention (Womelsdorf et Fries, 2007). L'activité gamma pourrait également jouer un rôle dans la mémoire à court terme et la mémoire de travail où la trace en mémoire en absence de la stimulation semble être établie par ces activités oscillatoires. La figure 1.6 illustre le scénario proposé par Jensen *et al.* (2007) concernant l'implication des synchronisations en bande gamma dans les tâches d'attention sélective et de mémoire de travail. Quant à la mémoire à long terme, les auteurs proposent que l'activité gamma intervient aussi bien dans la phase d'encodage des informations que dans le rappel des informations mémorisées à travers la modification de la plasticité synaptique : une mémorisation ou un rappel avec succès correspond à une augmentation des activités en bande gamma.

[2]Deux oscillations sont appelées cohérentes quand il y a une relation constante entre leurs phases au cours du temps. La synchronisation est un cas particulier où la différence de phase entre les oscillations est égale à zéro.

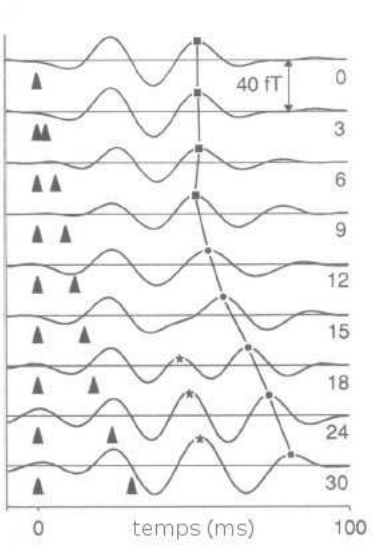


FIGURE 1.5 : Rôle des oscillations gamma dans le liage temporel : illustration de l'effet de l'intervalle interstimulus sur les activités MEG observées par Joliot *et al.* (1994). Les flèches représentent l'onset des deux clics. Les intervalles inter-stimulus (ISI) entre les clics sont affichés à droite. Pour les ISI<12 ms, une seule réponse correspondant au premier clic, autour de 40 Hz, a été observée tandis que pour les ISI plus long, une deuxième réponse apparaissait chevauchant celle évoquée par le premier clic. Les points représentent l'interaction entre le deuxième maximum de la réponse au premier clic et le deuxième maximum de la réponse au deuxième clic. Les étoiles représentent l'interaction entre le deuxième maximum de la réponse au premier clic et le premier maximum de la réponse au deuxième clic. Figure tirée de Joliot *et al.* (1994).

Il est également suggéré que l'activité oscillatoire dans les différentes bandes fréquentielles jouerait un rôle dans le liage perceptif entre les informations de nature sensorielle différente. Mishra *et al.* (2007) ont mis en évidence la présence des oscillations en bande gamma lors de l'intégration précoce des informations auditives et visuelles. Ils ont utilisé le paradigme de l'illusion visuelle induite par le son où un flash visuel accompagné par deux bips rapides est souvent perçu comme deux flash (figure 1.7 à gauche). Dans cette étude, les auteurs ont observé une augmentation des activités gamma dans le cortex occipital lors de la perception de deux percepts dans deux intervalles : 110-145 ms et 200-240 ms (figure 1.7 à droite).

Senkowski *et al.* (2008) proposent que le problème de liage intervient dans le traitement multisensoriel : premièrement parce que les informations doivent être intégrées à travers différentes aires corticales et sous-corticales et deuxièmement, parce que les signaux neuronaux de chaque canal sensoriel doivent être séparés des autres et en même temps, les signaux de différents canaux doivent être coordonnés ensemble d'une façon sélective. L'activité oscillatoire semble être le moyen adapté pour rendre compte de la sélectivité et de la flexibilité nécessaire pour le traitement des changements rapides des informations multisensorielles provenant du monde

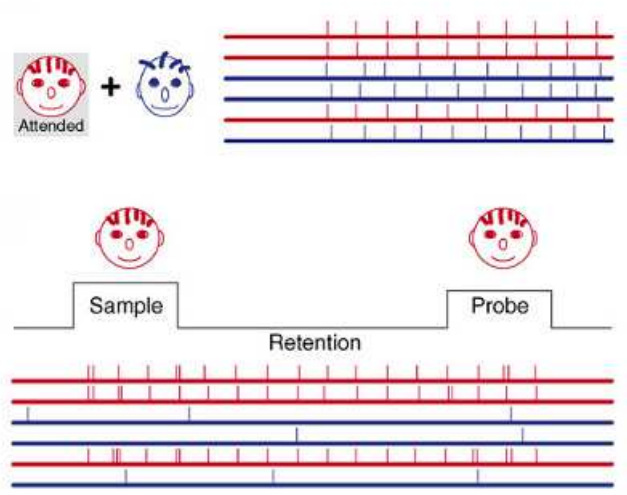


FIGURE 1.6 : La synchronisation en bande gamma pourrait être le mécanisme physiologique à la base de l'attention sélective et de la mémoire de travail (Jensen *et al.*, 2007). En haut : lorsque l'attention est dirigée vers le visage rouge, la synchronisation entre les neurones qui codent le visage rouge augmente. En bas : pendant la présentation du visage rouge, une synchronisation est établie entre les neurones encodant le visage rouge. Cette synchronisation est maintenue pendant l'intervalle de rétention, ce qui peut fournir le mécanisme à la base de la mémoire de travail. Figure tirée de Jensen *et al.* (2007).

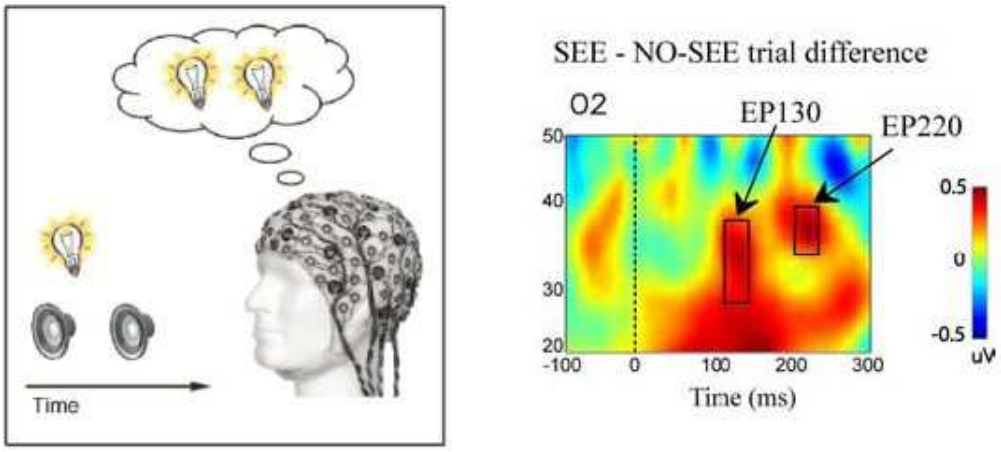


FIGURE 1.7 : À gauche : paradigme de l'illusion visuelle induite par le son. À droite : la différence entre l'activité d'une électrode dans le cortex occipital (O2) lors du percept de deux flashs (essai SEE) et percept d'un flash (essai NO-SEE). Figures tirées de Mishra *et al.* (2007) et Senkowski *et al.* (2008).

extérieur car ces oscillations permettent la connectivité fonctionnelle entre les populations de neurones, spatialement distribuées. Nous reviendrons sur ce point dans la section 2.3 sur l'intégration audiovisuelle de la parole où nous présentons des scénarios proposé par Senkowski *et al.* (2008) concernant le liage multisensoriel à travers des oscillations neuronales.

1.1.3 Les modèles psycho-attentionnels

Au-delà des niveaux phénoménologique et neuronal présentés ci-dessus, le 3ème niveau où doit se poser la question du liage est celui de mécanismes cognitifs assurant la mise en correspondance et l'unification des traces perceptives mono ou multisensorielles, et la cohérence des objets. En restant dans le champ de la perception visuelle, le modèle FIT (*Feature Integration Theory*) (Treisman et Gelade, 1980; Treisman, 1988) est un modèle cognitif qui essaie d'expliquer le problème de liage visuel en soulignant le rôle de l'attention spatiale dans la combinaison de différentes caractéristiques (*feature*) d'un objet en un tout.

Le modèle FIT, illustré sur la figure 1.8, contient deux types de cartes, une carte principale de localisation (*master map of location*) qui enregistre la localisation des objets dans différentes régions sans fournir l'accès aux caractéristiques qui définissent ces objets (par exemple, la couleur, la profondeur, etc.) et une série de cartes pour différentes caractéristiques (*feature maps*). Les cartes signalent si une caractéristique est présente dans la zone (drapeau sur la figure 1.8) et donnent des informations sur sa localisation spatiale. Selon ce modèle, le liage n'est pas nécessaire pour toutes les tâches, par exemple, lorsqu'on cherche à savoir si la couleur rouge est présente dans une scène visuelle, car on peut avoir un accès direct à cette information via les drapeaux. En revanche, le liage est nécessaire pour connaître par exemple la localisation d'un objet rouge et les autres caractéristiques de cet objet dans la scène. Le modèle FIT propose qu'une fenêtre attentionnelle se déplace sur la carte principale de localisation et sélectionne toutes les caractéristiques qui sont connectées à cette zone (voir figure 1.8), ce qui permet l'intégration des différentes caractéristiques d'un objet.

Le phénomène de la conjonction illusoire confirme certaines hypothèses du modèle FIT : lorsque les sujets font une recherche rapide dans une scène visuelle quand leur attention est chargée ou distraite, il devient plus difficile de se rappeler les attributs et ils peuvent faire des combinaisons qui sont fausses par exemple ils regroupent souvent les caractéristiques de deux objets différents en un seul objet. Dans une expérience, Treisman et Schmidt (1982) ont observé qu'après une présentation rapide de la lettre O en rouge, T en bleu et E en vert accompagnée par une tâche de distraction attentionnelle, certains sujets signalent l'observation d'une lettre E en rouge et T en vert (conjonction illusoire entre la couleur et la forme). En accord avec le modèle FIT, ce phénomène propose que les caractéristiques sont codées séparément et que l'attention joue un rôle dans le liage des différentes caractéristiques d'un objet.

Le liage perceptif par l'attention n'est pas en conflit avec les hypothèses de liage par les oscillations cohérentes. Nous avons vu dans la sous-section 1.1.2 que certains

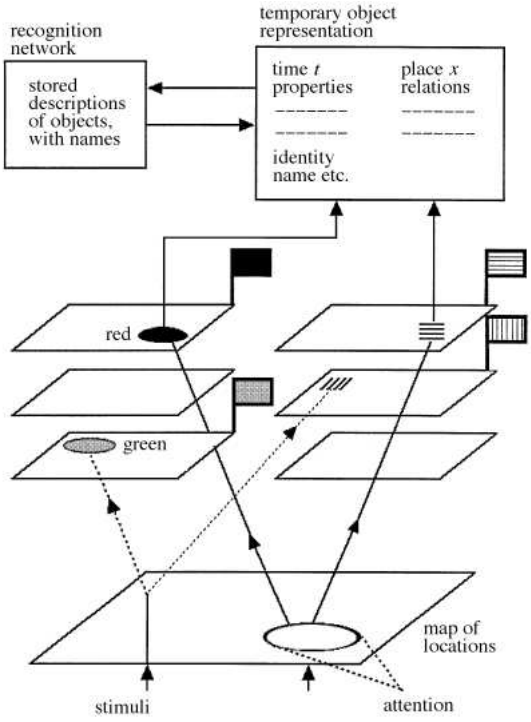


FIGURE 1.8 : Modèle FIT (Feature Integration Theory) de Treisman. Figure tirée de Treisman (1998).

auteurs proposent la synchronisation comme le mécanisme physiologique qui serait à la base de l'attention. Il existe des modèles qui combinent ces deux niveaux de description pour expliquer le liage perceptif. Par exemple, Hummel (2001) propose un modèle computationnel de liage perceptif par la sélection spatiale et la synchronisation neuronale dans le cadre de la reconnaissance des objets visuels et la perception de la forme. Ce modèle est un réseau de neurones artificiels au sein duquel différentes caractéristiques du même objet sont liées l'une à l'autre par la synchronisation des décharges neuronales. Dans ce modèle, l'attention visuelle est nécessaire pour établir les interactions inhibitrices qui conduisent à la désynchronisation des caractéristiques des différents objets les unes avec les autres.

1.2 Les objets sonores et l'analyse de scènes auditives

Dans la vie de tous les jours, nous sommes en permanence confrontés à des environnements sonores complexes où nos oreilles reçoivent un mélange de signaux émis par différentes sources sonores. Malgré ce mélange, nous sommes capables de percevoir séparément les différentes sources présentes dans l'environnement et de suivre indépendamment leurs émissions. Un exemple connu démontrant cette capacité de notre système perceptif est l'effet « cocktail party » (Cherry, 1953) : jusqu'à un certain point, nous sommes capables de suivre la conversation d'une personne alors que d'autres parlent en même temps ou de suivre une voix particulière en pré-

sence des sons ou des bruits (de fond) dans l'environnement. Les questions sur cette capacité essentielle de la perception auditive sont anciennes. Par exemple, en 1877, Helmholtz se demandait déjà comment on pouvait percevoir la qualité individuelle des instruments de musique dans un orchestre.

1.2.1 Primitives et schémas

En s'inspirant de l'analyse de scènes visuelles et la psychologie de la forme, Bregman propose un cadre, sous le nom d'analyse de scènes auditives (*auditory scene analysis*), afin d'expliquer la capacité de notre système perceptif d'intégrer et de séparer des sons et leurs composantes. Selon Bregman (1990), deux types de mécanismes sont impliqués dans l'analyse de scènes auditives, les mécanismes basés sur les primitives et ceux basés sur les schémas.

Les mécanismes basés sur les primitives sont conduits par les caractéristiques physiques du signal auditif. La plupart des processus primitifs peuvent être reliés à des principes de la psychologie de la forme. Une de ces primitives auditives est le degré de proximité fréquentielle ou temporelle : les éléments d'un son étant proches du point de vue fréquentiel ou temporel sont groupés dans un seul flux et génèrent un objet auditif distinct (voir figure 1.10). Ce principe est similaire à la loi de proximité en vision. Une autre primitive jouant un rôle dans l'analyse de scènes auditives est l'harmonicité. Une grande partie des sons de notre environnement sont des sons harmoniques (ex. les voyelles, beaucoup de sons musicaux, les cris des animaux). Pour des raisons liées à la production des sons, les différentes composantes spectrales des sons harmoniques sont des multiples entiers d'une fréquence principale. Le système auditif peut utiliser une stratégie pour exploiter cette régularité de sorte que les composantes harmoniques soient perçues comme faisant partie d'une même source sonore (Bregman, 1993). La synchronisation temporelle est également une primitive permettant l'organisation perceptive d'une scène auditive : les composantes fréquentielles dont le début est synchrone sont perçues la plupart du temps comme un seul objet auditif (voir Bregman, 1990, pour les autres primitives). Ces mécanismes basés sur les primitives sont de type bottom-up, automatiques et pré-attentifs (Bregman, 1990).

Le groupement perceptif ne peut pas toujours être basé sur les mécanismes primitifs. Pour démontrer cela, Bregman (1990) donne l'exemple de deux voyelles synthétiques ayant la même fréquence fondamentale dont le début et la fin sont synchronisés et qui proviennent de la même localisation spatiale. Les indices primitifs de l'analyse de scènes auditives prédisent un groupement de ces deux sons complexes en un seul flux. Pour cette scène sonore, Bregman suggère que l'utilisation des schémas est nécessaire afin de distinguer les deux voyelles. Un schéma mental est une structure de connaissance abstraite construite à partir de notre expérience sur notre environnement sonore (Neisser, 1967). Ces processus consistent à la sélection de l'information auditive en fonction de connaissances spécifiques de l'individu (par exemple, l'identification de son prénom dans un mélange de sons). Dans l'exemple ci-dessus, le schéma de chaque voyelle sélectionne la partie du spectre lui correspondant sans que les primitives soient impliquées dans la séparation du spectre. Contrairement

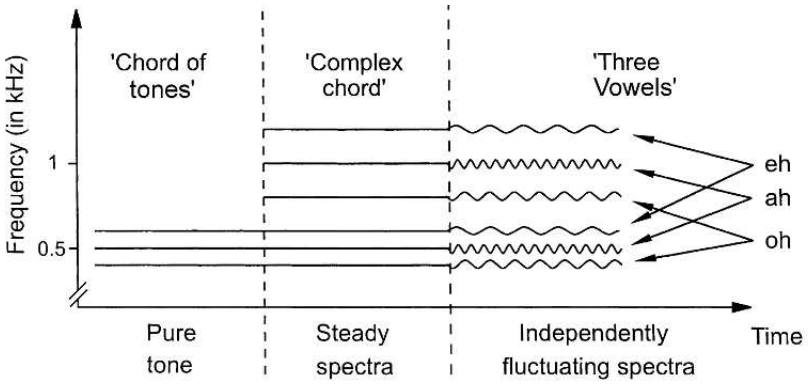


FIGURE 1.9 : Destin commun dans l'analyse de scènes auditives (Chowning, 1980). Figure tirée de Purwins *et al.* (2001).

aux mécanismes primitifs, l'analyse de scènes auditives par les schémas est de type top-down qui peut être automatique mais aussi volontaire et attentive (Bregman, 1990).

1.2.2 Le destin commun

Une des primitives utilisée dans l'analyse de scène auditive est le principe de « destin commun » (*common fate*) similaire à celui proposé par la psychologie de la forme. Selon le principe de destin commun en vision, les éléments de la scène visuelle dont les mouvements sont cohérents sont groupés ensemble et forment un seul objet visuel (voir figure 1.2, D). Ce principe peut également être appliqué à une scène auditive. Par exemple, lorsque deux composantes fréquentielles varient d'une manière synchrone et proportionnelle, il est très probable qu'elles sont des composantes d'un même son et qu'elles sont produites par la même source physique. Autrement dit, lorsque différentes parties d'un spectre subissent la même variation, fréquentielle ou d'intensité, au même moment, elles forment une seule unité perceptive et sont dissociées d'autres parties du spectre dont la variation est différente.

En 1980, Chowning a réalisé une expérience mettant en évidence le rôle du destin commun dans l'analyse de scène auditive (figure 1.9) : en premier lieu, trois sons purs ont été joués, ce qui a été perçu comme un accord contenant les trois hauteurs. Puis, les harmoniques de trois voyelles, « oh », « ah » et « eh », ont été ajoutées sans leurs fluctuations fréquentielles de sorte que les sons purs puissent être considérés comme leurs fréquences fondamentales respectives. Dans cette condition, un seul son a été perçu et l'existence des trois hauteurs différentes n'apparaissait pas à l'auditeur. Enfin, en ajoutant les fluctuations fréquentielles, la perception étaient cohérente avec les trois voyelles « oh », « ah » et « eh », chantées à trois hauteurs différentes.

1.2.3 Le modèle à canaux

Le phénomène de ségrégation du flux auditif (*auditory stream segregation*) ou streaming a été beaucoup utilisé dans les études comportementales et neurophysiologiques pour mettre en évidence les mécanismes impliqués dans l'analyse de scènes auditives. Comme présenté au début de ce manuscrit, Bregman et Campbell (1971) ont observé que, dans certaines conditions, une suite de sons purs alternant rapidement entre un son de fréquence basse et un son de fréquence haute entraîne la perception de deux flux distincts correspondant à un flux de son basse-fréquence et un flux de son haute fréquence. Ce phénomène est appelé streaming ou fission en comparaison avec la perception d'un seul flux qui alterne entre un son de basse et un son de haute-fréquence, phénomène appelé cohérence ou fusion.

En 1975, Van Noorden a publié une étude très complète sur les phénomènes de streaming et de cohérence. Il a utilisé une suite de sons purs basse-fréquence (A), haute-fréquence (B) et de silence (-) dans une configuration répétitive ABA-ABA-ABA-···. Il a observé que si la différence fréquentielle entre A et B est petite et le rythme de répétition est lente, les sujets perçoivent un seul flux ABA-ABA-···. En revanche, pour une différence fréquentielle suffisamment grande entre les sons A et B et un rythme rapide de répétition, la perception des sujets correspond à deux flux séparés de A-A-A-A-··· et B—B—··· (figure 1.10). Il a également montré que certaines combinaisons entre la différence fréquentielle et le rythme de répétition entraînent une perception bistable : la perception alterne de celle d'un seul flux (cohérence) à celle de deux flux distincts (streaming).

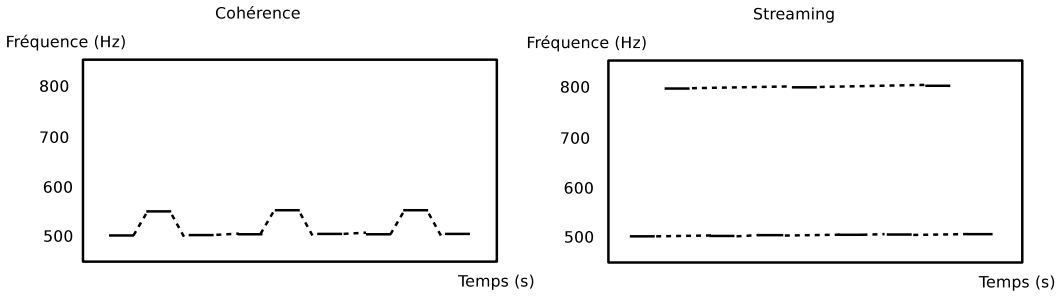


FIGURE 1.10 : Le phénomène de cohérence et de streaming. Les lignes en pointillé représentent les sons qui sont perçus connectés l'un à l'autre en un seul flux perceptif.

Une théorie influente expliquant le streaming auditif est la théorie des canaux (Hartmann et Johnson, 1991; Beauvois et Meddis, 1996). Selon cette théorie, le streaming est le résultat du traitement du signal auditif dans le système auditif périphérique : des filtres auditifs effectuent une analyse spectrale du signal et découpent le spectre du signal en différents canaux (bandes) fréquentiels. L'attribution des flux auditifs s'effectue selon la distance spectrale : les signaux dont les caractéristiques spectrales sont éloignées ont plus de chances d'être attribués à des flux auditifs différents.

Malgré la cohérence de plusieurs études comportementales avec la théorie des canaux et la pertinence des indices spectraux dans le streaming, certaines études ont

montré que le streaming peut également avoir lieu en absence des indices spectraux. Un des paradigmes utilisés pour mettre en évidence le rôle des indices non-spectraux dans le streaming est le paradigme des sons alternés de type ABA-ABA-$\cdots$, A et B étant des signaux complexes non-résolus. Les signaux non-résolus par le système auditif sont des signaux artificiels partageant les mêmes canaux spectraux mais perçus comme des signaux distincts. Par exemple, un ensemble d'études ont montré qu'une différence de hauteur entre A et B, entraînant une différence perceptive purement temporelle entre A et B, est un indice pertinent pour l'organisation perceptive des stimuli de type ABA-ABA-$\cdots$ (Vliegen et Oxenham, 1999; Grimault *et al.*, 2000). D'autres indices non-spectraux tels que la cadence de la modulation d'amplitude (Grimault *et al.*, 2002), ou le timbre (Cusack et Roberts, 2000), seraient également impliqués dans le streaming, ce qui met en défaut la théorie des canaux et le caractère seulement périphérique du streaming.

1.2.4 Corrélats neuronaux du modèle à canaux

Les études utilisant l'enregistrement unitaire et multi-unitaires des neurones chez l'animal ont mis en évidence le rôle de la sélectivité fréquentielle des neurones du cortex auditif primaire (A1) dans le streaming (e.g. Fishman *et al.*, 2001; Micheyl *et al.*, 2005) : pendant la présentation d'une suite ABA ou AB, lorsque la différence fréquentielle entre les sons A et B augmente, les neurones dont la fréquence caractéristique (FC, *best frequency*) correspond à celle du son A répondent de moins en moins au son B (voir figure 1.11). Ainsi, l'augmentation de Δf entre les sons A et B entraîne la ségrégation de ces deux sons, ce qui est évidemment cohérent avec le modèle à canaux.

Deux facteurs expérimentaux supplémentaires s'inscrivent également dans le cadre du modèle à canaux, en venant en raffiner le fonctionnement. Le premier est la suppression proactive (*forward suppression*) qui réfère à la réduction de la réponse neuronale à un stimulus par le stimulus qui le précède. Une des premières études montrant les effets combinés du rôle potentiel de la sélectivité fréquentielle et de la suppression proactive dans le streaming a été réalisée par Fishman *et al.* (2001). Ils ont enregistré les activités neuronales produites par des stimuli de type ABAB$\cdots$ dans le cortex A1 (cortex auditif primaire) des macaques éveillés. La fréquence du son A a été choisie de sorte qu'elle corresponde à la fréquence caractéristique (FC) du site cérébral enregistré. La fréquence du son B était variable mais elle était toujours loin de FC. Différents taux de présentation étaient utilisés (5, 10, 20 et 40 Hz) soit en variant les SOAs (*stimulus-onset asynchronies*, i.e. la durée entre l'onset de deux sons successifs) soit en modifiant les ITIs (*inter-tones intervals*, i.e. la durée entre la fin d'un son et l'onset du son suivant). Les auteurs ont observé que lorsque la présentation était lente (5 et 10 Hz), le site cortical enregistré ayant une FC correspondant au son A, répondait aux deux sons, A et B. En revanche, pour les présentations rapides (20 et 40 Hz), le site enregistré répondait seulement au son A. Une illustration de ces résultats est présentée sur la figure 1.12. Ce résultat est cohérent avec les résultats comportementaux montrant qu'une alternance rapide entre A et B entraîne la perception de deux flux séparés (streaming).

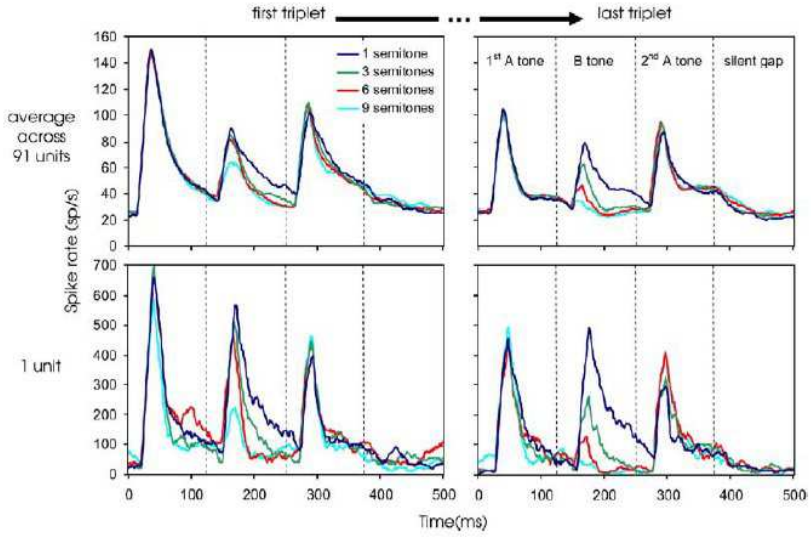


FIGURE 1.11 : Réponses neuronales à une suite auditive ABA chez le macaque éveillé, A et B étant des sons purs (fréquence caractéristique : celle du son A). Les lignes verticales en pointillé représentent le début ou la fin de chaque son. Figure tirée de Micheyl *et al.* (2005).

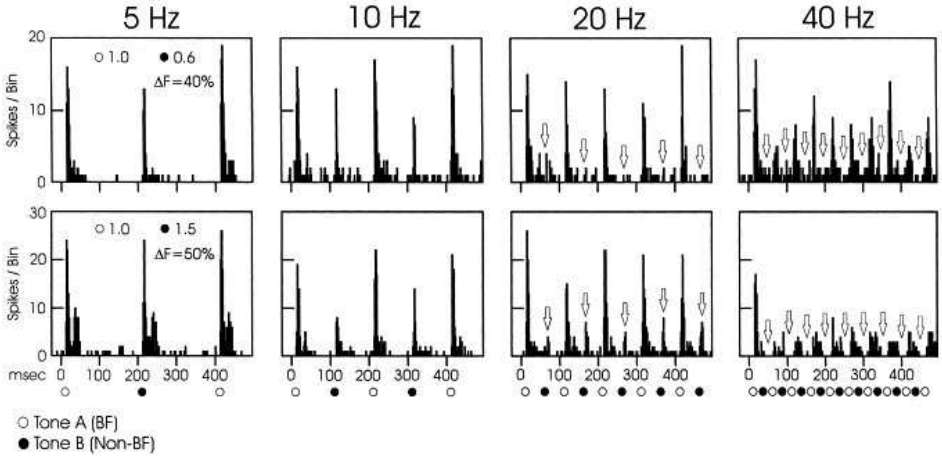


FIGURE 1.12 : Réponses neuronales enregistrées par deux électrodes dans le cortex auditif primaire pendant la présentation de stimulus ABAB$\cdots$. En haut : différents taux de présentation, 5, 10, 20 et 40 Hz. BF : fréquence caractéristique. Δf : différence fréquentielle entre les sons A et B par rapport à la fréquence caractéristique du site enregistré. Les flèches représentent la baisse ou l'absence de réponse au son B par rapport à celle pendant la présentation à 5 Hz (premières figures à gauche). Figure tirée de Fishman *et al.* (2001).

Le second mécanisme neuronal qui pourrait avoir un rôle dans le streaming est l'habituation. Le terme d'habituation a été utilisé par Micheyl et collègues pour désigner la baisse générale de réponse neuronale aux deux sons A et B (figure 1.11). Ce phénomène est différent du mécanisme de suppression proactive qui correspond à la baisse de réponse au son B mais pas au son A (figure 1.12, figures à gauche vs. figures à droite). Micheyl *et al.* (2005) ont observé l'habituation pendant la perception des stimuli de type ABA-$\cdots$ en prenant en compte la phase « *build up* » de streaming. Le *build up* concerne la phase où le streaming devient perçu, quelques secondes après le début du stimulus (le streaming n'a pas lieu au début du stimulus). Les auteurs ont observé que la réponse aux deux sons diminue au cours du stimulus (voir figure 1.11). Cette baisse est plus forte pendant les deux premières secondes du stimulus. Les auteurs ont proposé que cette baisse jouerait un rôle dans le streaming et pourrait expliquer le mécanisme de *build up*, comme on va le voir.

En effet, Micheyl *et al.* (2005) ont proposé un modèle pour expliquer les résultats comportementaux sur le streaming à partir des réponses neuronales aux stimuli de type ABA-$\cdots$. Selon leur modèle, une assemblée de neurones, à partir des activités neuronales dans le cortex auditif primaire, indique si un flux auditif (cohérence) ou deux flux auditifs (streaming) ont lieu. Cette assemblée se comporte comme un classificateur binaire et prend une décision en comparant le nombre de spikes évoqués par les sons A et B consécutifs avec un seuil (*threshold*) : si le nombre de spikes évoqués par le son A et le son B est plus grand que le seuil, un seul flux sera perçu mais lorsque le nombre des spikes évoqués par le son A dépasse le seuil mais pas le nombre de spikes évoqués par le son B, le modèle prédit deux flux séparés. Ainsi, la phase de *build up* du streaming peut être expliquée par l'habituation : les réponses qui dépassent le seuil au début du passage du stimulus peuvent baisser sous le seuil au cours de la présentation du stimulus à cause de l'habituation, ce qui conduit à la perception de deux flux.

Dans une étude récente enregistrant le noyau cochléaire du cochon d'Inde anesthésié, Pressnitzer *et al.* (2008) ont montré que la suppression proactive et l'habituation impliquées dans le streaming sont observables déjà dans le système auditif périphérique, et en ont déduit une implémentation du modèle à canaux dès ce niveau périphérique de la chaine de traitement.

1.2.5 Mécanismes attentionnels et contextuels

Outre les indices acoustiques présents dans le stimulus, des processus cognitifs de haut niveau peuvent également être impliqués dans le streaming. Carlyon *et al.* (2001) ont montré que le rôle de l'attention est important pour la perception du streaming. Ils ont observé que les sujets avec une négligence unilatérale gauche perçoivent moins de streaming par rapport aux sujets normaux lorsque les signaux de type ABA-ABA-$\cdots$ ont été présentés à l'oreille gauche. Quant à la présentation à l'oreille droite, il n'y avait pas de différence entre les deux groupes de sujets.

Le contexte qui précède le stimulus peut également influencer le streaming. Une expérience récente réalisée par Snyder *et al.* (2008) met en évidence cet effet de contexte sur le streaming des stimuli de type AB-AB-$\cdots$. Dans cette expérience, la

fréquence du son A était constante mais quatre différentes fréquences ont été utilisées pour le son B : la fréquence du son A et trois autres fréquences plus hautes que celle du son A. Le résultat de cette expérience montre un effet de contraste du contexte : les sujets ont plus tendance à percevoir le streaming si dans les stimuli précédents, la différence fréquentielle entre A et B était plus petite que dans le stimulus présent.

1.3 La spécificité de l'analyse de scènes de parole

Les indices primitifs permettant le groupement et la ségrégation auditive sont nombreux dans les signaux de parole. Est-ce qu'une analyse basée sur ces primitives conduit à un groupement exact des signaux de parole ? À cette question, Remez *et al.* (1994) ont répondu par la négative, dans un article qui présente clairement les arguments en faveur d'une « spécificité de l'analyse de scènes de parole ». Rappelons-en les principaux éléments.

Considérant des principes primitifs proposés dans le cadre de l'analyse de scènes auditives, il est surprenant que des portions différentes d'une même source de parole soient perçues comme cohérentes (Arons, 1992). Prenons l'exemple de la phrase « Why lie when you know I'm your lawyer ? » dont le spectrogramme est présenté sur la figure 1.13. Dans cet exemple, Remez *et al.* (1994) font remarquer la présence des indices primitifs suivants : une continuité du premier formant (F1), la discontinuité dans les fréquences hautes (F2, F3 et les formants nasaux), l'absence de similarité entre les trajectoires des fréquences hautes et l'absence de coïncidence temporelle des variations de ces fréquences. Selon les principes de l'analyse de scènes auditives, F1 serait considéré comme un flux continu, F2 aurait formé un flux discontinu avec des grandes variations fréquentielles et F3 et les formants nasaux seraient également séparés avec des variations fréquentielles moins importantes que celles de F2.

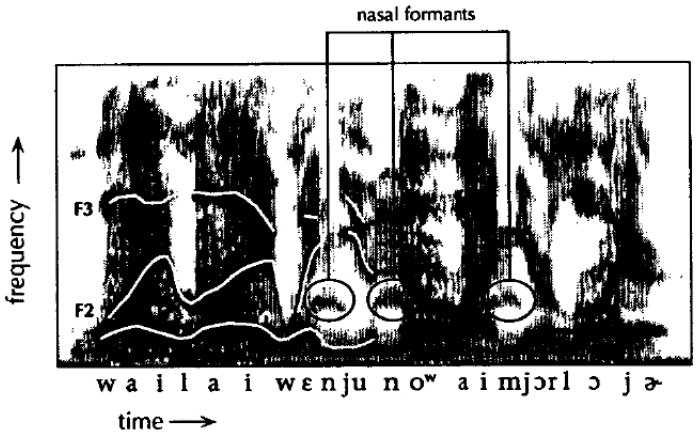


FIGURE 1.13 : La perception de la parole et les primitives auditives. Le spectrogramme de « Why lie when you know I'm your lawyer ? ». Voir le texte pour plus de détails. Figure tirée de Remez *et al.* (1994).

Dans cet exemple, le groupement de ces formants peut néanmoins être expliqué par le principe du destin commun proposé par Bregman (1990) dans le chapitre

consacré à la perception de la parole : ces résonances en effet ont une origine commune, celle de la mise en forme par le conduit vocal des vibrations laryngées, ce qui impose des liens harmoniques et une modulation d'amplitude commune entre ces différents formants. Ces deux indices, harmonicité et modulation commune, pourraient permettre le groupement des formants en un seul flux (Remez *et al.*, 1994).

Dans un deuxième exemple illustré sur la figure 1.14, Remez *et al.* (1994) indiquent qu'en appliquant le principe de la similarité spectrale et fréquentielle, la phrase « The steady drip is worse than a drenching rain » serait séparée en 10 flux distincts (voir Remez *et al.*, 1994, pour les détails). Mais cette fois, même en appliquant le principe de destin commun, utilisé dans le premier exemple, les quatre flux apériodiques correspondant aux sons fricatifs non-voisés, aux consonnes affriquées et aux déclenchements consonantiques (*release*) seraient séparés du reste du spectrogramme. Cet exemple suggère que les primitives auditives ne peuvent pas expliquer la cohérence perceptive du signal de parole.

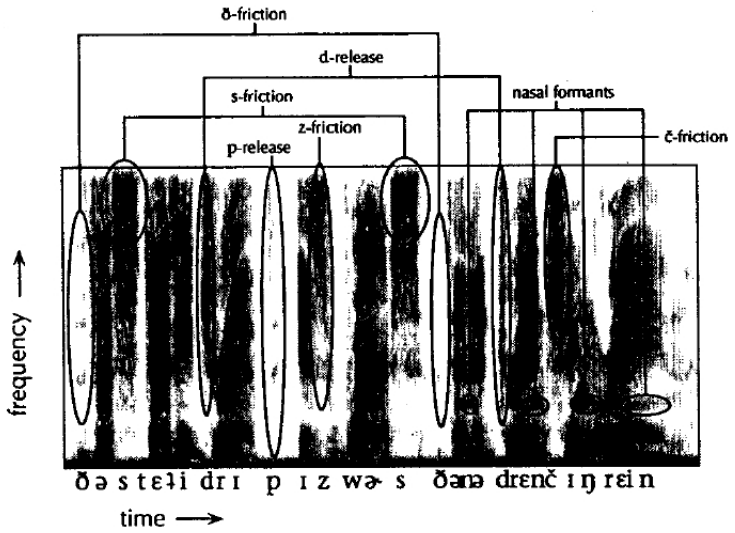
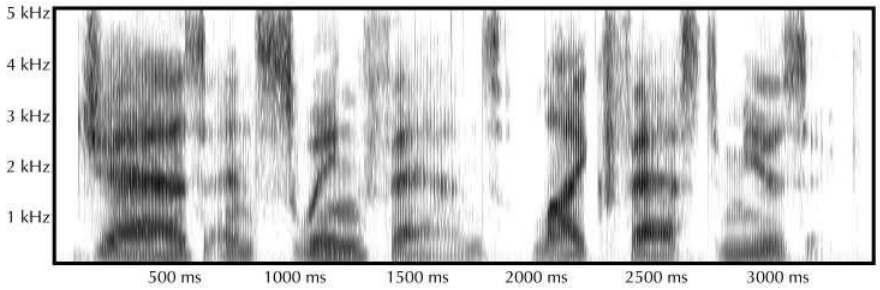


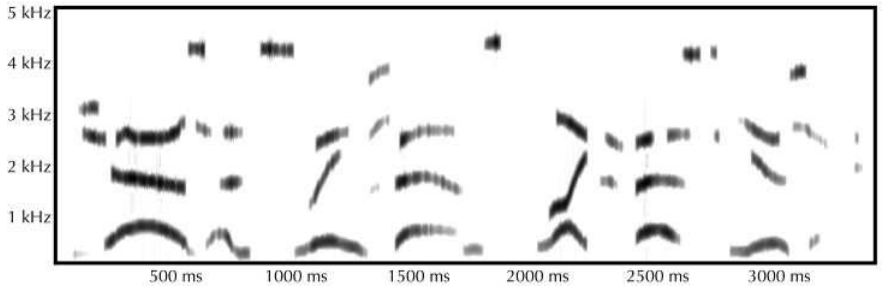
FIGURE 1.14 : La perception de la parole et les primitives auditives. Le spectrogramme de « The steady drip is worse than a drenching rain ». Voir le texte pour plus de détails. Figure tirée de Remez *et al.* (1994).

L'intelligibilité des signaux sinusoïdaux de parole (*sine-wave speech*, Remez *et al.*, 1981) met également en défaut l'idée de l'organisation perceptive de la parole basée sur les principes de l'analyse de scènes auditives. Les signaux sinusoïdaux de parole sont des signaux acoustiques synthétiques comprenant trois ou quatre signaux sinusoïdaux qui reproduisent les patterns de fréquence et d'amplitude des formants des signaux naturels de parole (pour un exemple, voir figure 1.15). Remez *et al.* (1994) indiquent que ni les primitives auditives ni les schémas à la Bregman ne sont capables d'expliquer l'intelligibilité des signaux sinusoïdaux de parole.

Sur la figure 1.15(b), on peut constater que la cohérence temporelle permettrait, dans la plupart des cas, le regroupement des trois premières composantes sinusoïdales. Cependant, un tel regroupement ne serait pas possible pour les composantes hautes fréquences isolées. De plus, la continuité globale de ce signal ne peut pas être



(a) Le spectrogramme de l'énoncé naturel « Jazz and swing fans like fast music ».



(b) Le spectrogramme de l'énoncé sinusoïdal « Jazz and swing fans like fast music ».

FIGURE 1.15 : Signal de parole naturel et son équivalant sinusoïdal. Figures tirées de Pardo et Remez (2006).

assurée par les principes de l'analyse de scènes auditives.

Est-ce qu'une analyse basée sur les schémas peut corriger cette prédiction ? L'hypothèse de l'implication des schémas dans la perception des signaux sinusoïdaux de parole n'est pas cohérente avec la définition des schémas. Nous rappelons que les schémas sont les patterns stockés en mémoire dont la formation est basée sur les expériences antérieures des individus. Malgré l'intelligibilité des signaux sinusoïdaux de parole, ces signaux ne ressemblent pas aux signaux de parole et ils ne sont pas perçus comme de la « vraie » parole. A l'issue d'une série d'expériences sur les signaux sinusoïdaux de parole, Remez *et al.* (1994) concluent que la parole est **spécifique** et proposent que d'autres principes que ceux de l'analyse de scènes auditives sont à la base de l'organisation perceptive de la parole et de la construction d'un objet spécifique de la perception auditive (et d'ailleurs multisensorielle, nous y reviendrons) que l'on peut dénommer « objet parole ». Nous utiliserons à partir de maintenant à dessein ce terme « d'objet parole », pour référer à ce qui est l'unité de traitement de la perception de la parole, et pour bien insister sur un ingrédient essentiel de cette thèse, l'idée que ces unités sont construites dans la perception par des mécanismes de constitution des objets impliquant des processus de liage généraux ou spécifiques, que nous cherchons à mettre au net. En retour, et c'est un autre ingrédient essentiel de cette thèse, nous considérons que la nature même de ces objets sera révélée ou en tout cas éclairée par la nature des processus de liage qui leur donnent forme dans la cognition humaine.

L'objet parole a été historiquement considéré soit comme un objet auditif (théories auditives de la perception de la parole) soit comme un objet de nature motrice

(théories motrices de la perception de la parole). Selon les théories auditives, les connaissances sur la façon dont les sons sont produits par le système articulatoire ne sont pas nécessaires pour la perception des sons. En revanche, les théories motrices considèrent que les objets parole sont des gestes articulatoires et non des objets auditifs. Du point de vue théorique, Remez et collègues inscrivent leur article sur la spécificité de l'organisation perceptive de la parole, dans le cadre des théories motrices de la perception de la parole proposant une indépendance entre l'analyse de scènes auditives et l'organisation phonétique (Remez *et al.*, 1994, p.151). Nous étudions les différentes théories sur la perception de la parole dans la section 2.1.

1.4 Conclusion

Nous avons vu dans ce chapitre les différents mécanismes, tant neuronaux que cognitifs, qui pourraient être à la base du liage perceptif en vision et en audition, et par là, responsables de la construction de l'objet visuel et l'objet auditif. Quant au liage perceptif en parole, nous avons présenté quelques éléments montrant sa spécificité par rapport aux mécanismes du liage auditif. En effet, nous avons vu que l'objet parole serait basé à la fois sur des principes auditifs généraux (primitives auditives et schémas mémorisés) et des principes phonétiques propres. Utilisant les termes proposés par Bregman, ces principes phonétiques spécifiques pourraient être considérés soit comme des primitives perceptuo-motrices soit comme des schémas phonétiques. Nous reviendrons sur ce point dans le chapitre 2 en présentant des éléments de réponses théoriques à la question de la nature de l'objet parole et des mécanismes de l'analyse de scène de parole. Mentionnons à ce point de notre parcours théorique que l'on constate bien à ce stade l'intérêt de reposer les questions sur la nature des objets de la perception de la parole à la lumière des paradigmes de structuration perceptive du type de l'analyse de scènes auditives, et d'autres sur lesquels nous reviendrons dans le chapitre 3.

Chapitre 2

La perception de la parole

Sommaire

La première section de ce chapitre porte sur une revue de la littérature sur la nature de l'objet parole selon les différentes théories de la perception de la parole. La deuxième et la troisième section concernent respectivement les modèles anatomiques fonctionnels de la perception de la parole et l'aspect multisensoriel des objets parole.

2.1 L'objet parole, entre les théories motrices et les théories auditives

La nature des objets de la perception de la parole (les « objets parole ») fait l'objet d'un débat classique dans la littérature, ce qui a donnée naissance à trois types de théories de la perception de la parole. Le premier type de théorie concerne la théorie motrice et la théorie réaliste directe selon lesquelles les objets parole sont de nature articulatoire. Le deuxième type recouvre les théories auditives qui proposent que les objets parole sont de nature auditive. Il est important de noter que nous utilisons le

terme « théories auditives » pour les distinguer par rapport aux théories motrices. Les théories auditives peuvent donc être cohérentes avec l'aspect multisensoriel des objets parole (ex. audio-visuel) que nous présentons à la fin de ce chapitre. Diehl *et al.* (2004) utilisent le terme « approches générales de la perception de la parole, auditives et basées sur l'apprentissage envers la perception de la parole » (*general auditive and learning approaches to speech perception*) pour désigner ces théories. Enfin, un troisième type de théorie concerne les théories perceptuo-motrices qui envisagent une nature perceptuo-motrice pour les objets parole. Dans cette section, nous faisons une revue de ces trois types de théories.

2.1.1 La théorie motrice et la théorie réaliste directe

Dans les années 50, Liberman et collègues proposent la théorie motrice de la perception de la parole (*Motor Theory of Speech Perception*) suggérant que les objets de la perception de la parole sont plutôt des événements articulatoires qu'auditifs (Liberman *et al.*, 1967; Liberman et Mattingly, 1985; Liberman et Whalen, 2000). Selon cette théorie, les événements articulatoires récupérés par les auditeurs sont des commandes neuro-motrices transmises aux articulateurs tels que la langue, les lèvres et les cordes vocales. Ces événements articulatoires ne sont pas les réels mouvements des articulateurs mais ce sont les gestes intentionnels (*intended gesture*). Selon la théorie motrice, ce sont les commandes neuro-motrices qui garantissent l'invariance des objets parole malgré la variabilité de leurs propriétés acoustiques. Liberman *et al.* (1967) illustrent ce propos en fournissant l'exemple des patterns de deux formants synthétiques qui sont perçus comme des syllabes /di/ et /du/ (figure 2.1). Dans cet exemple, les parties transitoires du début des syllabes portent des informations sur la consonne : la partie montante de F1 indique que la consonne est une plosive voisée comme /b/, /d/ ou /g/ et la partie montante de F2 dans /di/ et descendante dans /du/ donnent des informations sur le lieu d'articulation, ici une alvéolaire /d/. Malgré la différence de ces deux patterns au niveau acoustique, ils conduisent à la perception du même phonème /d/. Ces résultats conduisent ainsi Liberman et collègues à supposer l'implication de connaissances motrices implicites sur la coarticulation.

La théorie motrice propose également que la perception de la parole ne peut pas être expliquée par les mécanismes généraux de la perception auditive. Elle suggère qu'un module spécifique à la parole est à l'origine de la perception de la parole. Ce module est spécifique à l'homme et fait partie du système biologique spécialisé du langage chez l'humain (Liberman, 1996, chapitre 1). Ce module accomplit la perception de la parole par des processus similaires aux processus de la production de la parole :

> [T]he candidate signal descriptions are computed by an analogue of the production process - an internal, innately specified vocal trace synthesizer ... - that incorporates complete information about the anatomical and physiological characteristics of the vocal tract and also about the articulatory and acoustic consequences of linguistically significant gestures. (Liberman et Mattingly, 1985, p.26)

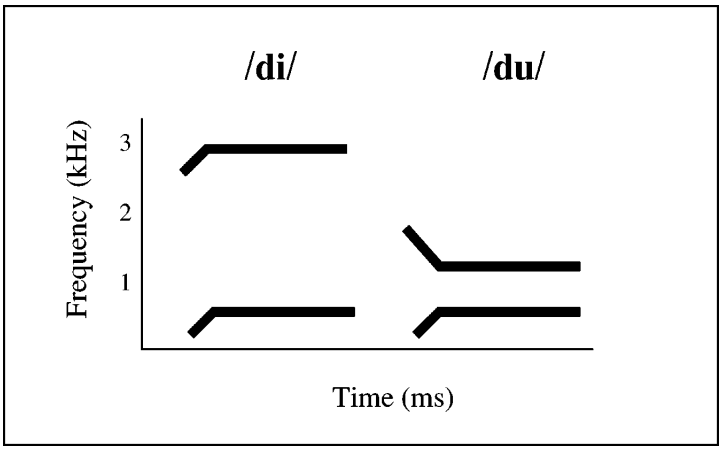


FIGURE 2.1 : Patterns formantiques de deux syllabes synthétiques /di/ et /du/. Les parties transitoires de F2 des deux patterns sont différentes, cependant, elles signalent la même information sur le lieu d'articulation de la consonne malgré leur variabilité acoustique. Figure tirée de Liberman (1996).

Dans les année 80, Fowler propose une autre alternative à la théorie motrice appelée la théorie réaliste directe de la perception de la parole (*Direct Realist Theory of speech perception*) (Fowler, 1986a,b, 1994). De la même manière que la théorie motrice, la théorie réaliste directe propose que l'objet parole est articulatoire mais elle suggère que ces objets articulatoires sont les vrais gestes et les vrais mouvements des articulateurs et pas les commandes neuro-motrices comme proposé par la théorie motrice. Une autre différence entre ces deux théories concerne l'hypothèse de la spécificité de la parole proposée par la théorie motrice. Fowler considère que la perception de la parole n'est pas spécifique et qu'elle peut être expliquée par les mêmes mécanismes que ceux, par exemple, de la perception visuelle ou tactile :

> Perceptual systems have a universal function. They constitute the sole mean by which animals can know their niches. [· · ·] even though it is the structure of the media (light for vision, skin for touch, air for hearing) that sense organs transduce, it is not the structure of those media that animals perceive. Rather, essentially for their survival, they perceive the components of their niche that caused the structure. (Fowler, 1996, p. 1732)

Les gestes articulatoires structurent donc le signal acoustique, ce qui sera ensuite utilisé comme médium de l'information et permet aux auditeurs de récupérer les gestes. Selon la théorie réaliste directe, aucun processus d'inférence ou de vérification d'hypothèse n'est utilisé pour la perception car l'information dans le signal acoustique est suffisante pour préciser les gestes qui structurent le signal (« directement »). Cette théorie est appelée réaliste car, selon elle, les auditeurs récupèrent les véritables causes physiques de la stimulation sensorielle, les gestes articulatoires, sans l'intervention des commandes neuro-motrices ou des processus cognitifs ou perceptifs.

Concernant le phénomène de la coarticulation, qui est un des défis expérimentaux ayant motivé la naissance de la théorie motrice de la perception de la parole, la théorie réaliste directe propose que le chevauchement temporel entre les voyelles et les consonnes est une conséquence de la coproduction des gestes vocaliques et consonantiques. Ces gestes structurent le signal acoustique d'une manière indépendante et les auditeurs sont donc capables de les récupérer dans le signal acoustique malgré leur superposition articulatoire. Dans les expériences mentionnées par Fowler et Smith (1986) sur l'identification et la discrimination des phonèmes, les sujets ont utilisé les informations anticipées de la coarticulation d'un segment pour la première partie de ce segment et ils ont perçu la deuxième partie, également influencée par la coarticulation, sans aucune influence contextuelle. Fowler et Smith (1986) suggèrent que comme dans les principes de « l'analyse vectorielle », introduits par Johansson (1973) dans la perception du mouvement biologique, les événements complexes tels que les séquences coarticulées peuvent ainsi être séparés en composantes indépendantes.

La découverte des neurones miroirs a été considérée comme un soutien de poids à l'hypothèse de la nature articulatoire de l'objet parole. Les neurones miroirs ont été observés pour la première fois par l'équipe de Giacomo Rizzolatti dans l'aire F5 du cortex prémoteur ventral du macaque. Ces neurones déchargent aussi bien lors de l'exécution par le macaque d'une action spécifique (ex. saisir un objet) que lors de l'observation par le macaque de l'exécution d'une action similaire par l'expérimentateur. Il a été proposé que ces neurones forment un système perceptif de reconnaissance des gestes impliqué dans la compréhension des actions perçues (Rizzolatti *et al.*, 2001). Les neurones miroirs ont été observés lors de la présentation de stimuli visuels liés à des actions transitives (c'est-à-dire impliquant un effecteur et un objet, comme casser une cacahuète avec la main) et lors de la présentation de stimuli auditifs associés (comme le bruit d'ouverture d'une cacahuète) (Kohler *et al.*, 2002). Ces neurones ont été appelés les neurones miroirs audio-visuels. De manière intéressante Ferrari *et al.* (2003) ont également montré l'existence de neurones miroirs associés aux actions intransitives orofaciales communicatives (figure 2.2, à droite) ou non (figure 2.2, à gauche). Cette découverte démontre que certains des neurones miroirs de la région F5 du macaque, homologue de l'aire de Broca chez l'humain, sont reliés à des actions orofaciales impliqués dans les fonctions communicatives (Ferrari *et al.*, 2003).

FIGURE 2.2 : Exemple des actions orofaciales pour lesquelles l'activation de neurones miroirs a été observée dans l'aire F5 du macaque. Figure tirée de Ferrari *et al.* (2003).

En ce qui concerne la parole, le système miroir chez l'humain a été suggéré comme

ayant un rôle fondamental dans l'émergence du langage et donc dans la perception de la parole (Rizzolatti et Arbib, 1998). Cette proposition nécessite l'implication des aires de planification et de production des gestes de la parole et des aires somatosensorielles dans la perception de la parole. Nous reviendrons sur ce point dans la section 2.2 qui concerne la neuroanatomie fonctionnelle de la perception de la parole. Des études utilisant la méthode de Stimulation Magnétique Transcranienne (TMS) suggèrent l'implication d'un système miroir chez l'homme dans la perception de la parole. Fadiga *et al.* (2002) ont enregistré les potentiels électromyographiques (EMG) des muscles de la langue lors de l'écoute de mots et pseudo-mots contenant des phonèmes dont la production articulatoire dépend ou non de ces muscles. Ils ont observé une augmentation sélective des potentiels électromyographiques lors de la stimulation de l'aire associée dans le cortex moteur primaire gauche, en cas d'écoute de mots ou pseudo-mots impliquant la langue. Watkins *et al.* (2003) ont observé un résultat similaire pour l'audition mais aussi la vision d'actions labiales (avec renforcement sélectif des potentiels électromyographiques des muscles des lèvres). Une étude récente en IRMf montre également l'activation de patterns somatotopiques similaires dans le cortex moteur primaire et le cortex prémoteur lors de la production et de la perception du phonème [p] associé aux mouvements des lèvres et du phonème [t] associé aux mouvements de la langue (Pulvermüller *et al.*, 2006). Ces résultats ont été interprétés en faveur d'un système miroir chez l'homme (Rizzolatti et Craighero, 2004) et en faveur de l'implication des représentations motrices dans la perception de la parole (Galantucci *et al.*, 2006) (mais voir aussi Hickok (2009a) et Dinstein (2008) pour des débats sur l'existence du système miroir chez l'homme et Lotto *et al.* (2009) pour une critique sur le lien entre les neurones miroirs et la théorie motrice de la perception de la parole).

2.1.2 Les théories auditives

Une alternative à la théorie motrice et la théorie réaliste directe concerne les théories auditives de la perception de la parole. Les théories auditives proposent que les objets parole sont plutôt de nature acoustique-auditive que de nature articulatoire. Les théories auditives de parole ne nient pas l'existence des liens entre la perception et la production de la parole et elles proposent que la production suit la perception et la perception suit la production (« production follows perception and perception follows production ») (Diehl *et al.*, 2004, p. 167). Cependant, selon ces théories, les connaissances motrices n'interviennent pas dans les processus de perception de la parole et la perception suit la production dans la phase d'apprentissage où on apprend les sons produits par nos partenaires de communication. En revanche, c'est la production qui suit la perception d'une manière active pour produire des gestes acoustiquement adaptés sur la base de la dispersion ou des invariances naturelles que nous présentons dans la suite.

La théorie de la dispersion (Liljencrants et Lindblom, 1972) est basée sur la préférence dans les langues du monde pour les voyelles [i a u]. Selon la théorie de la dispersion, cette préférence est due à la distance entre ces voyelles dans l'espace vocalique. En effet, comme illustré sur la figure 2.3, ces voyelles sont placées aux trois

extrémités du triangle vocalique. Cette dispersion entraîne une distinction maximale entre ces voyelles, ce qui facilite la communication par la réduction de la vraisemblance de la confusion entre elles. Par la suite, Lindblom (1986, 1990) a modifié la théorie de la dispersion - initialement « maximale » - et proposé la théorie de la variabilité adaptative selon laquelle la dispersion n'a pas besoin d'être maximale mais doit être suffisante pour permettre une distinction perceptive.

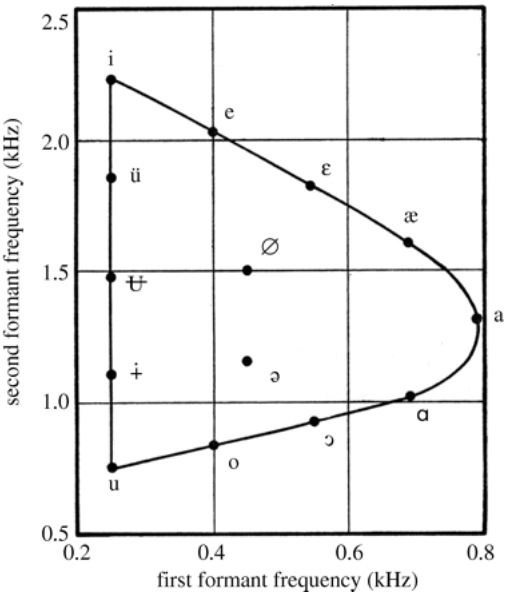


FIGURE 2.3 : L'espace des voyelles sur le triangle vocalique. Selon la théorie de la dispersion, la position des voyelles [i a u] aux extrémités du triangle vocalique entraîne une distinction maximale entre elles, ce qui peut expliquer leurs préférences dans les langues du monde. Figure tirée de Diehl (2008).

Diehl et Kluender (1989a,b) proposent, dans le cadre de l'hypothèse du renforcement auditif (*auditive enhancement hypothesis*), que le principe de dispersion est également utilisé pour les communications quotidiennes. Prenons l'exemple de la voyelle [u] caractérisée du point de vue acoustique notamment par la valeur de son deuxième formant (F2) de basse fréquence. Pour produire [u] dans un environnement bruité, les locuteurs choisissent typiquement une stratégie articulatoire pour baisser de plus en plus la fréquence de F2 (en retirant et montant la langue, en élargissant le pharynx, en avançant la racine de la langue, en baissant le larynx, etc.), ce qui rend la voyelle [u] la plus distincte possible par rapport aux autres voyelles. Selon cette théorie, l'objet parole est donc auditif et les gestes articulatoires sont choisis de sorte que les propriétés acoustiques qui permettent la distinction perceptive entre les différents sons soient préservées ou renforcées.

Stevens et Blumstein (1978) suggèrent quant à eux que la perception de la parole est basée sur des propriétés invariantes dans le signal acoustique (théorie de l'invariance acoustique). En accord avec cette théorie, Stevens (1972, 1989) propose la théorie quantique de la parole qui est basée sur la relation non-linéaire entre les paramètres acoustiques et articulatoires. La figure 2.4 illustre cette relation non-

linéaire. Selon Stevens, les langues du monde exploitent cette non-linéarité afin de renforcer le degré de distinction auditive.

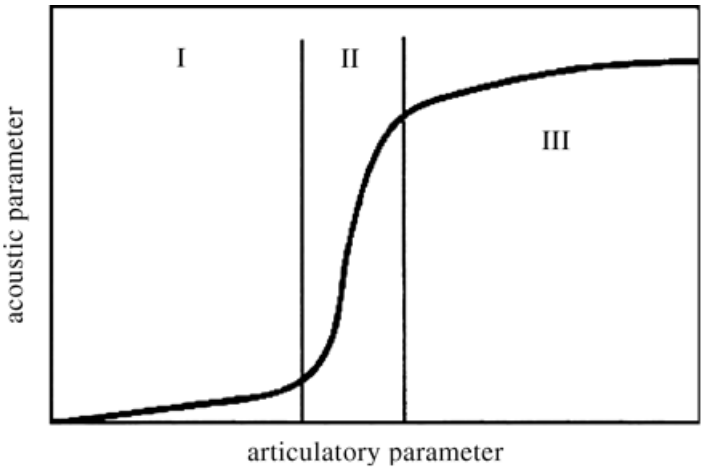


FIGURE 2.4 : La relation non-linéaire acoustique-articulatoire : un petit changement articulatoire dans la région I ou III provoque peu de changement acoustique tandis que le même changement provoque une modification acoustique importante dans la région II. Figure tirée de Stevens (1989).

Comme illustré sur la figure 2.4, pour certaines positions des articulateurs (lèvres, langues, cordes vocales, etc.), un petit changement a peu de conséquences acoustiques (région I et III) ; en revanche, pour d'autres positions, le même changement conduit à des modifications importantes au niveau des paramètres acoustiques (région II). La théorie quantique de Stevens propose que les langues du monde sont déterminées de sorte que les traits phonémiques soient associés avec les patterns acoustiques les plus stables : les deux plateaux I et III sur la figure 2.4 peuvent être associés à deux valeurs différentes du trait phonémique et la région de bascule II est évitée par la production, pour servir de frontière perceptive. On peut constater que, comme dans le cas de la dispersion, cette invariance est utilisée afin de produire des gestes acoustiquement adaptés pour la perception.

Selon les théories auditives, la perception suit donc des principes auditifs génériques et le développement se niche dans ces principes : la perception des sons coarticulés ou l'acquisition et la perception de la parole chez le nouveau-né seraient basées sur les mécanismes généraux de l'apprentissage perceptif (Diehl *et al.*, 2004) sans que les connaissances motrices soient nécessaires.

2.1.3 La théorie de la perception pour le contrôle de l'action

En ce qui concerne les théories perceptuo-motrices de la perception de la parole, nous nous focalisons sur la théorie de la perception pour le contrôle de l'action développée à l'Institut de la Communication Parlée et ensuite au GIPSA-lab. La théorie de la perception pour le contrôle de l'action propose que les objets parole ne sont ni purement sensoriels ni purement moteurs mais que ce sont des unités

perceptuo-motrices structurées par leurs valeurs perceptives et régularisées par l'action. Autrement dit, selon cette théorie, la perception forme l'action et l'action met des contraintes sur la perception (Schwartz *et al.*, 2002). Nous allons présenter ces deux concepts dans la suite.

La récupération des gestes articulatoires à partir du son est un problème « inverse », ce qui devient particulièrement problématique lors qu'il existe plusieurs gestes pour la production du même son (une relation articulation-son de type plusieurs-à-un : « many-to-one »). Le système de production de parole profite de ce fait lors qu'il prépare la cible suivante d'une façon anticipée, sans produire de changement audible du son (Schwartz *et al.*, 2002). La non-linéarité de la transformation articulatoire-à-acoustique peut également être problématique pour les théories motrices. Prenons l'exemple de l'arrondissement des lèvres : si on commence d'arrondir les lèvres en partant de la voyelle [i], on diminue progressivement la surface des lèvres, ce qui ne change pas le son au début, mais à un certain moment, le son change brusquement vers un son similaire à [y]. Ce phénomène suggère que l'arrondissement des lèvres ne peut pas être défini d'une manière absolue en termes purement articulatoires car l'arrondissement des lèvres signifie la baisse de la surface des lèvres jusqu'une valeur critique pour laquelle le son change brusquement (Schwartz *et al.*, sous presse). Les gestes de parole sont donc des gestes mis en forme par la perception.

Non seulement les gestes sont mis en forme par la perception mais ils sont également sélectionnés en lien avec leurs valeurs perceptives. Pour démontrer ce propos, Schwartz *et al.* (2007) présentent l'exemple des systèmes des voyelles des langues du monde dont la plupart contiennent les voyelles [i a u]. Du point de vue de la dispersion articulatoire, les voyelles [i a u], on l'a vu, sont de très bons choix. La figure 2.5 illustre l'espace articulatoire de ces voyelles. On peut constater sur cette figure que le système de voyelles avec [ɯ a y] est aussi bon qu'un système avec les voyelles [i a u]. Ce système combine l'arrondissement (non-arrondissement) des lèvres avec la position avant (arrière) de la langue, à l'inverse de [i a u]. Cependant, ce deuxième système de voyelles n'a jamais été observé dans les langues du monde. Schwartz *et al.* (2007) proposent que cette absence a des raisons auditives : [i u] sont meilleurs que [y ɯ] du point de vue de la dispersion acoustique (voir figure 2.5, à droite), ce qui suggère que les gestes sont choisis par rapport à leurs valeurs perceptives.

Malgré le fait que la théorie de la perception pour le contrôle de l'action suggère que les gestes de parole sont mis en forme par les traitements auditifs, cette théorie n'est pas une théorie auditive. En effet, la théorie de la perception pour le contrôle de l'action propose également que le système de production peut intervenir dans l'organisation perceptive des sons, ce qui distingue cette théorie par rapport aux théories auditives de la perception de la parole. Une série de données en faveur de cette proposition concerne les observations de Ménard *et al.* (2008) sur le contrôle idiosyncrasique de la hauteur des voyelles du français[1] pour les locuteurs âgés de 4 ans jusqu'à l'âge d'adulte. Ménard *et al.* ont observé que la distribution des voyelles par rapport à F1 est très variée d'un locuteur à l'autre (voir figure 2.6 pour deux

[1] [i y u] : voyelles hautes ; [e ø o] : voyelles mi-hautes ; [ɛ œ ɔ] : voyelles mi-basses ; [a] : voyelle basse.

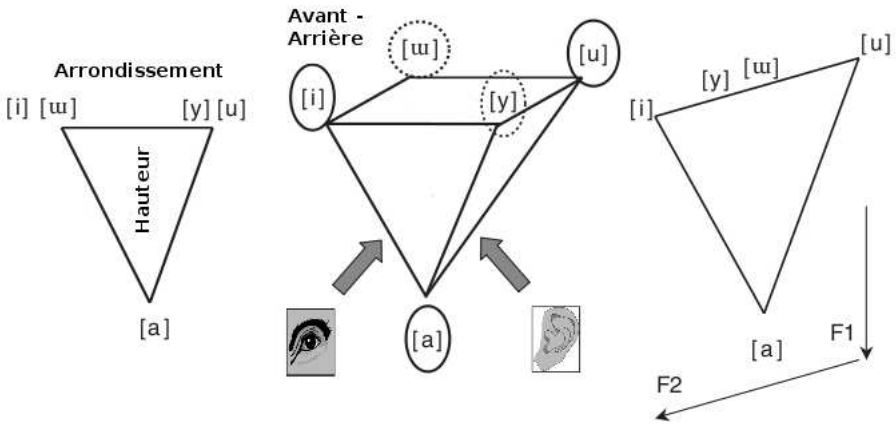


FIGURE 2.5 : Espace articulatoire pour les voyelles [i ɯ a u y] (au milieu) avec sa projection visuelle (à gauche) et auditive (à droite). Figure tirée de Schwartz *et al.* (2007).

exemples). Certains locuteurs produisent les voyelles hautes (ex. [i]) et mi-hautes (ex. [e]) proches les unes des autres et les voyelles mi-hautes loin des voyelles mi-basses (ex. [ɛ]). En revanche, pour certains autres locuteurs [i] et [e] sont très séparées mais les voyelles [e] et [ɛ] sont proches. Malgré ces différences entre les locuteurs, le système des voyelles de chaque locuteur semble suivre un pattern spécifique : les voyelles sont groupées par rapport aux valeurs stables de F1. Par exemple, si un locuteur produit [e] proche de [i], il produit également [o] proche de [u]. Or, il y a une organisation perceptive en miroir, avec des organisations perceptives différentes d'un sujet à l'autre et corrélées entre la perception et la production (Schwartz et Ménard, soumis).

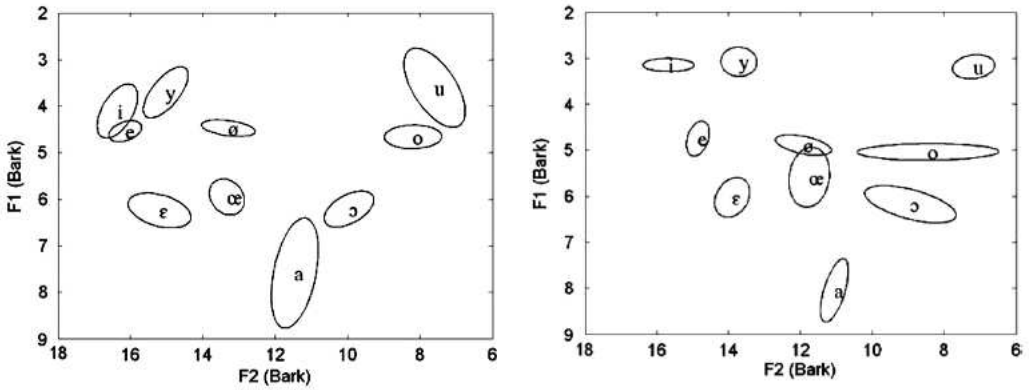


FIGURE 2.6 : L'espace des voyelles [i y a e ø o ɛ œ ɔ a] pour deux locuteurs parmi les participants à l'étude réalisée par Ménard *et al.* (2008). Figure tirée de Ménard *et al.* (2008).

Utilisant le modèle VLAM (*Variable Linear Articulatory Model*) (Maeda, 1979), les auteurs ont montré que les valeurs stables de F1 peuvent être caractérisées par la hauteur de la langue. Cette stratégie ne peut pas être expliquée en termes auditifs

(ex. par la théorie de dispersion adaptative ou l'hypothèse du renforcement auditif, etc.). En revanche, Ménard *et al.* (2008) proposent une explication en termes articulatoires. Ils suggèrent que les locuteurs, au cours du développement, acquièrent un contrôle suffisant de ces articulateurs à l'intérieur d'une série et ils transfèrent ensuite ce contrôle (suffisant et subjectif) à une autre configuration langue/lèvres, ce qui conduit aux séries stables de voyelles mi-hautes et mi-basses avec des hauteurs stables de langue et donc des valeurs stables de F1. Les implications articulatoires au cours du développement sont donc cruciales pour l'organisation du système vocalique de chaque individu. En accord avec la théorie de la perception pour le contrôle de l'action, ces données proposent que le système de perception et de de production de la parole sont structurés l'un par rapport à l'autre. Les objets de parole sont ainsi des gestes mis en forme par leurs valeurs perceptives.

2.2 Modèles anatomiques fonctionnels

La figure 2.7 illustre deux régions latéralisées classiques du traitement du langage chez l'homme : l'aire de Broca (cortex frontal inférieur, BA44/45/47) et l'aire de Wernicke (BA22, située dans la partie postérieure du lobe temporal supérieur). En 1861, l'aire de Broca a été mise en relation par Paul Broca avec la production de la parole. Après une dizaine d'année, l'aire de Wernicke a été mise en relation par Carl Wernicke avec la compréhension de la parole. Depuis, grâce aux techniques de neuroimagerie, le rôle d'autres régions a été mis en évidence dans les tâches liées à la parole et au langage, ce qui a conduit à la proposition de différents modèles anatomiques fonctionnels. Le but de cette section n'est pas de faire une revue exhaustive des études neuropsychologiques sur le traitement de la parole dans le cerveau, mais de présenter des architectures fonctionnelles proposées dans la littérature pour la perception de la parole. Les données récentes peuvent être trouvées dans les revues proposées par Scott et Johnsrude (2003), Demonet *et al.* (2005) et Hickok et Poeppel (2007).

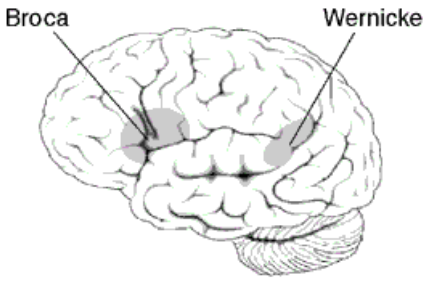


FIGURE 2.7 : L'aire de Broca et l'aire de Wernicke.

Dans cette section, nous nous limitons au modèle Wernicke-Lichtheim-Geschwind, le modèle classique de perception/production de la parole, et à un modèle récent de traitement de la parole proposé par Hickok et Poeppel (2004, 2007). Parmi les modèles anatomiques fonctionnels récents, le modèle proposé par Hickok et Poeppel

est le modèle qui est particulièrement basé sur les questions de la perception de la parole. D'autres modèles anatomiques fonctionnels du traitement du langage sont par exemple, le modèle proposé par Price (2000) et le modèle proposé par Friederici (2002) qui traitent respectivement des questions du traitement lexical et du traitement de la structure linguistique et de la phrase (voir Ben Shalom et Poeppel, 2008, pour une revue et une proposition d'unification de ces différents modèles).

2.2.1 Modèle Wernicke-Lichtheim-Geschwind

Le modèle neuro-anatomique classique de parole est le modèle Wernicke-Lichtheim-Geschwind (figure 2.8, en haut). Selon ce modèle, les zones de production de parole sont autour de l'aire de Broca et les zones de compréhension de la parole sont autour de l'aire de Wernicke. Ces deux aires sont connectées par un réseau de fibres appelé le faisceau arqué. Selon ce modèle, les aires auditives primaires et l'aire de Wernicke sont impliquées dans la perception de parole. L'aire de Broca est impliquée dans la production de la parole.

La figure 2.8, en bas, à gauche, illustre le flux de l'information pour la production d'un mot entendu selon le modèle Wernicke-Lichtheim-Geschwind. Le signal auditif est d'abord analysé dans les aires auditives primaires, puis il est transmis à l'aire de Wernicke. Pour la production, les informations sur la signification du mot sont envoyées grâce au faisceau arqué de l'aire de Wernicke vers l'aire de Broca où sont stockées les représentations articulatoires. Pour lire des mots écrits, outre l'aire de Wernicke et l'aire de Broca, Lichtheim a proposé que l'implication d'une troisième zone est nécessaire, ceci pour conserver les représentations mentales des objets et afin d'associer ces représentations avec les mots. Il a appelé cette zone « centre des concepts » (*concept center*) (Dewart, 1999). Geschwind (1972) a suggéré que cette zone se trouve dans le gyrus angulaire, d'où le nom de modèle Wernicke-Lichtheim-Geschwind. La figure 2.8, en bas, à droite, illustre le flux d'information pour la perception et la production d'un mot lu. Les informations visuelles sur le mot sont d'abord envoyées à partir des aires visuelles vers le gyrus angulaire (centre des concepts), puis sont envoyées vers l'aire de Broca pour la production.

2.2.2 Modèle à deux circuits de Hickok et Poeppel

Prenant en compte des données neuropsychologiques récentes et des études en neuroimagerie, Hickok et Poeppel (2004, 2007) proposent un modèle à deux circuits pour le traitement cérébral de la parole. Selon ce modèle, les processus acoustico-phonétiques de recodage des informations spectro-temporelles du signal acoustique d'entrée vers un code phonologique impliquent la partie dorsale et postérieure du gyrus temporal (STG) et du sulcus temporal supérieur (STS). Cette analyse est réalisée dans les deux hémisphères gauche et droit (Binder *et al.*, 2000). À partir de cette étape, le circuit ventral et le circuit dorsal se séparent. Le circuit ventral comprend la partie intermédiaire du gyrus temporal, le sulcus temporal inférieur et probablement le gyrus temporal inférieur (figure 2.9). Le rôle du circuit ventral est de projeter les représentations sensorielles/phonologiques vers les représentations

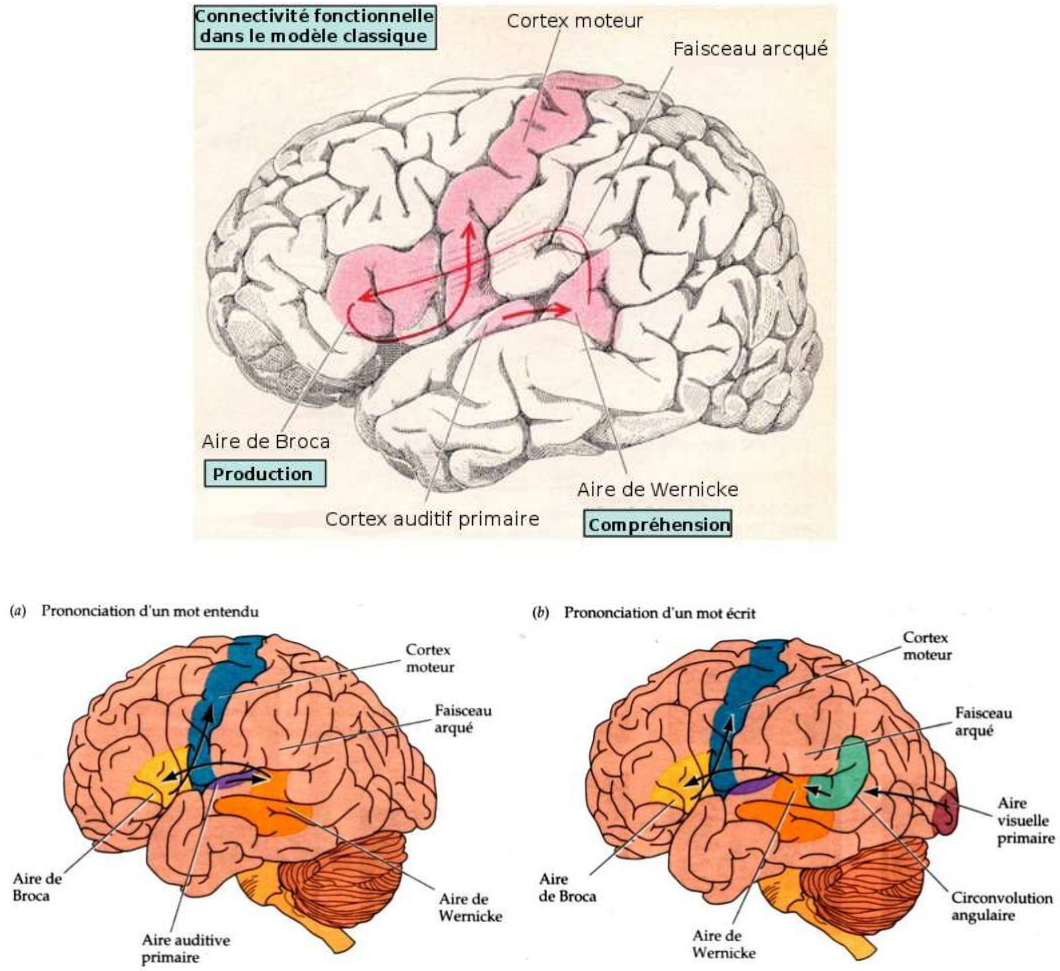


FIGURE 2.8 : En haut : modèle de Wernicke-Lichtheim-Geschwind (figure adaptée de Ben Shalom et Poeppel, 2008). En bas : la connectivité fonctionnelle dans le modèle pour la prononciation d'un mot entendu (a) et d'un mot lu (b) (figure adaptée du site « the brain from top to bottom » à « thebrain.mcgill.ca »).

lexicales ou conceptuelles (i.e. son → sens). Le circuit ventral semble être bilatéral (Hickok et Poeppel, 2007).

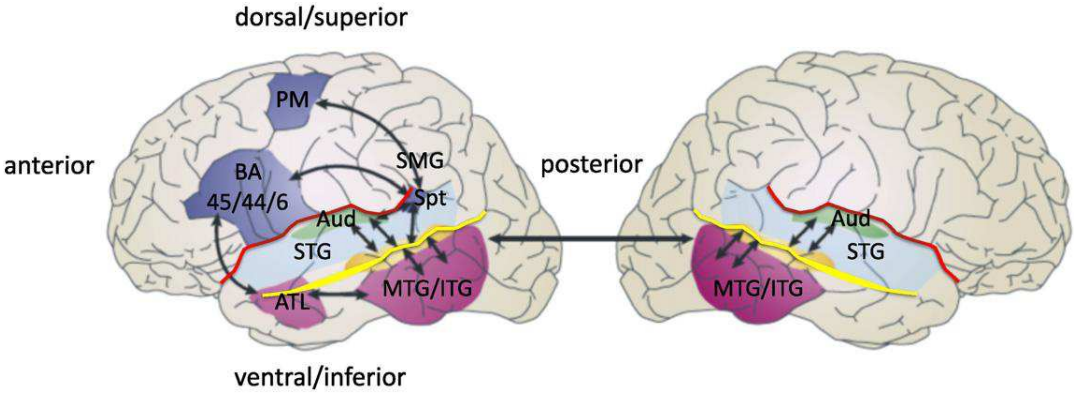


FIGURE 2.9 : Modèle proposé par Hickok et Poeppel (2004, 2007) pour le traitement cortical de la parole. Le circuit ventral, bilatéral et confié au lobe temporal, apparaît dans les deux hémisphères (des régions auditives primaires, en vert, et secondaires, en bleu clair, vers les régions inférieures du lobe temporal, en violet). Le circuit dorsal qui a une dominance gauche (et n'apparaît donc que dans l'hémisphère gauche) comprend des structures dans la jonction pariéto-temporale et le lobe frontal (en bleu foncé). ATL : lobe temporal antérieur ; Aud : cortex auditif (premiers étapes de traitement) ; BA45/44/6 : aires de Brodmann 45, 44 et 6, cortex prémoteur ventral ; MTG/ITG : gyrus temporal médian/gyrus temporal inférieur ; PM : partie dorsale du cortex prémoteur ; SMG : gyrus supramarginal ; Spt : région Sylvienne frontière entre le lobe pariétal et le lobe temporal ; ligne rouge : scissure de Sylvius ; ligne jaune : sulcus temporal supérieur (STS). Figure tirée de Hickok (2009b).

Le circuit dorsal comprend la région postérieure de la scissure de Sylvius à la frontière des lobes pariétal et temporal (aire Spt), le gyrus frontal inférieur (IFG), l'insula antérieure et le cortex prémoteur. Ce circuit fait une projection des représentations sensorielles et phonologiques vers les représentations articulatoires (i.e. son → articulation). Contrairement au circuit ventral bilatéral, le circuit dorsal est largement spécifique à l'hémisphère gauche. Le flux d'information dans ce circuit est le suivant : le codage sensoriel se réalise d'une façon bilatérale dans le STS et la traduction entre les représentations sensorielles et les représentations motrices de l'IFG gauche s'effectue par l'aire Spt gauche. Hickok et Poeppel (2007) soulignent le rôle essentiel du circuit dorsal de l'intégration auditivo-motrice dans l'acquisition du langage au cours du développement. Le circuit dorsal continue également d'être impliqué chez l'adulte. Ainsi, le phénomène d'adaptation sensori-motrice peut refléter l'activation du circuit dorsal même chez l'adulte : lors de la production des séquences de parole, un feedback auditif modifié influence la production pour compenser ce feedback modifié (Houde et Jordan, 2002). Le circuit dorsal serait également impliqué pour l'acquisition des nouveaux mots. Néanmoins, pour l'essentiel, ce circuit n'est, selon Hickok et Poeppel, pas requis lors de la compréhension en ligne.

2.2.3 Le rôle fonctionnel du système moteur dans la perception de la parole : éléments d'un débat

Nous avons vu dans cette section différents types d'hypothèses sur le lien entre le système moteur et le système auditif dans la perception de la parole. Alors que les théories motrices proposent une connexion directe entre l'entrée sensorielle et les représentations articulatoires, le modèle de Hickok et Poeppel suggère que cette connexion est seulement présente dans le circuit dorsal et pas dans le circuit ventral de la compréhension de la parole. Malgré les données comportementales et neuro-anatomiques démontrant l'existence des connexions perceptuo-motrices dans les tâches de la perception de la parole, le rôle fonctionnel de ces connexions fait l'objet de débats dans la littérature. Hickok et Poeppel font remarquer que les lésions du système moteur, par exemple dans le cas de l'aphasie de Broca, ne conduisent pas à des déficits significatifs de compréhension de la parole, traitée, selon leur modèle, par le circuit ventral de la compréhension de la parole (Hickok et Poeppel, 2007).

Quelques études récentes utilisant la méthode TMS ont montré que la perturbation du système moteur entraîne une modification de performance dans les tâches de perception de la parole. Meister *et al.* (2007) ont perturbé le cortex prémoteur gauche des sujets pendant une tâche d'identification des syllabes /pa/, /ba/ et /ta/ dans le bruit. Ils ont observé une baisse de taux de réponses correctes chez les sujets lors de la stimulation par rapport à la condition de contrôle. Les auteurs ont donc conclu que le cortex prémoteur jouerait un rôle « essentiel » (leur propre formulation) dans la perception de la parole. Cependant, la diminution de score est en réalité très faible (de l'ordre de 8% de réponses correctes). Il est également proposé que le cortex prémoteur gauche a un rôle fonctionnel dans la segmentation phonémique. Utilisant des stimuli non-bruités de type CVC (C : consonne, V : voyelle), Sato *et al.* (2009) ont observé que les stimulation TMS du cortex prémoteur gauche entraînent un temps de réaction plus grand lors de la discrimination phonémique qui demande une segmentation, mais pas pour les conditions d'identification phonémique ou discrimination syllabique.

D'Ausilio *et al.* (2009) ont étudié le rôle du système moteur dans une expérience de discrimination entre des phonèmes produits par la langue ([d] et [t]) et produits par les lèvres ([b] et [p]) en présence du bruit. Ils ont stimulé les aires liées aux représentations motrices de la langue ou des lèvres dans le cortex moteur des sujets avant la réalisation de tâche. Les auteurs ont observé que les stimulations des aires motrices en lien avec la langue facilitent l'identification des phonèmes [d] et [t] et inhibent l'identification des phonèmes [b] et [p] liés aux lèvres. Un pattern inverse a été observé lors de stimulation des aires motrices en lien avec les lèvres, ce qui mettrait en évidence le rôle fonctionnel du système moteur dans l'identification phonémique. Une critique envers ces études concerne la nature des tâches demandées aux sujets. Hickok (2009c) constate que ces études utilisent majoritairement les tâches métaphonologiques de discrimination ou d'identification, phonémique ou syllabique, qui impliquent la présence d'autres processus que ceux utilisés dans la compréhension et la reconnaissance normale des signaux de parole (par exemple l'implication de la mémoire de travail verbale).

En ce qui concerne le rôle des liens perceptuo-moteurs dans la perception de la parole, Skipper et collègues proposent, dans le cadre d'un modèle de l'intégration audio-visuelle de la parole, que ces liens interviennent pour prédire les entrées (multisensorielles) et pour moduler et contraindre ensuite notre interprétation de ces entrées (Skipper *et al.*, 2005; van Wassenhove *et al.*, 2005; Skipper *et al.*, 2007). Ce modèle sera présenté en détail dans la sous-section 2.3.5. Selon ce modèle, le rôle des liens perceptuo-moteurs serait spécialement important lorsque les informations sensorielles ne sont pas adéquates par exemple en présence du bruit ou dans la perception des langues étrangères. Cette prédiction est cohérente avec les études récentes qui montrent l'activation renforcée des régions frontales pendant la perception de la parole dans ce type de conditions (ex. Binder *et al.* (2004) dans le bruit ou Callan *et al.* (2004) pour les langues étrangères).

Il est proposé que les liens entre le système sensoriel et le système moteur peuvent être implémentés par les cartes sensori-motrices (Schwartz *et al.*, sous presse). Dans une expérience sur la discrimination du degré d'ouverture de la mâchoire, de la position de la langue ou de l'arrondissement des lèvres sur des voyelles synthétiques, Schwartz *et al.* (2008b) ont observé que les participants ont accès aux caractéristiques articulatoires des voyelles lors de la perception, ce qui peut être expliqué par l'existence de cartes sensori-motrices des voyelles. Les études IRMf réalisée par Obleser *et al.* (2006) et Pulvermüller *et al.* (2006) sur les cartes perceptives ou articulatoires des voyelles ou des consonnes mettent en évidence l'existence de ce type de cartes. Le résultat de D'Ausilio *et al.* (2009) présenté ci-dessus propose également que les cartes motrices de type effecteur↔son seraient à la base des liens sensorie-moteurs.

Le rôle fonctionnel du système moteur dans la perception de la parole est donc une question ouverte. Scott *et al.* (2009) proposent que l'implication du système moteur n'est pas essentielle pour la perception de la parole mais est essentielle lors d'une conversation. Dans une conversation, les interlocuteurs co-ordonnent leur respiration et la prononciation présente des éléments de convergence entre partenaires. De plus, les participants de la conversation prennent la parole la plupart du temps sans qu'il y ait une pause ou un chevauchement. Les auteurs expliquent ces phénomènes de convergence et de prise de parole coordonnée par l'implication du système moteur : lorsque les régions temporales de la voie ventrale cherchent à comprendre ce qui est dit dans une conversation, le système moteur stimulé par la voie dorsale s'adapterait à la vitesse d'élocution et au rythme du locuteur afin que l'auditeur puisse prendre la parole d'une manière non-brusque (figure 2.10). Ainsi, Scott *et al.* (2009) proposent que les représentations motrices et le système moteur joueraient un rôle essentiel dans la communication orale, ce qui permet de communiquer l'un avec l'autre sans à-coups même quand on est dans une conversation sans le support visuel (par exemple, au téléphone).

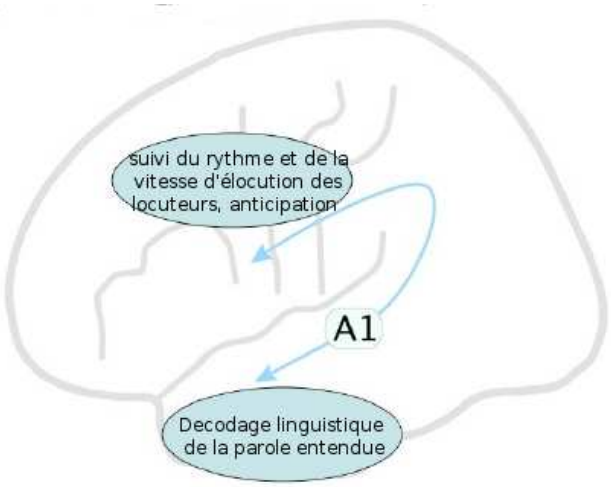


FIGURE 2.10 : Le rôle du système moteur dans une conversation, proposé par Scott *et al.* (2009). A1 : cortex auditif primaire. Le flux du signal vers les parties antérieures des aires temporales supérieures correspond au processus de compréhension dans la voie ventrale (la transformation du son vers le sens). Le flux vers le gyrus frontal inférieur en passant par la partie postérieure du gyrus temporal supérieur, le cortex pariétal inférieur et les aires motrices et sensorielles correspondent à la transformation du son dans la voie dorsale, vers les caractéristiques permettant la coordination des partenaires d'une conversation. Figure adaptée de Scott *et al.* (2009).

2.3 La multisensorialité des objets parole

Nous ne percevons pas la parole uniquement à partir du signal sonore : les informations visuelles et tactiles sont aussi utilisées pendant la perception de la parole. La modalité visuelle intervient essentiellement quand la tâche de perception/compréhension est difficile par exemple en présence du bruit (Sumby et Pollack, 1954; Benoit *et al.*, 1994). En ce qui concerne la perception tactile de la parole, la méthode Tadoma met en évidence l'apport du toucher dans une communication langagière chez les personnes sourdes-aveugles. Cette méthode est basée sur la réception des mouvements articulatoires réalisés pendant la production de la parole. La main du récepteur doit être placée sur le visage du locuteur de sorte que le pouce soit sur les lèvres et les autres doigts soient sur la joue et le cou (voir figure 2.11). Les sourds-aveugles pratiquant la méthode Tadoma peuvent acquérir les compétences pour la perception et la production de la parole (voir Reed, 1996).

Cette section porte sur l'aspect audio-visuel de l'objet parole. Une brève revue de la littérature sur l'intégration audio-visuelle pendant la perception de la parole sera présentée dans la suite. Il est à noter que la lecture labiale pure ne fait pas directement l'objet de cette section.

FIGURE 2.11 : Méthode Tadoma de la perception de la parole par la voix tactile.

2.3.1 Les objets parole audio-visuels

Plusieurs expériences comportementales sur la parole audio-visuelle ont montré que les informations visuelles augmentent l'intelligibilité de la parole dans un environnement bruité. Sumby et Pollack (1954) ont observé que la présentation de la modalité visuelle en plus de la modalité auditive était équivalente à une amélioration du rapport signal sur bruit de 15 dB. Une série d'études ont été réalisées par Summerfield et collègues pour mettre en évidence le rôle de différentes composantes du visage sur la perception de la parole (Summerfield, 1979; MacLeod et Summerfield, 1987). Ils estiment un gain de 11 dB apporté par la vision des lèvres sur l'intelligibilité de la parole. Dans une autre étude, ils ont également observé la contribution des dents dans la discrimination des voyelles (Summerfield *et al.*, 1989). L'amélioration de l'intelligibilité de la parole par la vision semble être obtenue grâce à la complémentarité des informations fournies par la modalité auditive et la modalité visuelle. En effet, Benoit *et al.* (1994) ont observé que l'intelligibilité de la voyelle [a] est plus importante que celle de la voyelle [i] et que celle de la voyelle [y] en modalité auditive alors qu'en modalité visuelle seule, celle de [y] est plus importante que celle de [a] et [i].

La modalité visuelle est bénéfique même lorsque le signal auditif n'est pas bruité. Reisberg *et al.* (1987) ont demandé à des sujets de répéter le plus rapidement possible un texte difficile à comprendre (un passage du « Critique de la raison pure » de Kant) en présentant en condition audio seule et audio-visuelle. Ils ont observé une amélioration de temps de réaction en condition audio-visuelle par rapport à la condition audio seule bien que le texte soit parfaitement audible et présenté sans bruit. Dans une série d'expériences sur la détection auditive de phrases, Grant et Seitz (2000) ont observé que le seuil de détection pouvait être réduit de 1.6 dB en moyenne lors de la présence du signal visuel cohérent avec le stimulus auditif.

Outre les expériences sur l'influence de la modalité visuelle sur l'intelligibilité

de la parole, les expériences de type McGurk mettent en évidence l'interaction des informations auditives et visuelles en perception de la parole. En 1976, McGurk et MacDonald ont reporté une série d'expériences utilisant les stimuli audio-visuels non-cohérents qui mettent en évidence l'interaction entre la modalité auditive et la modalité visuelle. Par exemple, lorsque la séquence /ga/ visuelle est présentée en concordance avec la séquence /ba/ auditive, la séquence sera souvent perçue /da/, qui n'est pas pourtant présent dans le flux auditif ni dans le flux visuel. Les informations auditives et visuelles semblent donc être intégrées lors de la perception de la parole. Cette intégration a été aussi observée au niveau lexical où les informations auditives et visuelles conflictuelles conduisent à la perception d'un troisième mot différent du mot auditif et du mot visuel (Dodd, 1977).

2.3.2 Nature et spécificité des informations visuelles

Si ces expériences montrent clairement que les modalités auditive et visuelle interagissent dans la perception de la parole, elles ne nous renseignent pas directement sur la nature des indices visuels utilisés dans la fusion. Un certain nombre de travaux nous éclairent en partie sur cette question. D'abord, les expériences de Summerfield (1979) montrent d'une part le rôle des lèvres, de la langue et des dents sur la perception de la parole audio-visuelle, mais aussi que le remplacement de la dynamique labiale par le mouvement d'une forme géométrique (elliptique) de même dynamique temporelle ne fournit pas de gain d'intelligibilité, au contraire du mouvement labial réaliste d'origine. De même, dans une expérience sur la détection auditive des syllabes dans le bruit, Bernstein *et al.* (2004) ont observé que la vision des mouvements des lèvres augmente la détection auditive de stimuli acoustiques cohérents avec les mouvements labiaux, mais pas la vision d'une forme simple animée de la même dynamique. Schwartz *et al.* (2004) ont fait la même observation dans une tâche d'identification de stimuli de type [Cy] où C était une plosive voisée ou non voisée : la vision du mouvement des lèvres, qui ne contenait pas intrinsèquement d'information sur le voisement consonantique, augmentait néanmoins l'identification, en permettant aux sujets de connaître la zone temporelle où pouvait se situer la barre de voisement ; mais un signal visuel différent, fournissant la même information temporelle, ne produisait pas de gain d'intelligibilité. Ainsi, il semble bien que le caractère réaliste du stimulus visuel, permettant de l'interpréter effectivement comme de la parole, soit nécessaire à une bonne fusion. De même, en enregistrant des activités neuronales unitaires (*unit activity*) et des champs de potentiels locaux (*local field potential*), Ghazanfar *et al.* (2005) ont également observé chez les singes qu'un disque simulant les mouvements des lèvres ne peut pas produire le même effet que lors de la vision réelle du mouvement des lèvres.

Par contre, Jordan et Sergeant (2000) ont obtenu un effet McGurk dans une tâche où les lèvres n'étaient pas parfaitement visibles en raison de la distance importante entre les participants et le stimulus. L'enjeu n'est donc pas la visibilité de l'information labiale, mais bien sa qualité « parole ». Cette information est précisée par d'autres travaux, montrant l'importance de la cinématique du visage. Des points lumineux ayant la même cinématique qu'un visage, en absence d'autres indices fa-

ciaux, peuvent augmenter l'intelligibilité de la parole bruitée (Rosenblum *et al.*, 1996) et produire l'effet McGurk (bien que l'effet soit moins important qu'avec le vrai visage) (Rosenblum et Saldaña, 1996).

Pour expliquer les différentes données sur les indices visuels utiles dans la perception de la parole, Campbell (2008) propose deux modes par lesquels la modalité visuelle influence la perception de la parole : mode complémentaire (*complementary mode*) et mode corrélé (*correlated mode*). Le mode complémentaire sert à désambiguïser la parole quand les informations auditives ne sont pas assez claires (par exemple : [m] et [n] visuel sont facilement distinguables tandis que leur signaux acoustiques sont très similaires). Pour ce mode, la visibilité de la partie basse du visage (ex. lèvres, langue, etc.) serait nécessaire. Pour le mode corrélé, l'information importante concerne les indices spectro-temporels qui peuvent révéler les parties ayant une dynamique similaire dans le signal auditif et le signal visuel. C'est dans ce mode que la visibilité des mouvements de visage serait nécessaire (et peut-être suffisante).

2.3.3 Les modèles de fusion audio-visuelle

Robert-Ribes *et al.* (1995) et Schwartz *et al.* (1998) proposent quatre architectures possibles pour la fusion entre les modalités auditive et visuelle illustrées sur la figure 2.12.

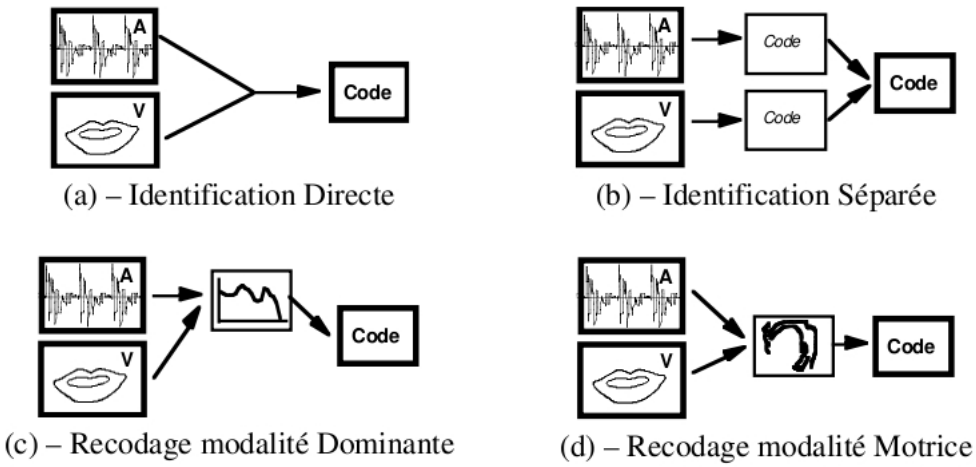


FIGURE 2.12 : Modèles de fusion audio-visuelle en perception de la parole (Robert-Ribes *et al.*, 1995; Schwartz *et al.*, 1998). Figure tirée de Schwartz (2004).

- Modèle à « Identification directe » (ID) : la classification se fait directement sur les flux auditifs et visuels sans aucune mise en forme commune des données. La décision est basée sur les règles bimodales apprises par le système.
- Modèle à « Identification séparée » (IS) : la classification du flux auditif et

du flux visuel se fait séparément. La fusion s'effectue par une intégration de décisions prises pour chaque modalité (fusion tardive).

- Modèle à « Recodage commun dans la modalité dominante » (RD) : les informations visuelles sont codées d'abord sous un format compatible avec la représentation en modalité auditive qui est considérée comme la modalité dominante.

- Modèle à « Recodage commun dans la modalité motrice » (RM) : inspiré par la théorie motrice et la théorie de la perception réaliste directe (voir section 2.1.1), ce modèle propose que les informations auditives et visuelles sont codées sous une nouvelle représentation commune de nature articulatoire qui est ensuite fournie à un processus de classification.

Parmi les architectures présentées ci-dessus, les modèles ID et IS sont ceux qui sont plus fréquemment utilisés dans la reconnaissance de la parole (Schwartz, 2004). En comparant ces quatre architectures pour une tâche de reconnaissance des voyelles en français, Teissier *et al.* (1999) ont conclu que les modèles ID et IS sont plus performants que les modèles RD et RM. En revanche, Schwartz (2004) argumente que les résultats expérimentaux à l'issue des études en psychologie expérimentale sont plus en faveur du modèle RM que des autres modèles. Nous reviendrons sur ce point dans la section suivante en présentant quelques appuis neuronaux en faveur du modèle RM.

Outre le choix de l'architecture d'intégration audio-visuelle, le choix de l'opérateur de fusion est également essentiel pour un modèle de fusion audio-visuelle. Bloch (1994) distingue trois types d'opérateurs de fusion qui diffèrent selon leur comportement.

- Opérateurs indépendants du contexte et à comportement constant (ICCC). Ce type d'opérateurs effectue une intégration entre le flux auditif et le flux visuel en appliquant une lois fixe indépendamment du contexte par exemple l'addition ou la multiplication des informations correspondant à chaque modalité.

- Opérateurs indépendants du contexte et à comportement variable (ICCV). La loi de fusion dans ce type d'opérateurs est variable et dépend des valeurs d'entrée (et pas du contexte) par exemple une fusion additive pondérée avec des coefficients de pondération dépendant du niveau des entrées.

- Opérateurs dépendants du contexte (DC). Ces opérateurs prennent en compte des connaissances sur l'environnement extérieur, qui peuvent lui permettre par exemple de pondérer plus ou moins une entrée selon la nature de ces informations contextuelles (ex. lors de la présence du bruit dans le signal auditif ou visuel).

Le modèle FLMP (*Fuzzy Logical Model of Perception*) (Oden et Massaro, 1978; Massaro, 1998), un des modèles les plus connus dans la littérature de l'intégration

audio-visuelle, est défini par une architecture de type IS qui utilise des opérateurs de fusion de type ICCC. Ce modèle effectue une estimation séparée des classes phonétiques dans chaque modalité. La fusion se fait ensuite par une multiplication des vraisemblances auditives et visuelles.

2.3.4 De la fusion au liage audio-visuel

L'introduction d'un mécanisme de contrôle de la fusion par un « contexte » ouvre sur une question importante, qui est celle de l'automaticité de la fusion. Dans la revue de questions sur les modèles de fusion audio-visuelle en perception de parole, Schwartz *et al.* (1998) décrivent deux types de contexte : les contextes « indépendant du stimulus » et « dépendant du stimulus ». Dans le premier type, on trouve d'abord tous les facteurs de variabilité inter-sujets, et notamment les différences inter-culturelles et linguistiques (la possibilité que certaines cultures attachent moins d'importance à la modalité visuelle : voir Sekiyama et Tohkura, 1993; Burnham, 1998) ; ainsi que les différences inter-individuelles pures (la possibilité que certains sujets soient plus orientés vers une modalité que vers une autre). On trouvera dans Schwartz (2010) une mise en évidence mathématique de ces différences inter-individuelles pures dans le cadre du modèle FLMP, en allant vers un modèle « Weighted FLMP » (WFLMP) où les modalités sont pondérées par des facteurs dépendant du sujet.

On trouve ensuite, toujours dans les facteurs indépendants du stimulus, l'ensemble des mécanismes attentionnels. La fusion audio-visuelle a été historiquement considéré comme un processus automatique et pré-attentif (voir Soto-Faraco *et al.*, 2004). Cependant, selon certaines études utilisant des distracteurs visuels ou auditifs, la fusion audio-visuelle serait sensible aux mécanismes attentionnels. Dans une expérience de type McGurk, Tiippana *et al.* (2004) ont ajouté au stimulus visuel une feuille parcourant le visage de la locutrice jouant le rôle d'un distracteur visuel. En comparant cette condition avec la condition de base (sans distracteur), ils ont observé que l'effet McGurk diminue lors de la présence du distracteur visuel. Ainsi, les mécanismes attentionnels joueraient un rôle dans la fusion ou non des informations auditives et visuelles. Alsius *et al.* (2005) ont également observé que la fusion audio-visuelle est moins efficace en présence d'un distracteur visuel. Un distracteur auditif diminue également l'effet McGurk. Cette baisse due au distracteur auditif est accompagnée par l'augmentation des percepts cohérents avec le flux auditif. Ce dernier résultat propose que la charge attentionnelle supplémentaire influence la fusion audio-visuelle et non les étapes précédant la fusion.

Ces expériences modifient bien sûr le contenu de la situation expérimentale, mais sans changer la stimulation elle-même : c'est par l'ajout de stimuli ou de tâches distracteurs que la fusion est perturbée. Par contre, ce que Schwartz *et al.* (1998) nomment « effet dépendant du stimulus » réfère à des situations contrôlées par le stimulus lui-même. C'est là qu'interviennent des mécanismes de liage au sens où nous les avons rencontrés dans le chapitre 1. Ainsi, dans l'expérience de Schwartz *et al.* (2004) décrite précédemment, la stimulation visuelle peut guider le processus d'extraction de l'information auditive adéquate. Récemment, Nohorna (2009) est

allée plus loin dans cette voie, en proposant un paradigme permettant de distinguer les mécanismes de liage audio-visuel et ceux de la fusion dans une expérience sur l'effet McGurk. Ses travaux montrent qu'un contexte préalable cohérent (présentation d'une série de syllabes audio-visuelles congruentes) permet la fusion audio-visuelle de la syllabe auditive /ba/ et de la syllabe visuelle /ga/, ce qui entraîne la perception de /da/ (effet McGurk). En revanche, l'effet McGurk diminue d'une manière importante lorsque les sujets étaient préalablement exposés à des signaux audio-visuels incohérents où une phrase visuelle quelconque était montée sur une série de syllabes auditives. Cette dernière condition favoriserait un « décrochage » du lien audio-visuel, ce qui pourrait conduire à la baisse ou la suppression de fusion audio-visuelle.

Ces expériences mettent bien en évidence un niveau de liage préalable à la fusion, qui en conditionnerait l'existence : ce mécanisme de liage audio-visuel permet de regrouper les éléments d'information pertinents, faisant ainsi émerger dans les scènes audio-visuelles des « objets paroles audio-visuels » cohérents.

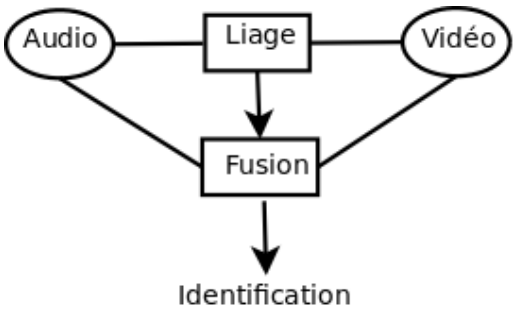


FIGURE 2.13 : Les mécanismes de liage audio-visuel permettent la fusion ou non de la modalité auditive et visuelle de la parole.

Ce mécanisme de liage et de constitution des objets audio-visuels de la parole est apparemment « spécifique », au sens de Remez *et al.* (1994) (voir section 1.3). En effet, si l'effet McGurk semble dépendre de mécanismes attentionnels et de principes de liage conditionnel dépendant du contexte, il est par contre robuste à des incohérences non phonétiques multiples : incohérence de localisation spatiale entre sources auditive et visuelle (Bertelson *et al.*, 1994), asynchronie entre les sources dans un domaine assez large, jusqu'à 200 ms (McGrath et Summerfield, 1985; van Wassenhove *et al.*, 2007), et même incohérence sur l'identité de la source, avec un visage de femme superposé sur une voix d'homme (Green *et al.*, 1991).

2.3.5 Mécanismes cérébraux de la fusion audio-visuelle

Le boucle temporo-occipitale

Des études récentes sur les mécanismes cérébraux de la perception audio-visuelle de la parole proposent que l'intégration audio-visuelle se réalise dans la partie postérieure du sulcus temporal supérieur (*posterior superior temporal suclus*, pSTS) (ex. Calvert *et al.*, 2000; Wright *et al.*, 2003). Dans une étude IRMf, Calvert *et al.* (2000)

ont observé un effet supra-additif dans le pSTS lors de la présentation des stimuli en modalité audio-visuelle congruente (AV > A + V) et un effet sub-additif lorsque les stimuli étaient incongruents (AV < A + V) (mais voir aussi Hocking et Price, 2008, pour une critique). Calvert *et al.* (2000) proposent que le pSTS joue un rôle clé dans l'intégration des informations audio-visuelles. Les auteurs ont également observé une activité plus importante dans le cortex auditif primaire et le cortex occipital en modalité bimodale par rapport à une présentation unimodale. Selon les auteurs, cette amélioration reflète une rétroprojection (*feedback*) de l'activité correspondant à l'intégration bimodale vers les zones unimodales auditives et visuelles, l'activité au sein de la région STS modulant ensuite les activités des cortex sensoriels correspondants. L'interaction entre les régions STS de l'intégration audio-visuelle et le cortex auditif a été également observée chez les singes (Ghazanfar *et al.*, 2008).

Ces mécanismes d'interaction audiovisuelle avec modulation de la réponse du cortex auditif produisent des effets visibles également en EEG, et se traduisant par des modifications précoces des potentiels évoqués auditifs, avec notamment une accélération et une diminution d'amplitude du pic de réponse à la stimulation auditive (Besle *et al.*, 2004; Colin *et al.*, 2002). Ces effets semblent là encore spécifiques de la cohérence audio-visuelle des signaux de parole, spécificité qui est en réalité une spécificité écologique, liée à la cohérence des mécanismes de production des signaux sonores et visuels. Ainsi Stekelenburg et Vroomen (2007) montrent que les effets de modulation de la réponse électrophysiologique du cortex auditif à un signal auditif par une entrée visuelle conjointe est produite par des stimuli non phonétiques mais organisés par des principes semblables, tels que des claquements de mains. Dans ce cas, ce qui compte est la capacité du signal visuel à prédire le début du signal auditif. Au contraire, dans le cas de stimuli non prédictibles, l'entrée visuelle ne produit pas de modulation électrophysiologique de la réponse du cortex auditif au stimulus sonore.

L'étude réalisée par Bernstein *et al.* (2008) ajoute un autre élément aux propositions ci-dessus. Dans cette étude IRMf, ils ont présenté aux participants trois types de stimuli audio-visuels ayant différents degrés d'incongruité audio-visuelle : LI (*Low Incongruity*) où les stimuli auditif et visuel étaient congruents, MI (*Medium Incongruity*) où les stimuli auditif et visuel étaient moyennement incongruents et HI (*High Incongruity*) où les stimuli auditif et visuel étaient fortement incongruents. Les stimuli ayant une incongruité basse étaient $A_{ba}V_{ba}$ et $A_{la}V_{la}$. Les stimuli avec une incongruité moyenne étaient $A_{ba}V_{da}$, $A_{la}V_{va}$, $A_{ba}V_{ga}$ et $A_{la}V_{wa}$ et ceux très incongruents étaient $A_{ba}V_{va}$, $A_{la}V_{ba}$, $A_{ba}V_{wa}$ et $A_{la}V_{da}$. Ces stimuli entraînent différents types de percepts : le percept cohérent avec la modalité auditive, le percept cohérent avec la modalité visuelle, combinaison des deux consonnes présentes dans la modalité auditive et visuelle (ex. /gi/ auditif + /bi/ visuel → /bgi/ audio-visuel) ou le percept de type McGurk. Dans cette étude, le gyrus supramarginal (SMG) gauche était la seule région qui a montré des activités différentes en fonction du degré d'incongruité des informations auditives et visuelles (HI>MI>LI). Les auteurs suggèrent que le SMG jouerait un rôle dans l'analyse fine de la relation entre les entrées phonétiques auditive et visuelle. Cette analyse pourrait être implémentée par une comparaison entre les représentations perceptives auditives et visuelles et la

connaissance mémorisée d'une relation normale (congruente) des patterns auditifs et visuels.

La boucle temporo-pariéto-frontale

Contrairement au modèle présenté ci-dessus, certaines études proposent que l'intégration entre les informations auditives et visuelles en parole se réalise par le biais des interactions sensori-motrices. Selon Skipper *et al.* (2007), les mouvements articulatoires visibles des articulateurs pourraient activer chez l'auditeur, grâce au système miroir, un plan moteur qu'il aurait pu utiliser comme locuteur pour produire ces mêmes mouvements observés (Skipper *et al.*, 2005). En accord avec cette proposition, Sams *et al.* (2005) ont observé que l'effet McGurk peut également se produire lorsque les sujets articulaient silencieusement des syllabes et regardaient leur visage dans un miroir. Ils ont observé un effet moins fort lors de l'articulation silencieuse sans le feedback visuel. Plusieurs études en neuroimagerie ont mis en évidence le rôle des régions impliquées dans la production de la parole, surtout le gyrus frontal inférieur, lors de la perception audio-visuelle de la parole (ex. Callan *et al.*, 2003; Ojanen *et al.*, 2005). Ces données sont cohérentes avec le modèle selon lequel l'intégration audio-visuelle se réalise par l'intervention du système moteur.

Skipper et collègues proposent un modèle basé sur « l'analyse-par-synthèse » pour la perception audio-visuelle de la parole (Skipper *et al.*, 2005; van Wassenhove *et al.*, 2005; Skipper *et al.*, 2007). Ce modèle suggère que les représentations précoces de parole tirées du son et des mouvements faciaux du locuteur sont plutôt des hypothèses multisensorielles et pas les interprétations finales concernant les phonèmes produits par le locuteur. Ces hypothèses sont projetées sur l'espace des commandes motrices utilisées dans la production de parole. Les commandes activées effectuent ensuite une prédiction sur les conséquences acoustiques et somato-sensorielles de la réalisation de ces mouvements par la copie d'efférence, ce qui peut ensuite moduler l'analyse et l'interprétation phonétique de l'information sensorielle d'entrée.

La figure 2.14 illustre les aires corticales proposées correspondant au modèle de Skipper et collègues. Ces aires concernent les régions visuelles, le cortex auditif primaire (A1), les parties supérieures postérieures des régions temporales (STp), le gyrus supramarginal (SMG), les cortex somato-sensoriels (SI/SIII), le cortex prémoteur ventral (PMv) et pars opercularies (POp). Le processus de perception audio-visuelle commencerait par une hypothèse dans les aires STp (aire visuelle $\rightarrow$ STp $\leftarrow$ A1). Cette hypothèse serait spécifiée en terme de cible motrice (*motor goal*) du mouvement observé dans le POp, région considérée comme équivalente de F5 chez le macaque où les neurones miroirs ont été trouvés (STp $\rightarrow$ POp). Cette cible motrice serait ensuite projetée vers les commandes motrices qui pourraient avoir généré les mouvements observés dans le cortex PMv (POp $\rightarrow$ PMv $\leftarrow$ M1). Ces commandes motrices conduiraient à une prédiction sur les conséquences auditives (PMv $\rightarrow$ STp) et somatosensorieles (PMv $\rightarrow$ SI/SII $\rightarrow$ SMG $\rightarrow$ STp) de ces commandes. Ces prédictions seraient finalement utilisées pour favoriser une interprétation/hypothèse dans STp.

Considérant l'activation conjointe des régions pSTS et des régions frontales in-

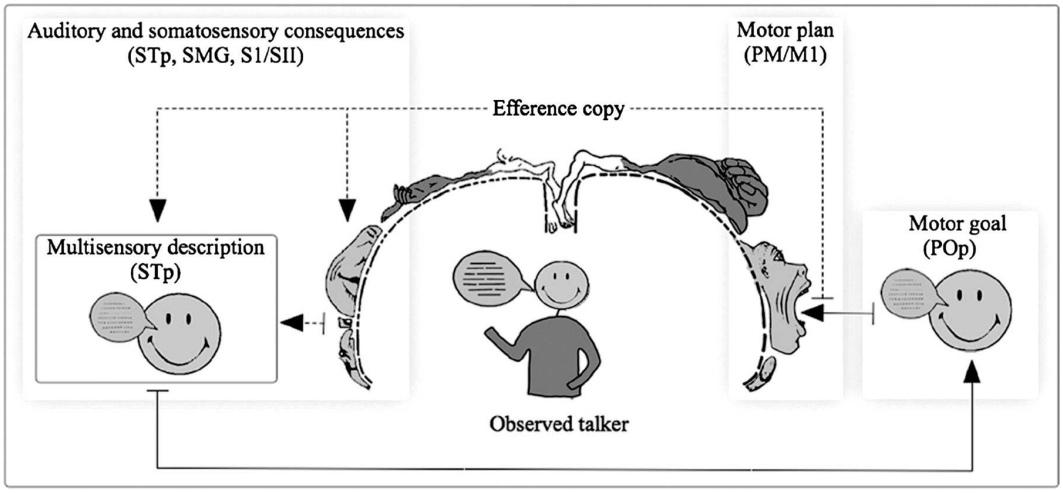


FIGURE 2.14 : Mécanismes cérébraux proposés par Skipper *et al.* (2007) pour la perception audio-visuelle de la parole. Figure tirée de Skipper *et al.* (2007).

férieures pour la perception audio-visuelle de la parole, ces régions semblent refléter différents types d'analyse sur les informations audio-visuelles. Par exemple, Miller et D'Esposito (2005) proposent deux processus distincts d'intégration audio-visuelle, l'un pour la comparaison entre les informations auditives et visuelles du stimulus et l'autre pour la perception des informations audio-visuelles unifiées. Ils ont étudié ces processus dans une expérience IRMf en présentant aux sujets les stimuli audio-visuels synchrones ou avec un offset, ce qui conduisait respectivement à une fusion audio-visuelle ou à une perception successive des événements auditifs et visuels. Ils ont observé une activité plus importante dans le STS et le gyrus de Heschl lors de la fusion des flux auditif et visuel. Dans le sulcus intrapariétal (IPS), ils ont observé une augmentation d'activité pour la fusion et une baisse lorsque les deux modalités ne sont pas fusionnées. Le gyrus frontal inférieur montre une tendance inverse, i.e. son activité baisse lors de la fusion audio-visuelle des stimuli et augmente quand les modalités auditive et visuelle ne sont pas fusionnées.

En résumé, les résultats présentés ci-dessus suggèrent que le réseau de l'intégration audio-visuelle comprend les régions Heschl, STS, IPS et IFG. Ces régions semblent faire partie du réseau de l'intégration sensori-motrice pour les tâches liées à la parole (Hickok *et al.*, 2003; Buchsbaum *et al.*, 2005). Dans une expérience IRMf, Okada et Hickok (2009) étudient l'implication ou non des aires de l'intégration sensori-motrice dans la perception de la parole visuelle. Les auteurs ont observé une activation latéralisée gauche dans le gyrus frontal inférieur, le cortex prémoteur dorsal, la scissure de Sylvius dans la frontière des aires temporale et pariétale (Spt) et dans le sulcus temporal supérieur à la fois dans une tâche de lecture labiale et une tâche sensori-motrice. L'activité correspondant à la tâche sensori-motrice était calculée par la soustraction suivante : écoute→répétition − écoute→repos. Ce résultat est cohérent avec le modèle de Skipper *et al.* présenté ci-dessus. Il est important de noter que la région Spt semble être impliquée à la fois dans les tâches sensorielles

et dans les tâches motrices dans le domaine de la parole par exemple pendant la lecture et la perception des pseudo-mots (Buchsbaum *et al.*, 2005). Une étude IRMf réalisée par Pa et Hickok (2008) suggère que Spt jouerait un rôle d'interface entre le système auditif et l'effecteur vocal (en comparaison avec la partie antérieure de l'IPS qui ferait l'interface entre l'effecteur manuel et le système auditif).

Outre les activités dans les régions frontales, Okada et Hickok (2009) ont observé l'activation dans certaines régions du lobe temporal supérieur uniquement dans la tâche de lecture labiale et pas lors de la tâche sensori-motrice. Ce résultat est cohérent avec le modèle selon lequel le liage audio-visuel se réalise par une intégration cross-sensorielle dans le pSTS (Calvert *et al.*, 2000; Ghazanfar *et al.*, 2008). Les auteurs proposent que chacun de ces réseaux fournissent des informations indépendantes pour l'analyse de parole.

2.3.6 Liage audio-visuel : un modèle corrélationnel

Dans le chapitre précédent, nous avons vu les propositions selon lesquelles le liage perceptif, uni-modal ou cross-modal, serait basé sur les activités oscillatoires cohérentes des neurones (voir sous-section 1.1.2). Prenant en compte des modèles anatomiques fonctionnels de l'intégration audio-visuelle, Senkowski *et al.* (2008) suggèrent quelques scénarios possibles de type neuro-corrélationnel pour le liage audio-visuel de la parole. Le scénario le plus simple consiste en un changement de la synchronisation neuronale entre les régions sensorielles en lien avec le liage perceptif (figure 2.15, a). Une autre possibilité peut être le changement de la cohérence neuronale ou de l'intensité de réponse dans les régions multisensorielles par exemple dans les régions pariétales ou temporales supérieures (figure 2.15, b). Les scénarios a et b peuvent être tous les deux présents : le changement de synchronisation entre les régions uni-modales peut être associé à l'amélioration de l'activité oscillatoire dans les régions multisensorielles (figure 2.15, c). Les régions frontales peuvent avoir une influence modulatrice sur les régions temporo-pariétales par un couplage oscillatoire (figure 2.15, d). Le scénario le plus probable selon Senkowski *et al.* (2008) consiste en l'implication des régions uni-modales aussi bien que des régions frontales et temporo-pariétales et des structures sous-corticales dans ce réseau d'oscillation cohérente du liage audio-visuel (figure 2.15, e et f).

2.4 Conclusion

La revue de la littérature présentée dans ce chapitre propose que l'objet parole est intrinsèquement perceptuo-moteur et multisensoriel. Il est plongé dans une architecture cérébrale temporo-pariéto-frontale pour le liage, la fusion et la prise de décision. Au cours de cette thèse, nous avons souhaité obtenir une meilleure compréhension de cette nature intrinsèque et de ce réseau cérébral. Pour cela, nous interrogeons l'objet parole par un paradigme naturellement bien adapté à la question du liage : la multistabilité, que nous allons aborder dans le chapitre suivant.

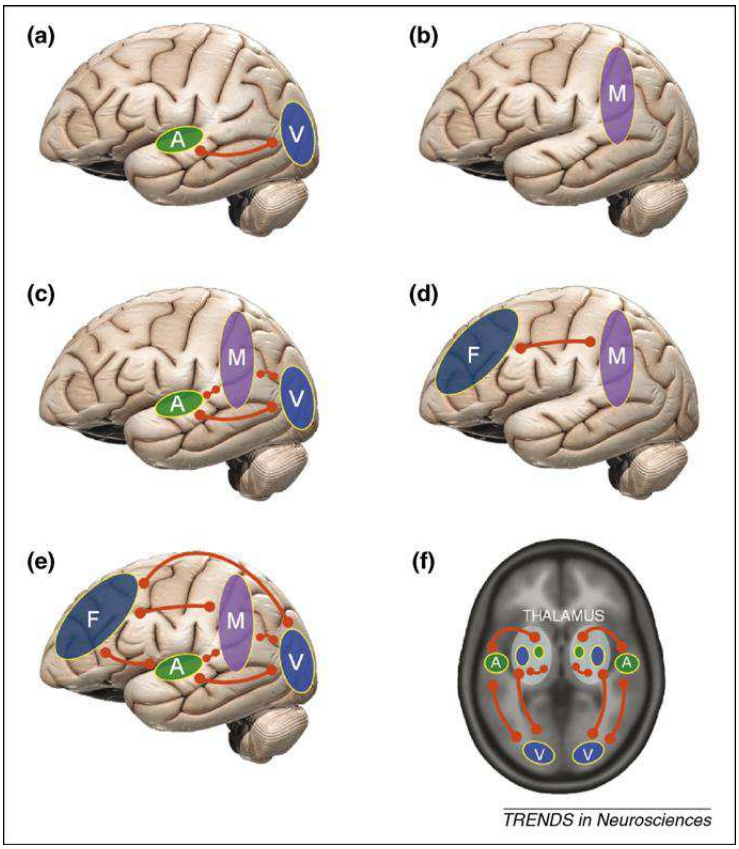


FIGURE 2.15 : Liage audio-visuel par oscillations cohérentes des neurones. Chaque image représente un scénario possible proposé par Senkowski *et al.* (2008). A : cortex auditif ; V : cortex visuel ; M : régions multisensorielles de haut-niveau ; F : cortex préfrontal. Figure tirée de Senkowski *et al.* (2008).

Chapitre 3

Multistabilité perceptive

Sommaire

Ce chapitre présente une revue de la littérature sur la multistabilité perceptive visuelle, auditive et phonétique (l'effet de transformation verbale). Dans la littérature sur la multistabilité perceptive, les recherches sur la multistabilité visuelle occupent une place prépondérante. La plupart des idées théoriques ont été proposées à l'issue de résultats expérimentaux en vision utilisant différents types de paradigmes tels que la rivalité binoculaire, les figures réversibles et les plaids dynamiques. Dans ce chapitre, nous allons d'abord faire une revue des études expérimentales et des idées théoriques expliquant le phénomène de la multistabilité perceptive en vision. Puis, une revue de la littérature sur la multistabilité perceptive auditive sera présentée. La dernière partie de ce chapitre concerne l'effet de transformation verbale où nous plaçons la multistabilité perceptive en parole en lien avec la multistabilité perceptive en général, d'une part, et avec la spécificité de la parole, d'autre part.

3.1 Multistabilité perceptive visuelle

Certaines recherches sur le phénomène de multistabilité perceptive, au travers d'expériences comportementales et neurophysiologiques, proposent que la multistabilité perceptive est un phénomène sensoriel et le résultat de processus corticaux de type bottom-up. Un des scénarios possibles cohérents avec cette idée consiste en une compétition entre des populations neuronales dans le système visuel. Selon ce

scénario, différentes populations de neurones représentant chacune un des percepts possibles entrent en compétition en s'inhibant l'une l'autre, ce qui amène une population à gagner et les autres à perdre. C'est cette population gagnante qui représente le percept actif. Ainsi, les bascules perceptives correspondent à la perte d'une population et l'activation d'une autre (pour une revue, voir Blake et Logothetis, 2002). En ce qui concerne les mécanismes responsables des bascules perceptives entre les percepts en compétition, il est proposé que l'adaptation neuronale au niveau local dans le système visuel jouerait un rôle (ex. Lehky, 1988, pour la rivalité binoculaire). Selon cette hypothèse, l'adaptation au niveau de la sortie synaptique ou du taux de décharge de la population dominante entraînerait des bascules perceptives vers un autre percept en compétition. Le rôle du bruit dans le système a été également invoqué dans les bascules perceptives. Certaines études proposent que l'adaptation doit être accompagnée par du bruit car sinon les bascules seraient régulières. Il est également proposé que le bruit serait directement la cause des bascules perceptives (Moreno-Bote *et al.*, 2007).

À l'inverse, une autre proposition est que les bascules perceptives seraient le résultat des processus non-sensoriels de type top-down initiés en dehors du système visuel. Différents mécanismes top-down ont été proposés comme responsables des bascules perceptives tels que la fluctuation de l'attention (Kawabata et Mori, 1992) ou le lancement aléatoire et itératif d'un processus de réorganisation de la perception (Leopold et Logothetis, 1999). De plus en plus, les études sur la multistabilité perceptive en vision proposent que les processus bottom-up et top-down sont tous deux impliqués dans la perception multistable, ce qui a donné naissance aux modèles hybrides (pour une revue, voir Sterzer *et al.*, 2009; Long et Toppino, 2004).

Dans cette partie, une revue sur les expériences comportementales et les données neurophysiologiques est présentée. Cette revue est séparée en trois sous-sections : processus bottom-up, processus top-down et modèles hybrides. La citation d'un article dans une sous-section ne signifie pas forcement que les auteurs de l'article ont une conclusion complètement cohérente avec le type de processus invoqué dans la sous-section en question.

3.1.1 Processus bottom-up : le rôle des mécanismes sensoriels

Études comportementales

Plusieurs études sur la multistabilité perceptive en vision ont montré que le taux de bascules perceptives augmente en fonction de la durée de simulation (Long et Toppino, 2004). Ce résultat peut être interprété comme supportant les modèles qui proposent l'adaptation comme la cause des bascules perceptives : dans chaque phase de stabilisation, les populations neuronales non-actives ne peuvent pas avoir totalement récupéré (*recovered*) par rapport au cycle précédent. Une fois la population activée, elle devient fatiguée plus tôt par rapport au cycle précédent, ce qui amène à une augmentation du taux de bascules perceptives.

Une série d'expériences a été réalisée par Toppino et Long pour démontrer que cette augmentation de nombre de bascules perceptives est due à l'adaptation et la

fatigue neuronale (voir Toppino et Long, 2005, pour une revue). Par exemple, dans une expérience utilisant le paradigme des images réversibles, ils ont montré que cet effet de fatigue est valable si l'exposition prolongée du stimulus est sur la même position rétinale et qu'en changeant l'endroit de l'exposition du cube de Necker sur la rétine, l'effet d'augmentation de nombre de bascules disparaît (Toppino et Long, 1987). Ce résultat est cohérent avec le modèle d'adaptation-récupération de la perception multistable car pendant une exposition sur le même endroit de rétine, les mêmes neurones sont stimulés tandis qu'en changeant la position rétinale de stimulation, d'autres neurones, non encore adaptés, seront impliqués.

Un autre résultat expérimental qui peut être interprété en faveur des processus bottom-up concerne la non-possibilité de contrôle total des bascules perceptives. Plusieurs études ont montré que les sujets peuvent, dans une certaine mesure, maintenir des percepts stables ou basculer plus rapidement d'un percept à l'autre (Leopold et Logothetis, 1999) mais ce contrôle n'est pas parfait et les sujets ne peuvent pas tout à fait éviter les bascules spontanées (Toppino et Long, 2005). Cette observation propose que, outre les processus top-down tels que le contrôle volontaire que nous verrons dans la suite, les processus bottom-up sensoriels seraient également impliqués dans la perception multistable[1] (Toppino et Long, 2005).

Un des défis par rapport à ce modèle d'adaptation-récupération a été souligné par Leopold et Logothetis (1999) : si le nombre de bascules augmente systématiquement en fonction de l'exposition à la stimulation, il faut que les durées des percepts successifs soient corrélées tandis que les données expérimentales ne montrent pas une tendance similaire. La figure 3.1 illustre la relation entre deux durées consécutives observées dans une tâche de perception multistable en vision (plaids dynamiques, Pressnitzer et Hupé, 2005) montrant une absence totale de corrélation.

Sur la figure 3.2, en haut, nous présentons les durées des percepts (*phase duration*) pendant l'observation d'un cube de Necker (à gauche) et une tâche de rivalité binoculaire (à droite) (Zhou *et al.*, 2004). Sur la même figure, en bas, nous présentons les densités de probabilité correspondantes qui sont compatibles avec une distribution gamma ou log-normale (voir Zhou *et al.*, 2004, pour une revue). Zhou *et al.* (2004) expliquent cette observation en soulignant le principe statistique suivant : si on considère que la perception est une tâche complexe composée de n sous-tâches indépendantes et que la probabilité de réussite dans la tâche complexe est le produit de la suite des n sous-tâches, alors, les réussites de la tâche auront une distribution log-normale pour n suffisamment grand. Ce principe statistique est plus cohérent avec la proposition de Leopold et Logothetis (1999) concernant le rôle d'un processus aléatoire et itératif de réorganisation de la perception dans la perception multistable. En résumé, les études comportementales montrent que les processus bottom-up sont impliqués dans la perception multistable mais ils ne peuvent pas être les seuls processus en jeu dans la stabilisation des percepts et les bascules d'un percept à l'autre.

[1] Il est à noter que Leopold et Logothetis (1999) interprètent ce résultat en faveur de leur modèle, que nous présenterons plus loin, selon lequel un processus aléatoire et itératif de l'exploration de la scène perceptive et de réorganisation de la perception génère les bascules perceptives.

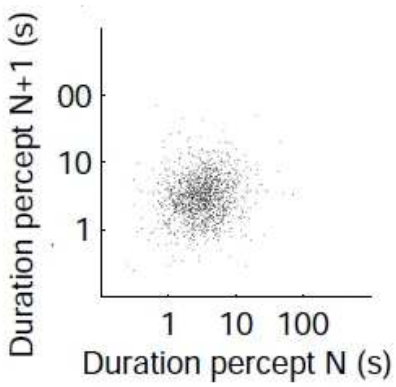


FIGURE 3.1 : Absence de corrélation entre les durées des percepts successifs (à l'échelle logarithmique) pendant la présentation des plaids dynamiques (Pressnitzer et Hupé, 2005). Figure tirée de Pressnitzer et Hupé (2005).

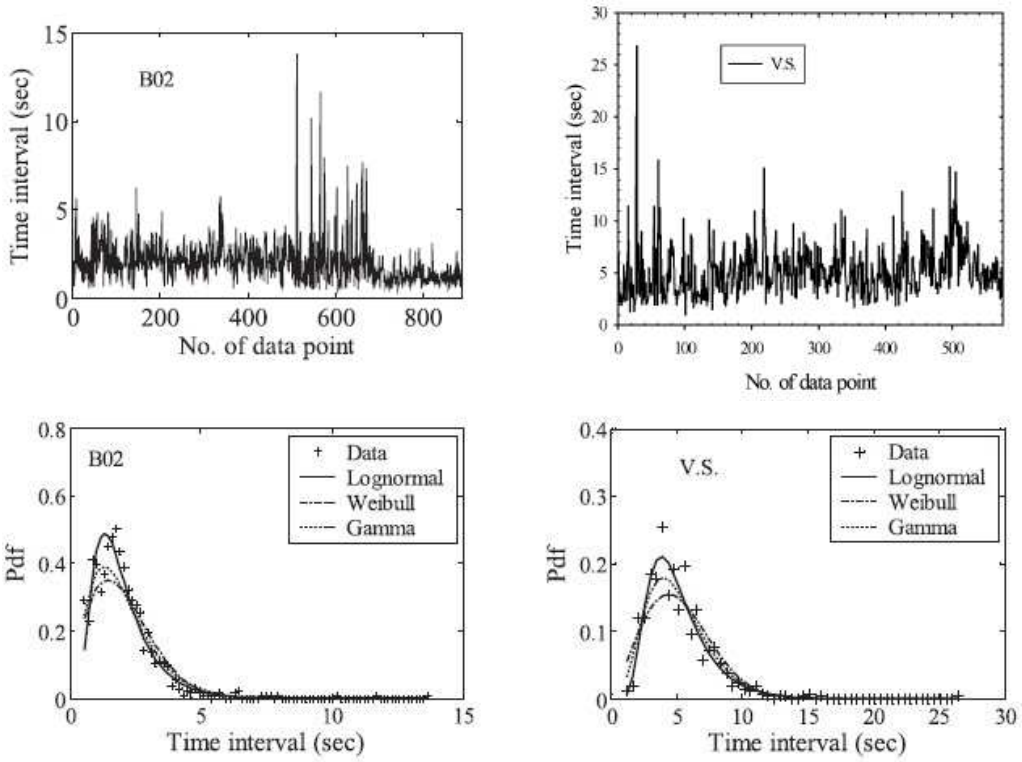


FIGURE 3.2 : Dynamique temporelle des percepts multistables dans une étude réalisée par Zhou *et al.* (2004) pour deux sujets pendant la présentation d'un cube de Necker (à gauche) et une tâche de rivalité binoculaire (à droite). En haut : les durées des percepts en fonction du temps pendant le passage du stimulus. En bas : les fonctions de densité de probabilité correspondantes. Les croix sont les résultats expérimentaux. La distribution log-normale semble permettre le meilleur ajustement. Figure tirée de Zhou *et al.* (2004).

Études neurophysiologiques

L'implication ou non de l'aire visuelle primaire (V1) a été proposée par plusieurs auteurs comme le reflet du rôle des processus bottom-up dans la perception multistable (e.g. Sterzer *et al.*, 2009; Leopold et Logothetis, 1999). Dans une expérience IRMf, Tong et Engel (2001) ont étudié l'activation du point aveugle en V1 lors d'une tâche de rivalité binoculaire. Le point aveugle correspond à la partie de la rétine qui ne possède pas de photorécepteur. Dans le cortex visuel primaire, le point aveugle correspond à une région monoculaire relativement large qui reçoit l'entrée de l'œil ipsilatéral (de même côté) et pas de l'œil controlatéral (de l'autre côté). La figure 3.3 illustre le stimulus et les conditions utilisées dans cette expérience. Les auteurs suggèrent que si la rivalité est basée sur une compétition interoculaire, l'activation de la partie du point aveugle en V1 doit augmenter lorsque les sujets perçoivent la grille qui est présentée à l'œil ipsilatéral et elle doit diminuer pendant que la grille perçue est celle présentée à l'œil controlatéral. De plus, le même pattern d'activation doit être observé pendant la présentation d'un stimulus avec les alternances physiques entre la grille ipsilatérale et la grille controlatérale.

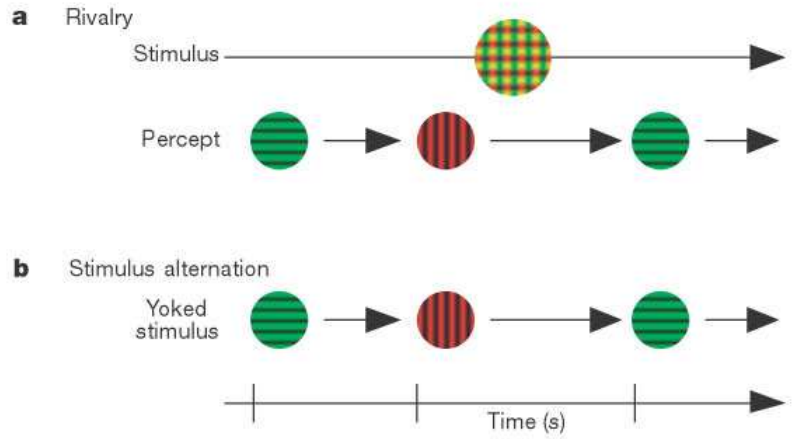


FIGURE 3.3 : Les stimuli utilisés par Tong et Engel (2001). (a) La condition de rivalité binoculaire pendant laquelle le stimulus a été présenté avec un filtre vert à un œil et avec un filtre rouge à l'autre ce qui conduit respectivement à la perception d'une grille verte et d'une grille rouge. (b) La condition des alternances physiques. Figure tirée de Tong et Engel (2001).

En accord avec les hypothèses sur la compétition interoculaire, l'activité IRMf du point aveugle augmente quand le percept correspond à la grille ipsilatérale et diminue lorsque la grille controlatérale devient dominante dans les deux conditions, rivalité binoculaire et alternances physiques. Les auteurs concluent que les bascules perceptives sont basées sur une compétition interoculaire entre les neurones monoculaires en V1.

Une autre étude montrant l'implication de V1 dans la perception multistable a été réalisé par Polonsky *et al.* (2000). Dans une tâche de rivalité binoculaire utilisant

deux patterns, l'un plus contrasté que l'autre, les auteurs ont observé que l'activité de V1 augmente lorsque les sujets observent le pattern plus contrasté et baisse pendant la perception du pattern moins contrasté.

Par ailleurs, une expérience IRMf faite par Lee *et al.* (2005) a montré que l'activité en V1 reflète la dynamique spatiotemporelle de la perception dans une tâche de rivalité binoculaire. La figure 3.4(a) illustre les grilles présentées à chaque œil dans cette expérience. Il est à noter que la transition perceptive entre ces deux grilles ne se produit pas d'une manière brusque, au contraire, elle se présente sous forme d'une onde progressive (*traveling wave*) (figure 3.4(a), à droite). Une condition normale de rivalité ne permet pas de détecter où les ondes commencent, cependant, une manipulation locale du contraste des grilles peut induire une onde progressive se produisant d'un endroit contrôlé à un moment précis (Wilson *et al.*, 2001). Dans cette expérience, Lee *et al.* (2005) ont pu initier les transitions grâce à une brève augmentation du contraste dans une petite région en haut de la grille la moins contrasté, ce qui a amené les sujets à percevoir une onde progressive de la forme du pattern le moins contrasté qui se propageait du haut vers le bas en supprimant le pattern le plus contrasté (figure 3.4(a), à droite). Si les activités en V1 reflètent la dynamique spatiotemporelle de la rivalité binoculaire, il faut donc observer une onde progressive en V1 correspondant à l'onde perçue par les sujets. Notamment, il faut observer un délai temporel de l'activité IRMf sur la carte rétinotopique de V1 en fonction de la distance corticale avec la région représentant le haut de l'anneau (début de l'onde progressive). Par exemple, le maximum de la réponse IRMf de la région en V1 correspondant à l'emplacement du cercle bleu illustré sur la figure 3.4(b) doit avoir lieu plus tardivement que celui correspondant au cercle rouge dû à la progression de l'onde à partir du cercle rouge vers le cercle bleu. La figure 3.4(c) représente les réponses IRMf de ces deux régions qui sont compatibles avec l'hypothèse décrite ci-dessus. Ces données montrent donc que l'activité en V1 peut refléter la dynamique spatiotemporelle de la perception pendant la rivalité binoculaire.

Une étude récente utilisant la figure réversible du vase de Rubin semble confirmer également l'implication de V1 dans la perception multistable (Parkkonen *et al.*, 2008). Dans cette étude, les deux percepts possibles (deux visages face à face ou un vase) ont été taggés par deux fréquences différentes, 12 Hz ou 15 Hz. Utilisant la méthode MEG, les auteurs ont observé en V1 des signaux de fréquence 12 Hz ou 15 Hz présentant des modulations d'amplitude corrélées temporellement à la perception du sujet (renforcement de la fréquence correspondant à la forme perçue par le sujet dans chaque période inter-bascule) (figure 3.5).

Bien que les résultats cités ci-dessus montrent une activation de V1 pendant la perception multistable en vision, ils ne peuvent pas clarifier si ces activités représentent un mécanisme bottom-up (e.g. une compétition locale entre les populations neuronales) ou au contraire, si elles représentent la présence des signaux feedback venant des aires de plus haut niveau (top-down). Nous reviendrons sur ce point dans la sous-section suivante.

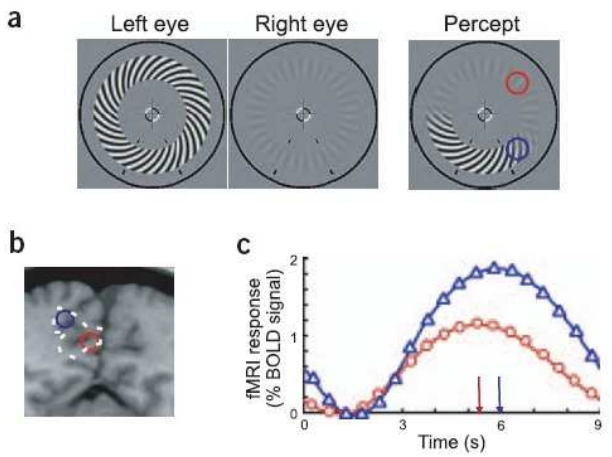


FIGURE 3.4 : Onde progressive de l'activité IRMf en V1. (a) À gauche : les deux grilles présentées à chaque œil. À droite : onde progressive perçue par les sujets. (b) Les régions rétinotopiques en V1 correspondant au cercle bleu et au cercle rouge et (c) les réponses IRMf de ces régions. Les flèches représentent le temps correspondant au maximum des réponses IRMf. Figure tirée de Lee *et al.* (2005).

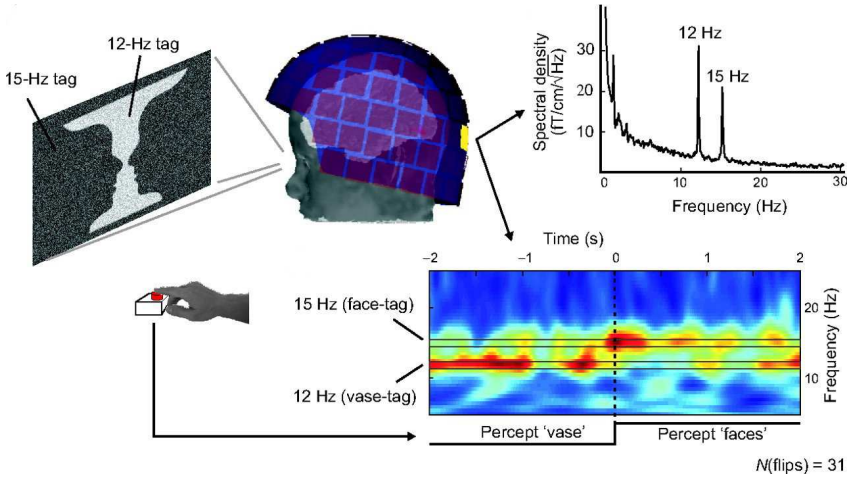


FIGURE 3.5 : Signaux MEG taggés en V1 pour un sujet dans l'étude réalisée par Parkkonen *et al.* (2008). En haut : stimulus utilisé taggé par deux fréquences différentes et le spectre de l'amplitude calculé pour un des capteurs. En bas : représentation temps-fréquence de la moyenne des signaux MEG pour 31 bascules perceptives. Figure tirée de Parkkonen *et al.* (2008).

3.1.2 Processus top-down : attention, mémoire

Études comportementales

Un des résultats expérimentaux en faveur des modèles top-down de la mutltistabilité perceptive concerne la possibilité de contrôle volontaire pendant la perception multistable. Plusieurs études ont montré que les sujets peuvent, d'une manière volontaire, stabiliser un percept ou en revanche, basculer plus souvent d'un percept à l'autre (e.g. Van Ee *et al.*, 2005) (pour une revue, voir Toppino et Long, 2005). La connaissance préalable des sujets sur la tâche influence également le nombre de bascules. Dans une tâche de perception multistable, Rock et Mitchener (1992) ont observé que la plupart des sujets non informés sur la possibilité de bascules perceptives ne signalent pas de bascule tandis qu'une fois informés, les mêmes sujets basculent d'une manière régulière et fréquente d'un percept à l'autre.

La présentation discontinue du stimulus peut également influencer la dynamique de la perception multistable. Ce mode de présentation, introduit par Orbach *et al.* (1963), consiste à enchaîner un intervalle de présentation du stimulus avec un intervalle de non-présentation. Leopold *et al.* (2002) ont observé qu'en utilisant ce mode de présentation, le nombre de bascules perceptives baisse et les bascules peuvent même disparaitre. Ils ont également observé que cet effet de stabilisation ne dépend pas de la suppression sensorielle du stimulus : la suppression illusoire du stimulus (méthode MIB, *motion induced blindness*) stabilise également les percepts. Ce résultat n'est pas cohérent avec les modèles proposant l'adaptation comme la cause des bascules perceptives car malgré la présence de stimulation, le nombre de bascules baisse. Les auteurs proposent que des mécanismes actifs de réorganisation de percept sont à l'origine de la génération des bascules, plutôt que des processus purement sensoriels (Leopold et Logothetis, 1999).

Une des explications du phénomène de stabilisation des percepts pendant la présentation discontinue du stimulus concerne l'existence de la mémoire perceptive qui persiste lors des périodes où le stimulus est absent (Leopold *et al.*, 2002; Maier *et al.*, 2003). Plusieurs études ont été réalisées pour caractériser cette mémoire (pour une revue, voir Pearson et Brascamp, 2008). Par exemple, Chen et He (2004) ont étudié l'effet de certaines caractéristiques de l'objet visuel sur cette mémoire pendant une tâche du cylindre pivotant bistable (figure dans laquelle 50% des points indiquent une rotation gauche, et 50% une rotation droite, figure 3.6, en haut). La figure 3.6, en bas, illustre les conditions expérimentales utilisées et la fréquence des bascules perceptives pour chaque condition. Cette étude montre que les caractéristiques jouant sur l'identité de l'objet telles que la couleur et la taille du stimulus n'ont pas d'effet sur la stabilisation du percept, mais que par contre, la mémoire impliquée dans la stabilisation des percepts est très sensible à la position de l'exposition du stimulus sur la rétine.

La présentation discontinue du stimulus n'a pas toujours un rôle stabilisant : pour les intervalles de non-présentation de stimuli courts (moins de 400 ms), les sujets ont tendance à basculer plus souvent (figure 3.7) (Kornmeier *et al.*, 2007; Orbach *et al.*, 1963), ce qui n'est pas cohérent avec la proposition ci-dessus sur le rôle de la mémoire.

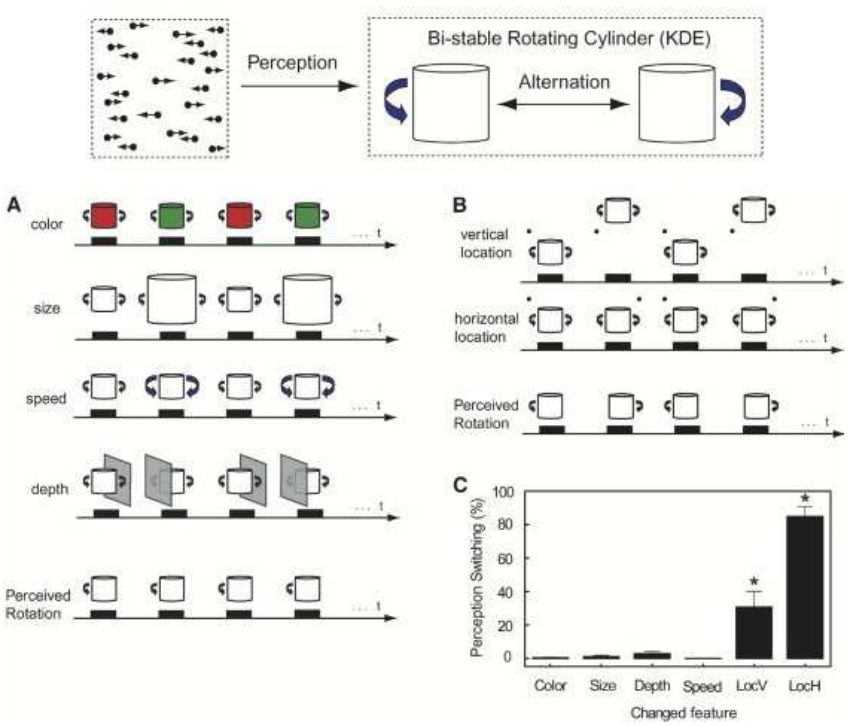


FIGURE 3.6 : Les facteurs influençant la stabilisation du percept dans l'étude réalisée par Chen et He (2004). En haut : le cylindre pivotant bistable. En bas : différentes conditions expérimentales (A) et (B) et le taux de bascules perceptives pour chaque condition (C). Seul le changement de position sur la rétine influence le taux de bascules perceptives. Figure tirée de Chen et He (2004).

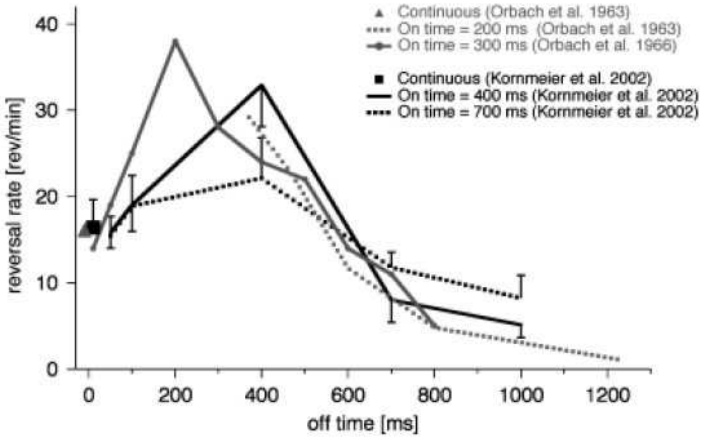


FIGURE 3.7 : Le taux de bascules perceptives pendant la présentation discontinue d'un cube de Necker (on time : présentation, off time : pause). Figure tirée de Kornmeier *et al.* (2007).

Récemment, Noest *et al.* (2007) ont proposé un modèle qui reproduit ces deux effets sans utiliser de processus top-down ni d'effet de la mémoire. Ce modèle est un réseau de neurones avec adaptation neuronale et inhibition entre les neurones représentant chaque percept. Le phénomène de mémoire peut être obtenu par l'interaction entre l'adaptation et la ligne de base neuronale (*neural baseline*). Cette interaction permet aux neurones représentant le percept dominant pendant la phase de présentation du stimulus de surmonter l'effet de l'adaptation et de redevenir dominant après la phase de non-présentation du stimulus. La figure 3.8 illustre la prédiction de ce modèle en fonction du degré d'adaptation des neurones représentant le percept 1 et le percept 2 au début de chaque phase de présentation du stimulus (*A1* et *A2*) et la ligne de base neuronale (β). Le degré d'adaptation au début de chaque phase de présentation dépend lui-même de la durée de présentation (T_{on}) et la durée de non-présentation (T_{off}). Si au début de la phase de présentation (« 0 »), le percept 1 est dominant, l'adaptation des neurones représentant ce percept augmente pendant T_{on} (flèche vers la droite) et baisse lors de T_{off} (flèche vers la gauche). Pour les T_{on} courts (a), une période courte de T_{off} est suffisante pour obtenir l'effet de mémoire mais pour les T_{on} long (b), le choix de percept dépend de T_{off} : les durées courtes de T_{off} entraînent une bascule (3) et les durées longues de T_{off} conduisent à la répétition du même percept (2).

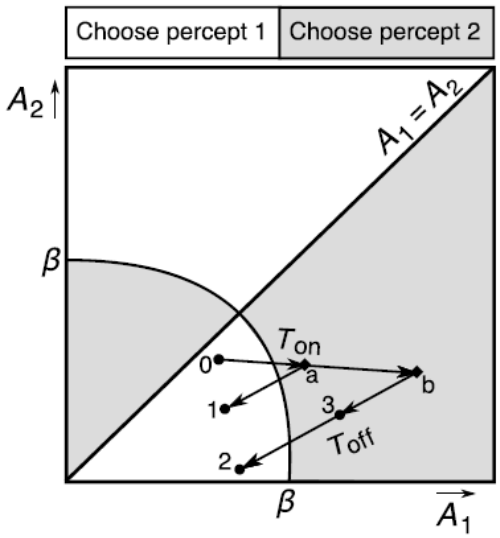


FIGURE 3.8 : Le modèle de Noest *et al.* (2007). *A1* et *A2* : adaptation des deux percepts concurrents (percept 1 et percept 2). β : ligne de base neuronale. T_{on} : durée de présentation. T_{off} : durée de non-présentation. Figure tirée de Klink *et al.* (2008).

Études neurophysiologiques

Plusieurs études neurophysiologiques montrent que les aires non-primaires de traitement visuel (cortex extrastrié) sont impliquées pendant différents paradigmes

de perception multistable chez le singe et l'humain. Il est proposé que ces régions sont spécialisées dans la représentation du contenu de la perception consciente (pour une revue, voir Leopold et Logothetis, 1999; Blake et Logothetis, 2002). Les aires fronto-pariétales sont également impliquées dans la perception multistable en vision. Dans la suite, une brève revue sera présentée sur les études montrant l'activation des aires fronto-pariétales pendant les tâches de multistabilité perceptive visuelle. Ces activations ont été observées aussi bien en lien avec les bascules perceptives et dans la phase de stabilisation des percepts que lors de la présentation discontinue des stimuli bistables.

Rôle dans les bascules perceptives : Deux types de mécanismes pourraient expliquer l'implication des aires fronto-pariétales dans les bascules perceptives (Sterzer *et al.*, 2009).

- Mécanisme feedforward : les activités des aires fronto-pariétales reflètent un flux d'activation feedforward, à partir des aires visuelles primaires vers les aires de plus haut niveau. Ce mécanisme serait identique à celui qui surviendrait lors de la perception non-multistable du changement réel dans le stimulus visuel. Cette interprétation est cohérente avec l'idée selon laquelle les bascules perceptives sont le résultat des processus bottom-up.

- Mécanisme top-down : Les activités pariéto-frontales peuvent également correspondre aux processus cérébraux top-down qui initialisent les bascules perceptives et entraînent la modulation des activités des aires visuelles.

Dans une étude IRMf sur la rivalité binoculaire, Lumer *et al.* (1998) ont observé, pour la première fois, des activations dans les aires fronto-pariétales correspondant aux bascules perceptives. Dans cette étude, les auteurs ont mis en relief deux conditions : la condition de rivalité (visage vs. grille) et celle de l'émulation de la rivalité. Dans cette deuxième condition, le stimulus consistait en une réplication des percepts des sujets (visage ou grille) pendant la condition de rivalité binoculaire. Pour chaque bascule, un mélange de deux patterns en rivalité a été présenté. Cette condition a permis une simulation de la perception binoculaire des sujets du point de vue temporel et qualitatif. Les auteurs ont observé que certaines régions fronto-pariétales dans le cortex droit étaient seulement actives en lien avec les bascules perceptives dans la condition de rivalité binoculaire et pas dans la deuxième condition d'émulation.

Un lien de causalité ou non-causalité entre les activités pariéto-frontales observées et les bascules perceptives pourra clarifier le mécanisme en jeu. Bien que les mesures neurophysiologiques ne puissent que démontrer l'éventuelle corrélation entre la tâche réalisée et les activités observées, une précédence temporelle est considérée comme une information en faveur de la causalité (Leopold et Logothetis, 1999). Autrement dit, il faut vérifier si les activités pariéto-frontales précèdent temporellement les bascules perceptives et les activités des aires visuelles extrastriées et primaires correspondant aux bascules perceptives.

Une étude IRMf récente réalisée par Sterzer et Kleinschmidt (2007) a montré qu'effectivement, les activités du cortex préfrontal droit précèdent celles de V5/MT

pendant la présentation du stimulus de type mouvement apparent (*apparent motion*) au contraire de ce que l'on observe dans la perception des changements réels dans le stimulus d'entrée non-bistable. Ce résultat suggère que les aires préfrontales contribueraient à la dynamique des percepts en initialisant les bascules perceptives (Sterzer et Kleinschmidt, 2007). Une étude iEEG (*intracerebral* EEG) a également montré que les activités neuronales dans le cortex préfrontal droit précédaient les bascules perceptives pendant que les sujets regardaient un cube de Necker (Britz *et al.*, 2009). La région observée dans cette étude a été aussi observée par Lumer *et al.* (1998).

Une autre donnée en faveur du rôle fonctionnel des aires fronto-pariétales dans les bascules perceptives vient des études qui montrent que le nombre de bascules perceptives baisse chez les sujets subissant des lésions focales dans les zones préfrontales (Windmann *et al.*, 2006) et pariétales (Bonneh *et al.*, 2004).

Une remarque intéressante soulignée par Sterzer *et al.* (2009) concerne la similarité entre les régions fronto-pariétales actives correspondant aux bascules perceptives et celles correspondant aux tâches de direction de l'attention vers les événements saillants. Nous reviendrons sur cette remarque dans la suite.

Rôle dans la stabilisation des percepts : Outre les études montrant l'implication des aires fronto-pariétales pendant les bascules perceptives, ces aires seraient également impliquées pendant la stabilisation des percepts. Dans une étude IRMf utilisant le mode de présentation discontinue du stimulus, Sterzer et Rees (2008) ont observé une corrélation entre les activités dans les aires fronto-pariétales et la tendance des sujets à stabiliser des percepts pendant les phases de non-présentation de stimulus. Les auteurs soulignent que les mêmes zones ont été observées dans plusieurs études sur la mémoire de travail visuelle et la sélection visuelle attentive.

Rôle fonctionnel des régions fronto-pariétales : Naghavi et Nyberg (2005) ont fait une revue sur les études IRMf et PET montrant l'implication des régions pariétales et frontales dans différentes tâches en lien avec l'attention, la mémoire de travail, la mémoire épisodique et la conscience[2]. Cette revue propose qu'il existe une similarité entre les régions actives pendant ces tâches spécialement dans le cortex pariétal bilatéral (BA7 et BA40 ; proche du sulcus intrapariétal) et le cortex préfrontal dorsolatéral (BA9 à droite et BA6 à gauche). Cette similarité peut être expliquée, selon les auteurs, par l'implication de fonctions cognitives proches dans ces tâches qui seraient basées sur les mêmes structures anatomiques. Les auteurs soulignent spécialement le rôle des processus d'intégration et de liage impliqués dans ces tâches. Ils suggèrent que l'intégration bas-niveau comme le liage entre différentes caractéristiques d'un objet visuel pourraient être basés sur les aires sensorielles du cerveau mais que l'intégration haut-niveau entre les représentations distribuées et multisensorielles impliquées dans les tâches sur la conscience, la mémoire de travail, la mémoire épisodique et l'attention serait basée sur l'activité des aires fronto-

[2]Il est à noter que la plupart des études dans la revue sur la conscience concernent celles utilisant le paradigme de la rivalité binoculaire et les figures réversibles.

pariétales.

Bien que l'implication des régions pariéto-frontales dans la perception multistable soit démontrée par plusieurs études, le rôle exact de ces régions n'est pas encore connu. Prenant en compte la similarité entre les régions actives pendant les bascules perceptives et celles correspondant aux tâches impliquant la mémoire de travail, attention et surtout la redirection de l'attention vers les événements saillants, Sterzer *et al.* (2009) ont proposé deux scénarios possibles :

- Feedback spontané vers les aires de traitement visuel : les régions fronto-pariétales qui sont impliquées dans les tâches d'attention volontaire généreraient des signaux de feedback d'une manière spontanée et automatique. Ces feedbacks pourraient servir à la réévaluation du percept comme proposé par Leopold et Logothetis (1999). Ces signaux initieraient ainsi une réorganisation perceptive de l'entrée lorsque l'équilibre des forces entre les populations neuronales codant les percepts dans les aires sensorielles déstabilise dû à l'adaptation ou l'inhibition mutuelle.

- Redirection de l'attention vers les événements saillants : la déstabilisation de l'équilibre des forces entre les populations neuronales dans les aires visuelles pourrait jouer le rôle d'un événement saillant qui activerait en conséquence les régions pariéto-frontales. Ces régions activées pourraient faire de sorte que l'attention soit redirigée vers l'entrée sensorielle et ainsi ils initialiseraient une réévaluation du percept pouvant entraîner une bascule perceptive.

De plus en plus, les études sur la multistabilité perceptive en vision proposent que les aires de traitement visuel et les aires pariéto-frontales interagissent d'une manière active pendant la perception multistable. Une brève revue sur ces propositions hybrides est présentée dans la suite.

3.1.3 Vers des modèles hybrides

Long et Toppino (2004) proposent dans un article de revue sur les études sur les figures réversibles, un modèle hybride multi-niveaux incorporant les deux types de mécanismes, bottom-up et top-down. Dans ce modèle, un premier niveau de traitement visuel extrait les caractéristiques du stimulus telles que sa taille, son orientation, etc. Deux niveaux intermédiaires reçoivent les signaux bottom-up du niveau d'extraction des caractéristiques visuelles et ceux de type top-down des mécanismes cognitifs de haut niveau comme l'attention et l'apprentissage (voir figure 3.9).

Ce cadre est cohérent à la fois avec les études comportementales montrant l'implication des effets de stimuli et des effets cognitifs de plus haut niveau et avec les études neuro-anatomiques proposant un réseau actif distribué correspondant à la perception multistable. Dans leur article de revue sur les bases neuronales de la multistabilité perceptive, Sterzer *et al.* (2009) proposent un cadre hybride dans lequel la perception multistable est le résultat des interactions continues entre les

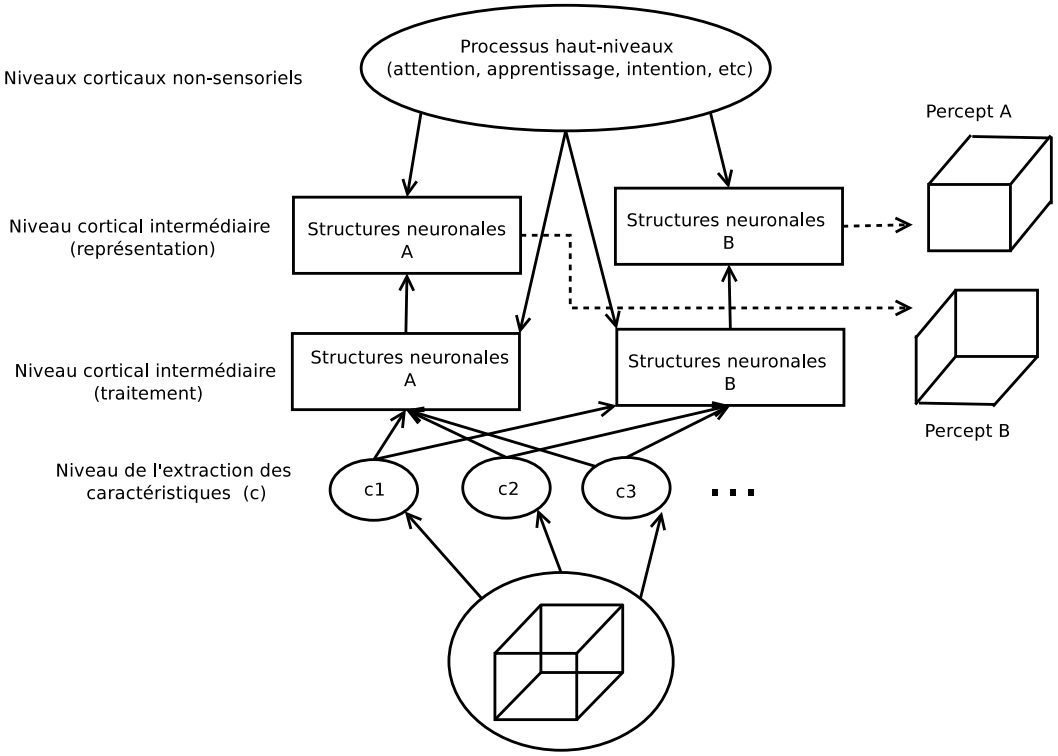


FIGURE 3.9 : Modèle hybride proposé par Long et Toppino (2004) pour la multistabilité perceptive des figures « réversibles ».

régions corticales sensorielles et les régions pariétales et frontales : lorsque les processus sensoriels amènent à la déstabilisation du percept dominant, les mécanismes de haut niveau interviendraient et initialiseraient une réorganisation perceptive.

3.2 Multistabilité perceptive auditive

La plupart des études sur la multistabilité perceptive en audition utilisent un paradigme similaire à celui proposé par Bregman et Campbell (1971) et Van Noorden (1975) que nous avons présenté dans la sous-section 1.2.3. La multistabilité auditive concerne ainsi des bascules perceptives entre un flux (cohérence) et deux flux (streaming) auditifs.

3.2.1 Mise en évidence

Une étude réalisée par Pressnitzer et Hupé (2006) propose que la multistabilité perceptive en audition présente plusieurs caractéristiques communes avec la multistabilité perceptive en vision. Dans cette étude, les auteurs ont utilisé des stimuli bistables auditifs et visuels. Le stimulus auditif était une suite ABA- (A : 587 Hz, B : 440 Hz, - : un silence de 120 ms) et le stimulus visuel était de type plaid dynamique (voir un exemple sur la figure 3.10). Dans les deux modalités, les sujets percevaient soit une seule forme groupée (ABA- pour la modalité auditive et un seul plaid se dé-

plaçant vers une direction pour la modalité visuelle) soit deux formes séparées (A-A- et B—B pour la modalité auditive et deux plaids se déplaçant vers deux directions différentes). Les auteurs ont montré que les distributions temporelles des percepts pour le stimulus auditif et le stimulus visuel ne sont pas différentes d'une façon significative d'une distribution log-normale. De plus, elles ne sont pas significativement différentes l'une avec l'autre, ce qui signifie que la dynamique temporelle de la perception est fortement similaire dans les deux modalités (Pressnitzer et Hupé, 2006).

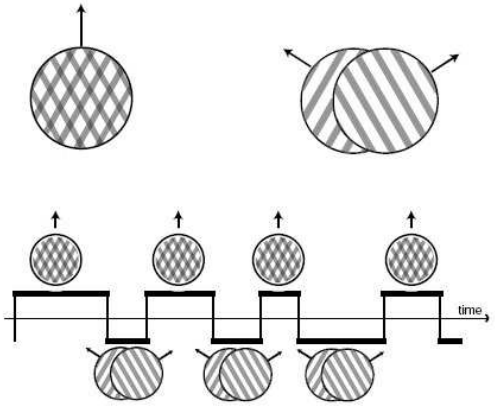


FIGURE 3.10 : Plaids dynamiques. En haut : les deux percepts possibles, à gauche, le percept cohérent et à droite, le percept transparent. En bas : la dynamique des bascules perceptives. Figure tirée de Rubin et Hupé (2005).

3.2.2 Interactions entre multistabilité auditive et visuelle

Le résultat ci-dessus sur la similarité entre la perception multistable en vision et en audition pourrait être interprété selon la proposition de Leopold et Logothetis (1999) présentée au début de la section 3.1 qui suggère qu'un mécanisme central de réorganisation de scène perceptive génère les bascules. Pour vérifier cette hypothèse, Hupé *et al.* (2008) ont réalisé une série d'expériences comportementales. Dans la première expérience, les auteurs ont utilisé les mêmes stimuli que dans leur précédente étude (Pressnitzer et Hupé, 2006) en modalité auditive seule, visuelle seule et bimodale (audio-visuelle), c'est-à-dire en superposant le flux auditif (ABA-) et le flux visuel (plaids). Dans la deuxième expérience, ils ont choisi des stimuli audio-visuels plus congruents : les sons A et B ont été présentés par des haut-parleurs, le son A du haut-parleur à gauche et le son B du haut-parleur à droite. Une diode de couleur rouge a été mise au centre de chaque haut-parleur. Chaque diode flashait de manière synchrone avec le son diffusé par chaque haut-parleur. Ce stimulus visuel était perçu soit comme un mouvement apparent (deux clignotants groupés) soit comme deux flash qui clignotaient d'une façon indépendante.

Dans cette étude, une interaction entre la modalité auditive et visuelle pourrait refléter l'implication d'un mécanisme central et commun. Si un mécanisme central

commun génère des bascules perceptives, soit il doit alterner entre les modalités, ce qui entraîne la baisse de la coïncidence entre les bascules auditives et visuelles pendant la présentation bimodale (par exemple, pendant une bascule auditive, on ne peut pas avoir de bascule visuelle), soit il regroupe les deux modalités ensemble et une synchronisation sera observée entre les bascules auditives et visuelles.

Les résultats des expériences présentées en haut n'ont pas montré d'interaction entre la modalité auditive et la modalité visuelle. Le temps passé dans un percept (groupé ou séparé) et le nombre de bascules n'étaient pas significativement différents en présentation unimodale par rapport à la présentation bimodale. De plus, aucun effet de la présentation bimodale n'a été observé sur la probabilité de la coïncidence des bascules auditives et visuelles. Ces observations suggèrent qu'un mécanisme central commun comme proposé par Leopold et Logothetis (1999) ne peut pas être la cause des bascules perceptives. Hupé *et al.* (2008) proposent un modèle de compétition distribuée pour la perception multistable en distinguant deux concepts : ce qui est en compétition (*what*) et comment la compétition/bascule perceptive se produit (*how*). Ils proposent que les effet top-down tels que l'attention agissent sur ce qui est en compétition (différents percepts possibles) et non pas sur le mécanisme de bascule. Selon cette proposition, ces effets top-down, d'une manière similaire aux caractéristiques physiques du stimulus, peuvent changer le poids de chaque interprétation, ce qui peut influencer par exemple la durée de chaque percept.

La compétition distribuée implique que le corrélat neuronal de la perception multistable soit observé dans différents niveaux de traitement neuronal. Cette proposition est cohérente avec les études neurophysiologiques sur la multistabilité perceptive en vision présentées dans les sous-sections 3.1.1 et 3.1.2 et celles sur la bistabilité auditive qui seront présentées dans la suite.

3.2.3 Architectures neuronales

Une des études neurophysiologiques sur le caractère bistable des stimuli de type ABA- montre un lien fort entre l'activité du cortex auditif et l'organisation perceptive de ces stimuli (Gutschalk *et al.*, 2005). Dans cette étude utilisant la méthode MEG, la différence fréquentielle entre les sons A et B et le rythme de présentation étaient choisis de sorte que la perception bascule entre deux formes, la cohérence (un flux) et le streaming (deux flux séparés). Il était demandé aux sujets de signaler leurs bascules perceptives lors de l'enregistrement des signaux MEG. Les auteurs ont mesuré les champs évoqués auditifs (*auditory evoked field*) dans le cortex auditif. Comme illustré sur la figure 3.11, la valeur de P_1m (50-70 ms après l'onset du stimulus) et N_1m (100-120 ms après l'onset du stimulus) correspondant au son B était plus importante lorsque les sujets percevaient deux flux séparés (streaming) par rapport à la perception d'un seul flux. En se basant sur la littérature, les auteurs proposent que ces activités de P_1m et N_1m ne soient pas générées dans le cortex auditif primaire mais dans le gyrus de Heschl latéral, le planum temporale et le gyrus temporal supérieur.

Dans une étude IRMf utilisant les stimuli bistables de type ABA-, Cusack (2005) a comparé l'activité cérébrale pendant la perception d'un flux et celle pendant la

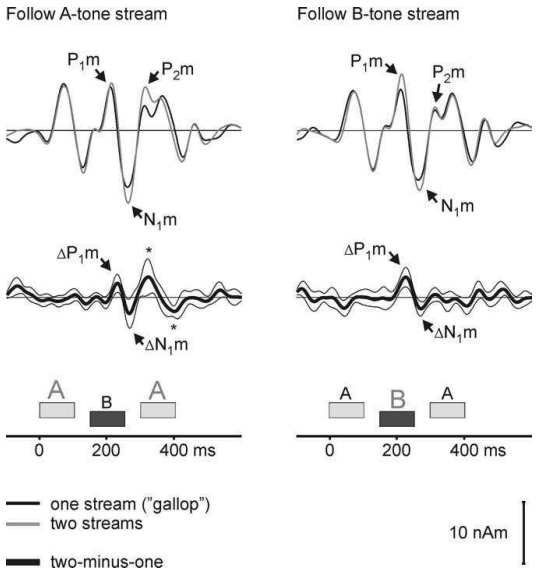


FIGURE 3.11 : Champ évoqué auditif pour un stimulus de type ABA- dans le cortex auditif lorsqu'il étaient demandé aux sujets de suivre le son A (à gauche) et le son B (à droite) (Gutschalk *et al.*, 2005). Figure tirée de Gutschalk *et al.* (2005).

perception de deux flux. L'auteur a observé que l'activation du suclus intraparietal (IPS) est plus importante lors de la perception de deux flux. Ce résultat a été interprété en invoquant le rôle général de l'IPS dans les mécanismes de liage perceptif (par rapport aux modalités auditive, visuelle ou tactile ou dans le cadre de l'intégration multisensorielle, Cusack, 2005) et dans l'encodage du nombre d'objets perçus (surtout la partie inférieur de l'IPS) (Cusack *et al.*, 2010).

En résumé, différentes aires corticales semblent être impliquées dans le streaming, ce qui suggère qu'un réseau distribué pourrait être en charge de l'organisation perceptive auditive. En ce qui concerne les aires préfrontales, à notre connaissance, elles ne sont pas signalées actives directement pour les stimuli auditifs bistables. Cependant, ces aires sont impliquées dans l'identification et la catégorisation des objets auditifs (eg. Lee *et al.* (2009) chez le singe et Binder *et al.* (2004) chez l'humain). Dans le cadre de leur modèle de streaming auditif présenté dans la sous-section 1.2.4, Micheyl *et al.* (2005) proposent que les neurones de décision pourraient se situer dans le cortex préfrontal.

3.3 Multistabilité perceptive en parole

L'effet de transformation verbale a été présenté par Warren et Gregory (1958) comme l'analogue de la multistabilité perceptive lors de l'observation continue des images réversibles. Cet effet concerne les bascules perceptives qui ont lieu pendant l'écoute en boucle ou lors de la répétition d'une séquence langagière. Par exemple, en écoutant une répétition rapide du mot anglais *life*, notre perception bascule du mot *life* au mot *fly* et ensuite de *fly* à *life* et ainsi de suite. L'effet de transformation

verbale a été présenté comme un phénomène d'illusion auditive au même titre que l'effet de restauration phonémique (Warren et Warren, 1970). Depuis sa découverte, différentes caractéristiques de l'effet de transformation verbale ont été étudiées par Warren et collègues (voir figure 3.12 pour une mise en place expérimentale). Il a été montré que le nombre de transformations pouvaient dépendre de la longueur des stimuli utilisés, de la durée de l'intervalle inter-stimulus et du nombre de répétitions (Warren, 1961b). L'age des participants semble être également un facteur important pour la perception des transformations verbales : les jeunes enfants ou les adultes âgés sont moins sensibles à cet effet (Warren, 1961a).

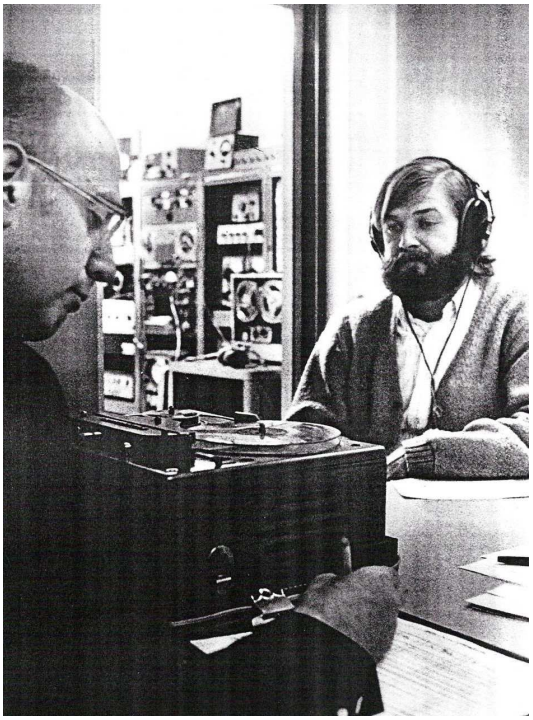


FIGURE 3.12 : Mise en place d'une expérience perceptive sur l'illusion auditive par Warren et collègues. Dans les expériences sur l'effet de transformation verbale, les sujets écoutaient les stimuli enregistrés sur une cassette et signalaient à l'expérimentateur ce qu'ils entendaient initialement et ensuite, ils signalaient leur percept chaque fois qu'ils entendaient un changement. Figure tirée de Warren et Warren (1970).

Les expériences sur l'effet de transformation verbale mettent en évidence différents types de transformation. En comparaison à la séquence répétée, les transformations perçues peuvent être compatibles avec une resegmentation du stimulus (ex. *life* → *fly*), avec des modifications phonétiques, lexicales ou sémantiques (ex. /poʤ/ → /pold/) ou avec un regroupement auditivo-perceptif de type streaming (ex. /skin/ → /kin/ ou /gin/ + /s/ en toile de fond) (Pitt et Shoaf, 2002). Nous allons passer en revue les processus à la base des différents types de transformations verbales et les facteurs influençant ces transformations.

3.3.1 L'effet de transformation verbale : les causes

L'effet de transformation verbale a été expliqué par Warren et Warren (1970) et Warren (1983) comme étant la conséquence normale des processus de l'organisation perceptive de la parole. Ils proposent que ces mécanismes, de nature constructive, garantissent une interprétation appropriée de l'entrée lorsque les informations auditives sont ambiguës, déformées ou incomplètes. En accord avec cette explication, l'absence de l'effet de transformation verbale chez les jeunes enfants pourrait être dû au fait qu'ils n'ont pas encore acquis les compétences nécessaires pour la réorganisation perceptive (Warren et Warren, 1970).

En analogie avec la multistabilité perceptive en vision, une des explications de l'effet de transformation verbale pourrait être l'habituation. Snyder *et al.* (1993) ont vérifié cette possibilité dans une expérience d'une durée de 6 minutes en comparant le nombre de transformations signalées pendant la première période de 3 minutes de passage de stimulus avec la second période de 3 minutes. Ils ont observé que les sujets percevaient plus de transformations pendant la seconde période de stimulus. Les auteurs ont interprété ce résultat comme le reflet de l'implication des processus d'habituation dans l'effet de transformation verbale. Il est cependant important de noter qu'il y a des données contradictoires sur l'évolution du nombre de transformations pendant le passage de stimulus. Par exemple, Lass et Golden (1971) ont observé une baisse du nombre de transformations au cours du temps.

Warren et Meyers (1987) proposent que l'habituation, au moins au niveau des mécanismes auditifs, ne peut pas être la cause de l'effet de transformation verbale (voir aussi Kaminska *et al.*, 2000, pour des données contre l'habituation auditive). Ils proposent deux processus à la base de l'effet de transformation verbale : la satiété et le changement de critère. La répétition d'une séquence entraîne la satiété de sa représentation en mémoire. En même temps, le critère utilisé pour catégoriser le stimulus change de sorte qu'une autre représentation est estimée avoir une meilleure correspondance au stimulus d'entrée que la représentation initiale, ce qui entraîne une transformation verbale. Ces processus se répètent tout au long de la répétition du stimulus. Il est proposé que la satiété opérait au niveau des représentations lexicales et sous-lexicales (ex. syllabiques). Une des propositions les plus élaborées expliquant l'effet de transformation verbale a été fournie par MacKay *et al.* (1993) dans le cadre d'une théorie de la perception/production de la parole appelée *Node Structure Theory* (MacKay, 1982, 1988). Ces auteurs proposent également que les transformations verbales se produisent essentiellement par les processus de saturation. Nous présentons en détail dans la sous-section 7.3.1 la *Node Structure Theory* et les propositions de MacKay *et al.* (1993) concernant les mécanismes à la base de l'effet de transformation verbale.

Les analyses détaillées de la nature des transformations signalées par les sujets sont une source riche d'information sur les processus responsables de l'effet de transformation verbale. Dans une expérience où la majorité des transformations correspondaient à la substitution phonémique, Pitt et Shoaf (2001) ont observé plusieurs régularités concernant cette substitution. Par exemple, pour les consonnes, la substitution portait sur le lieu d'articulation (ex. /b/→/d/ ou /m/→/n/) et pour

les voyelles, les voyelles antérieures avaient tendance à devenir plus basses et plus postérieures (ex. /i/→/I/). Ils ont interprété leurs résultats en terme de satiété des représentations segmentales : la présentation répétitive d'une séquence conduirait à la satiété de sa représentation, ce qui engendrerait un petit changement dans l'identité du phonème (on retrouve là les éléments classiques de littérature sur l'adaptation sélective phonétique : voir Cooper, 1974).

Il ne semble pas que la satiété puisse expliquer tous les types de transformations signalées. Les expériences réalisées par Pitt et Shoaf (2002) sur les transformations verbales de type streaming montrent clairement qu'un des processus impliqué dans l'effet de transformation verbale est le regroupement auditivo-perceptif. Dans cette étude, les auteurs ont montré que pour certaines séquences répétées, les sujets entendaient des flux auditifs distincts et la transformation verbale correspondait à un percept en avant-plan et un autre en arrière-plan. Par exemple, la répétition de /wɛm/ conduisait aux transformations telles que /wɛ/ ou /wæ/ et /m/ formait un flux en arrière-plan. Un autre stimulus utilisé par les auteurs était /pɛʧ/ qui formait les transformations de type /pɛt/ et /pɛ/ avec le flux /ʧ/ en arrière-plan. Les résultats de ces expériences montrent ainsi une correspondance étroite entre le percept en avant-plan et le contenu phonétique de la partie séparée en arrière-plan.

Cette observation a amené les auteurs à questionner sur la nature de ce mécanisme : est-ce que ce phénomène peut être expliqué par les mécanismes de l'analyse de scènes auditives « à la Bregman » ou est-ce qu'il est spécifique à la parole ? Pour répondre à cette question, Pitt et Shoaf (2002) ont utilisé des séquences sinusoïdales de parole (*sine-wave*, voir section 1.3) dans une expérience de transformation verbale. Les stimuli ont été présentés aux sujets dans une condition comme des séquences de parole produites par l'ordinateur et dans une autre condition, les stimuli n'ont pas été présentés comme de la parole. Les auteurs ont observé des transformations très similaires dans les deux conditions. En comparant ces deux conditions, ils ont également observé que les séquences sinusoïdales présentées comme de la parole sont plus résistantes face au streaming : les transformations sont plus longues et il faut deux fois plus de temps pour qu'elles émergent. Bien que ce résultat ne puisse pas être interprété en faveur de l'implication de processus spécifiques à la parole, il met en évidence l'implication de processus de type top-down dans l'organisation perceptive des transformations verbales.

Un des processus top-down impliqués dans l'effet de transformation verbale semble être basé sur les effets lexicaux : la mémoire lexicale pourrait influencer la puissance de groupement des éléments auditifs en un mot (Pitt et Shoaf, 2002). En accord avec cette hypothèse, Natsoulas (1965) a montré que le nombre de différentes formes de transformations perçues est moins important pour des mots que pour des pseudo-mots. Dans une série d'expériences, Shoaf et Pitt (2002) ont répliqué ce résultat et montré que cet effet n'était pas dû au nombre de transformations perçues car le nombre total de transformations signalées par les sujets n'était pas significativement plus important pour les stimuli de type pseudo-mot que pour les mots. Lorsque la séquence répétée était un mot, les auteurs ont observé que les sujets basculaient plus souvent vers la forme originelle répétée par rapport aux conditions où la séquence répétée était un pseudo-mot. Cet effet lexical pourrait être relié à la

mémoire lexicale qui n'a pas seulement le rôle d'un biais conduisant les sujets à percevoir un mot mais elle jouerait également un rôle dans la stabilisation du percept cohérent avec la séquence originelle répétée. Ce mécanisme, appelé par les auteurs mécanisme de récupération, est plus fort pour les mots que pour les pseudo-mots et pour les séquences phonologiquement légales que pour les séquences illégales.

Les résultats ci-dessus semblent donc globalement compatibles avec les principes de l'analyse de scènes auditives présentés dans la section 1.2 : les différents types de transformations perçues pendant les expériences perceptives sur l'effet de transformation verbale pourraient être expliqués par des mécanismes généraux de type primitif (ex. pour le streaming) et par des schémas phonétiques et lexicaux (ex. pour la substitution phonétique ou sémantique). Cependant, les contraintes articulatoires semblent également être impliquées dans ce phénomène, comme nous allons le voir.

3.3.2 Le rôle des contraintes articulatoires

La première démonstration du rôle des contraintes articulatoires dans l'effet de transformation verbale a été fournie par Reisberg *et al.* (1989). Dans une série d'expériences, les auteurs ont demandé à un groupe de sujets de répéter mentalement un mot et de signaler toute transformation perçue. Ils ont contrôlé le degré de subvocalisation lors de la répétition en ajoutant une condition de chuchotement et une condition de répétition silencieuse. Ils ont également ajouté trois conditions pendant lesquelles la possibilité d'*enactment* (simulation mentale motrice) était perturbée (conditions de blocage de la mâchoire, lèvres jointes et langue collée au palais, condition de suppression articulatoire conjointe[3] et condition de mâcher du chewing-gum). Les auteurs ont demandé à un autre groupe de sujets de signaler les transformations perçues lors de la répétition à haute voix du même mot, produite soit par l'examinateur soit par le sujet lui-même. Ils ont observé que le taux de transformations diminue de la condition d'externalisation complète vers la condition de répétition mentale en passant par les conditions d'externalisation partielle (chuchotement et répétition silencieuse). Le processus d'émergence dépendrait ainsi du degré de subvocalisation possible. Les données de ces expériences montrent également que les transformations verbales disparaissent dans les trois conditions de suppression d'enactment, ce qui suggère l'implication du système articulatoire dans les processus d'émergence des transformations.

Sato *et al.* (2006) font remarquer le manque d'études sur les éventuelles asymétries des transformations perçues, ce qui pourrait également refléter l'implication des contraintes articulatoires dans l'effet de transformation verbale. Les auteurs donnent l'exemple de la transformation de *life* vers *fly* qui semble être plus probable que celle de *fly* vers *life*. Cette asymétrie pourraient être liée à la cohésion articulatoire de ces deux formes : dans *fly*, les trois gestes peuvent être produits avec une synchronie importante tandis que la synchronisation de [l] et [f][4] dans *life* est impossible. Ce

[3] La suppression articulatoire conjointe consiste à répéter à haute voix une séquence dépourvue de sens pendant la répétition mentale d'un mot.

[4] Dans la suite, nous utilisons systématiquement la notation // sans distinguer les unités phonologiques des contenus phonémique (notations // vs. [])

phasage de gestes dans *fly* pourrait expliquer la facilitation de transformation de *life* vers *fly*.

Dans une série d'expériences sur l'effet de transformation verbale, Sato *et al.* (2006) ont étudié cette hypothèse de phasage articulatoire lors des conditions de répétition à haute voix, de répétition mentale et de l'écoute sans répétition en utilisant les séquences /psə/, /səp/,/əps/, /spə/, /pəs/ et /əsp/. Dans les conditions de répétition à haute voix et de répétition mentale, la séquence /psə/ était la séquence la plus attractive, i.e. le nombre de transformations de /səp/ ou /əps/ vers /psə/ était significativement plus grand que le nombre de transformations de /psə/ vers /səp/ ou /əps/. Les auteurs expliquent ce résultat en soulignant que dans la séquence /psə/, /p/ et /s/ peuvent être produits en synchronie grâce à l'anticipation du geste de /s/ : lorsque les lèvres s'ouvrent pour produire /p/, le bout de la langue est en contact avec le palais, ce qui permet la production de /s/ immédiatement après le /p/. Ce phasage n'est pas possible pour la séquence /səp/ car les lèvres doivent rester ouvertes pendant la réalisation de /s/ puis elles doivent se fermer et se rouvrir pour la production de /p/. Il est important de noter que l'existence de cette asymétrie dans la condition de répétition mentale suggère que les contraintes articulatoires sont également impliquées dans la construction des représentations verbales même lorsqu'il n'existe pas de signal auditif externe et sans la production des gestes. Dans la condition d'écoute pure, les auteurs n'ont pas observé la même tendance que dans les deux autres conditions. Ce résultat peut être dû à la présence importante dans la condition de l'écoute pure d'autres types de transformations (notamment d'analyse en flux /pə/ et /s/ ou /sə/ et /p/) par rapport au nombre de transformations de type resegmentation.

Le rôle de la cohérence articulatoire dans l'émergence des transformations verbales en condition d'écoute pure a cependant été démontré par Sato *et al.* (2007) en utilisant les stimuli de type CVCV (C : consonne, V : voyelle). Ces stimuli ont été construits avec les consonnes /p/ (consonne labiale) et /t/ (consonne coronale) et les voyelles /a/, /i/ et /o/, i.e. /pata/, /tapa/, /piti/, /tipi/, /poto/ et /topo/. L'objectif de cette étude était de vérifier s'il existe une asymétrie perceptive des transformations en faveur des formes Labial-Coronal. En effet, il existe une tendance en faveur des formes Labial-Coronal dans les langues du monde (MacNeilage et Davis, 2000) et lors de la production des premiers mots chez le jeune enfant (Davis et MacNeilage, 2003) (ex. la structure de type /bada/ est plus fréquente que /daba/). Cet effet, appelé l'effet LC, semble avoir une explication articulatoire : les séquences LC peuvent être réalisées en un seule cycle de mâchoire grâce à l'anticipation du geste de la consonne coronale, ce qui n'est pas possible pour les séquences CL (Rochet-Capellan et Schwartz, 2007). Une asymétrie en faveur des séquences LC dans une tâche de perception pure pourrait mettre en évidence le rôle des contraintes articulatoires dans l'émergence des transformations verbales. Le résultat de l'étude réalisée par Sato *et al.* (2007) montre que l'organisation des transformations est majoritairement celle d'un couplage par paires entre la séquence répétée et sa forme inverse (par exemple, les transformations /pata/ et /tapa/ lorsque la séquence répétée était /pata/). Les auteurs ont également observé que les transformations de type LC sont plus stables que celles de type CL. Ils proposent que les sujets réaliseraient

une segmentation perceptive de flux (...)LCLC(...) en unités LC en accord avec le pattern de mâchoire proposé dans le cadre de l'effet LC.

Un autre élément en faveur de l'implication des contraintes articulatoires dans l'effet de transformation verbale est apporté par les études sur la neuroanatomie fonctionnelle des transformations verbales. Dans une expérience IRMf, Sato *et al.* (2004) ont utilisé un paradigme de type bloc avec deux conditions pour mettre en évidence les régions impliquées dans l'effet de transformation verbale : la condition de base qui consistait en une répétition mentale des stimuli et la condition de transformation verbale qui consistait en une répétition des mêmes stimuli avec recherche active de transformations. La figure 3.13 illustre le réseau proposé par les auteurs comme étant en charge des transformations verbales. Le contraste entre la condition de transformation verbale et la condition de base montre l'implication d'un réseau avec une dominance gauche comprenant le gyrus temporal supérieur, le gyrus supramarginal et le gyrus frontal inférieur. Ce réseau est compatible avec le circuit dorsal de traitement de parole proposé par Hickok et Poeppel qui réalise une correspondance entre les représentations auditives et les représentations articulatoires (voir sous-section 2.2.2).

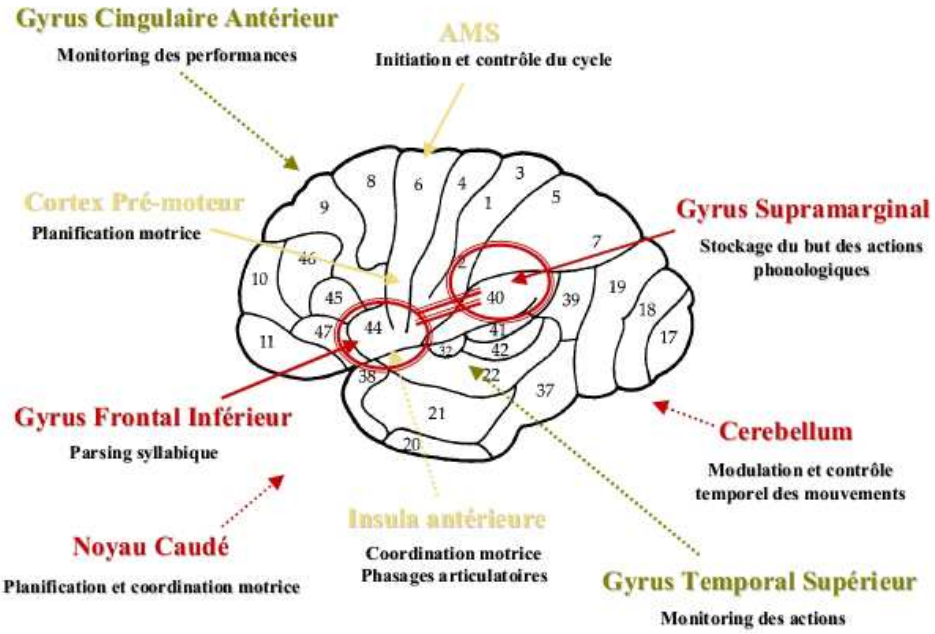


FIGURE 3.13 : Le réseau cérébral à la base des transformations verbales proposé par Sato *et al.* (2004). Figure tirée de Sato (2004).

Sato *et al.* (2004) proposent que le gyrus supramarginal serait impliqué dans la construction des représentations phonologiques. Ces représentations seraient recodées et envoyées au cortex prémoteur et au gyrus frontal inférieur puis renvoyées en retour au gyrus supramarginal lors de l'émergence d'une nouvelle transformation. Les auteurs proposent également que l'émergence des nouvelles représentations se-

rait basée sur les mécanismes de parsing syllabique (resegmentation) dans l'aire de Broca et les mécanismes de compétitions entre les représentations dans le cortex cingulaire antérieur. Nous reviendrons sur ce point dans la sous-section 3.3.3.

L'étude IRMf réalisée par Kondo et Kashino (2007) montre également l'implication des aires frontales en lien avec la production de la parole dans l'effet de transformation verbale. Dans cette étude, les auteurs ont utilisé un paradigme événementiel avec deux conditions, condition de détection du son et condition de transformation verbale. Ils ont observé que le cortex frontal inférieur gauche, le cortex cingulaire antérieur et le cortex préfrontal gauche étaient actifs seulement pendant la condition de transformation verbale. La figure 3.14 illustre l'intensité du signal observé en fonction du nombre de transformations verbales signalé par les sujets dans le cortex frontal inférieur gauche et le cortex cingulaire antérieur et en fonction du nombre de formes différentes perçues par les sujets dans le cortex préforntal gauche.

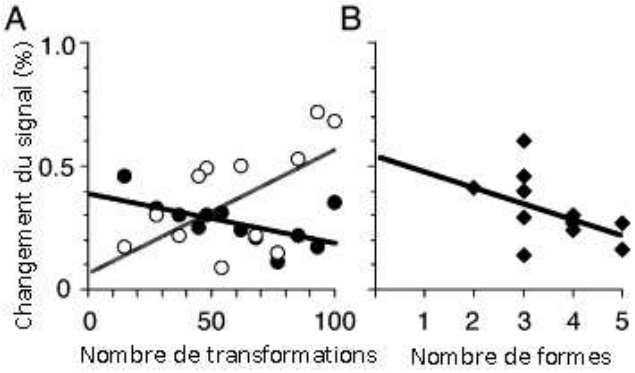


FIGURE 3.14 : La corrélation entre le nombre de transformations et le nombre de formes perçues et l'activation cérébrale observée par Kondo et Kashino (2007). (A) l'intensité du signal en fonction du nombre de transformations dans le cortex cingulaire antérieur gauche (cercles noirs) et le cortex frontal inférieur gauche (cercles blancs). (B) l'intensité du signal en fonction du nombre de formes différentes de transformations perçues dans le cortex préfrontal gauche. Figure tirée de Kondo et Kashino (2007).

Comme illustré sur la figure 3.14, à gauche, le nombre de transformations augmente avec l'augmentation de l'activation dans le cortex frontal inférieur gauche. L'activation de cette région, nous l'avons vu dans la section 2.2, est associée avec les représentations articulatoires de la parole. Les auteurs ont conclu que les transformations sont générées par des prédictions basées sur les gestes articulatoires et qu'elles sont mises à jour en permanence dans le cortex frontal inférieur gauche. En revanche, le nombre de transformations diminue en fonction de l'augmentation de l'activation dans le cortex cingulaire antérieur gauche. Les auteurs proposent que la partie dorsale du cortex cingulaire antérieur gauche est impliquée dans la stabilisation des percepts. L'activation du cortex préfrontal gauche dans cette étude a été associée aux mécanismes d'accès au lexique qui sont impliqués dans l'émergence des transformations verbales de type lexical. Les auteurs ont expliqué la corrélation négative entre l'activité de cette région et le nombre de formes différentes de trans-

formations perçues en invoquant le fait que les sujets qui signalaient un nombre faible de formes différentes, percevaient plus longtemps le mot « banana » (séquence mise en boucle dans leur stimuli). Selon les auteurs, l'activité plus importante du cortex préfrontal gauche pour ces sujets refléterait ce mécanisme de stabilisation lexicale. Les auteurs suggèrent qu'outre les effets tels que l'habituation auditive, l'organisation perceptive et les effets lexicaux, les transformations sont également liées aux réorganisations phonétiques par les contraintes articulatoires.

3.3.3 Mécanismes de prise de décision

Les mécanismes de prise de décision perceptive ont été généralement étudiés dans les tâches d'identification et de discrimination des objets visuels. Une autre catégorie de tâches pouvant mettre en évidence les mécanismes de prise de décision sont les tâches de multistabilité perceptive où la prise de décision concerne les bascules perceptives (Kast, 2001). De ce point de vue, l'activation du cortex cingulaire antérieur observée dans les tâches de transformation verbale pourrait refléter les mécanismes de prise de décision et, plus spécialement, ceux de compétition entre des représentations susceptibles d'émerger. Dans une étude IRMf, Carter *et al.* (1998) ont observé l'activation du cortex cingulaire antérieur dans les conditions qui impliquaient une compétition accrue entre les différentes réponses possibles. Une autre région pouvant être reliée aux mécanismes de prise de décision est le cervelet qui joue le rôle d'un système de contrôle articulatoire dans les tâches de mémoire de travail (Desmond *et al.*, 1997). Ainsi, Sato *et al.* (2004) suggèrent que le couplage fonctionnel entre, d'une part, le réseau temporo-pariéto-frontal présenté précédemment et, d'autre part, les régions du cortex cingulaire antérieur et du cervelet serait à la base de l'émergence des transformations verbales.

Le couplage partiéto-frontal observé dans les tâches de transformation verbale est également compatible avec les données sur les corrélats neuronaux de la prise de décision perceptive simple. Les études réalisées sur des singes montrent que les neurones dans la partie dorsolatérale du cortex préfrontal (Kim et Shadlen, 1999) et le cortex pariétal (LIP) (Shadlen et Newsome, 2001) sont impliqués dans la prise de décision (par exemple, lors d'une tâche de décision sur la direction du mouvement d'une série de points se déplaçant d'une façon aléatoire). Ces études proposent que la décision perceptive calculée par les aires préfrontales et pariétales est basée sur la différence entre le taux de décharge des neurones dans les aires sensorielles représentant les choix possibles (par exemple, les neurones sensibles au mouvement vers le haut et ceux sensibles à un mouvement vers le bas dans une tâche de décision sur la direction d'une série de points). La figure 3.15 illustre le circuit cérébral possible de la prise de décision perceptive simple (Platt, 2002). Selon ce circuit, les régions pariétales et préfrontales accumulent les informations basées sur les entrées sensorielles afin de pouvoir favoriser une alternative ou l'autre en calculant la vraisemblance de chaque choix. Les structures motrices produisent ensuite une réponse comportementale basée sur ce calcul. Les neurones dans les ganglions de la base, le cortex orbitofrontal et le cortex cingulaire réalisent une évaluation de cette sortie comportementale qui sera encodée et ensuite transférée au cortex pariétal et

préfrontal pour calculer la décision suivante.

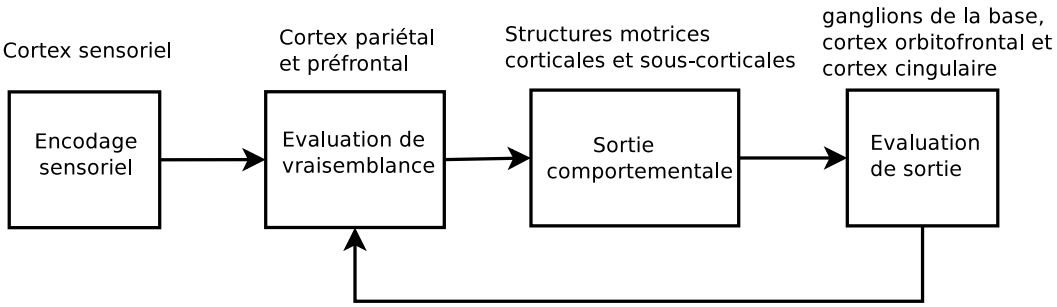


FIGURE 3.15 : Le réseau cérébral de prise de décision perceptive simple.

L'implication du cortex préfrontal a été également observée chez l'homme dans une expérience IRMf en lien avec la prise de décision. Heekeren *et al.* (2004) ont observé que l'activité de la partie dorsolatérale du cortex préfrontal gauche est plus importante pour les décisions faciles où il existe beaucoup d'apports sensoriels pour la catégorie donnée (choix entre un visage et une maison) par rapport aux décisions plus difficiles (choix entre les mêmes objets à partir des images bruitées). De plus, ils ont observé que l'activation du cortex préfrontal est corrélée avec la différence entre l'activation des régions du cortex temporal qui sont sélectives à l'image du visage et celle de la maison (pour une revue sur les corrélats neuronaux de la prise de décision, voir Kast (2001) et Platt (2002) et voir Smith et Ratcliff (2004) pour une revue sur les modèles mathématiques de la prise de décision simple).

En accord avec les études présentées ci-dessus, Leopold et Logothetis (1999) suggèrent que la réorganisation perceptive est assurée par les aires pariéto-frontales de l'intégration sensori-motrice aussi bien dans la perception normale que dans la perception multistable. Ces mêmes aires seraient responsables de l'envoi des commandes aux structures motrices qui pourraient venir en aide à la perception (par exemple, pour la saccade en réponse à une cible visuelle). Selon les auteurs, cette coordination sensorimotrice est nécessaire pour la conscience perceptive de l'environnement.

3.4 Conclusion

Dans le cadre du paradigme de multistabilité perceptive, les transformations verbales permettent donc d'étudier l'objet parole sous l'angle des processus à la base de son émergence : le liage perceptif, les bascules de liage entre différentes représentations et la prise de décision. Ces mécanismes en jeu dans la perception semblent être en partie similaires d'une modalité à l'autre. Les études sur l'effet de transformation verbale permettent à la fois de tester ces mécanismes généraux en parole et de comprendre leur éventuelle spécificité liée à la nature de l'objet parole. Ce paradigme peut nous permettre de mieux comprendre les interactions entre différentes modalités, voire entre différents domaines de la cognition, et conduire ainsi à mieux comprendre, plus généralement, la perception, et dans notre cas, la perception de la parole.

Deuxième partie

Expériences

Projet expérimental

À la lumière de ce qui précède, l'objectif de cette thèse était d'utiliser le paradigme des transformations verbales pour mieux comprendre la nature des processus de liage à l'œuvre pour la parole, entre primitives perceptuo-motrices multisensorielles et schémas phonétiques/linguistiques. Pour cela, un ensemble d'études comportementales et neurophysiologiques a été réalisé afin de répondre à certaines de nos questions portant notamment sur les processus perceptuo-moteurs et multisensoriels susceptibles de mettre en forme l'objet parole.

Au cours des quatre expériences comportementales présentées dans le premier chapitre de cette partie, nous nous sommes concentrés sur l'aspect multisensoriel des transformations verbales, ce qui n'a jamais été traité auparavant dans la littérature. Notre objectif était d'étudier l'influence de la modalité visuelle sur l'émergence et la stabilité des percepts phonétiques. Plus précisément, les questions que nous nous sommes posées dans la première expérience étaient les suivantes : est-ce que les informations visuelles sur les gestes articulatoires du locuteur peuvent influencer les transformations verbales ? Si oui, est-ce que des informations visuelles congruentes avec le signal auditif stabilisent les percepts cohérents avec la séquence auditive ? Au contraire, est-ce que des informations visuelles incongruentes entraînent la déstabilisation des percepts cohérents avec la séquence auditive ? Dans les deux cas, est-ce que cette stabilisation (déstabilisation) est accompagnée par la déstabilisation (stabilisation) du percept concurrent, cohérent avec la séquence visuelle ? Est-ce que les transformations verbales ont lieu pendant une présentation purement visuelle (en absence du flux auditif) ?

Si la modalité visuelle influence l'effet de transformation verbale, la question suivante porte naturellement sur les mécanismes à la base de cette influence, ce qui a fait l'objet de la suite de nos expériences comportementales. Ainsi, notre premier objectif était de vérifier la capacité des informations visuelles de contrôler, au cours du temps, l'émergence et la stabilité des transformations verbales. Pour cela, nous avons utilisé dans l'expérience 2 un flux visuel présentant des changements entre la séquence visuelle congruente ou non par rapport à la séquence auditive, tout en gardant le flux auditif stable.

La réponse à la question décrite dans le paragraphe précédent nous amène à une nouvelle question : comment les informations visuelles conduisent aux bascules perceptives en parole ? Notre hypothèse consiste en l'implication de l'onset visuel du geste d'ouverture de la mâchoire dans l'induction des nouveaux percepts, ce qui a été testé pendant l'expérience 3.

Enfin, une question complémentaire à celles présentées ci-dessus concerne la nature des informations visuelles ayant la capacité d'influencer l'émergence et la stabilité des transformations verbales. Autrement dit, il est intéressant de savoir si cette capacité est spécifique aux informations visuelles de type parole (gestes articulatoires) ou non. Pour cela, nous avons testé au cours de l'expérience 4, l'influence des flux visuels non-parole simulant les gestes d'ouverture de la mâchoire sur les

transformations verbales.

En accord avec la littérature sur la perception de la parole et particulièrement avec les travaux de M. Sato sur les transformations verbales (voir chapitre 2 et section 3.3), nous plaidons pour une nature perceptuo-motrice de l'objet parole et des transformations verbales. Dans ce sens, il serait nécessaire d'observer l'implication des aires cérébrales en lien avec les représentations articulatoires et les mécanismes sensori-moteurs dans l'émergence des nouveaux percepts de parole. Bien que les deux études IRMf sur l'effet de transformation verbale montrent une activation cérébrale cohérente avec une interprétation perceptuo-motrice des transformations (Sato *et al.*, 2004; Kondo et Kashino, 2007, voir sous-section 3.3.2), ces études ne permettent pas de préciser la dynamique temporelle des activités en lien avec l'émergence des transformations verbales. Une étude centrée sur la dynamique des transformations verbales vise à mieux comprendre les mécanismes impliqués dans les bascules perceptives en parole. Cette étude fait l'objet du deuxième chapitre de cette partie. En ce qui concerne l'acquisition des données cérébrales, nous avons utilisé la méthode EEG intracrânienne car ayant une très bonne résolution à la fois spatiale et temporelle, elle semble être très compatible avec les exigences de notre étude.

L'expérience iEEG réalisée dans le cadre de cette thèse a donc comme but la mise en évidence des activités cérébrales en lien avec l'émergence des transformations verbales. Pour cela, l'étude des activités juste avant des transformations est nécessaire. Notre prédiction est que ces activités sont présentes au sein du même réseau fonctionnel qui a été observé dans les études IRMf utilisant une tâche de transformation verbale. En se basant sur les études associant les oscillations cohérentes des neurones au liage perceptif (voir sous-section 1.1.2) et aux mécanismes de communication locale et globale entre les neurones (Fries *et al.*, 2007; Varela *et al.*, 2001), nous prédisons que les activités en lien avec les transformations verbales sont représentées par des oscillations neuronales en bande gamma.

Les résultats obtenus par ce projet expérimental peuvent apporter de nouveaux arguments en faveur de la théorie PACT (Schwartz *et al.*, 2002, 2007, voir sous-section 2.1.3), d'une part en montrant l'intervention active des informations multimodales dans l'organisation perceptive de la parole et d'autre part, en mettant en évidence le rôle des représentations articulatoires et sensori-motrices dans l'émergence des percepts parole. En outre, ces résultats expérimentaux peuvent fournir quelques éléments de modélisation computationnelle de l'effet de transformation verbale, ce qui fait l'objet de la partie III de cette thèse.

Chapitre 4

Transformations verbales audio-visuelles et rôle des onsets visuels dans le processus de liage

Sommaire

4.1 Expérience 1 : mise en évidence

Avec cette première expérience, nous cherchons à faire entrer pour la première fois le paradigme des transformations verbales (donc de la multistabilité phonétique) dans le domaine de la multimodalité, et notamment dans la modalité visuelle. De modalité marginale et en quelque sorte « réservée aux handicapés de l'audition », la vision de la parole (la « lecture labiale », ou *speech reading* dans le terme plus général proposé en anglais) a peu à peu pénétré tous les secteurs de la perception de la parole : identification segmentale, reconnaissance tonale, prosodie, lexique, attention, mémoire, émotions, développement. Il aurait été bien surprenant qu'elle

échappe à la multistabilité ! Notre objectif ici est donc, d'abord, de montrer qu'il y a dans la perception de la parole une multistabilité visuelle comme il y a une multistabilité auditive, d'en décrire la phénoménologie, et surtout de décrire comment audition et vision se combinent dans ces mécanismes de bascule. Nous avons bien évidemment en tête l'idée que nous comprendrons mieux ainsi comment fonctionne le liage des signaux de la parole, liage en toute vraisemblance multisensoriel, puisque ces signaux sont eux-mêmes multisensoriels.

Concrètement, l'objectif de cette expérience était de tester si l'on pouvait produire des transformations verbales visuelles, et si les informations visuelles sur les gestes articulatoires du locuteur pouvaient influencer les transformations verbales auditives. Pour cela, deux séquences langagières ont été présentées en boucle aux sujets en quatre modalités différentes : audio pure, vidéo pure, audio-visuelle congruente et audio-visuelle incongruente. Dans cette dernière modalité, le flux auditif consistait en une mise en boucle d'une des séquences et le flux visuel était la répétition synchrone de l'autre séquence. Une différence entre ces différentes modalités de présentation pourrait refléter l'intervention des informations visuelles dans l'effet de transformation verbale.

4.1.1 Méthode expérimentale

Sujets

Quinze personnes volontaires, neuf femmes et six hommes (moyenne d'âge ± écart-type : 27±7), ont participé à cette expérience. Ils étaient tous de langue maternelle française ne présentant pas de troubles auditifs ou articulatoire et ayant une vision normale ou corrigée. Ils se sont présentés individuellement à l'expérience sans avoir été au préalable renseignés sur l'objectif de cette étude.

Matériel phonétique

Deux séquences monosyllabiques, /psə/ et /səp/, ont été choisies parmi celles utilisées par Sato (2004). Ces séquences combinent la plosive non voisée labiale /p/ avec la fricative non voisée coronale /s/, sur une composante vocalique « neutre » /ə/. Ces deux syllabes ne sont pas dans le lexique de la langue française, ce qui minimise l'influence des effets lexicaux sur les transformations perçues (Shoaf et Pitt, 2002, voir sous-section 3.3.1). Cependant, la structure de ces séquences est phonotactiquement valide en Français. Les analyses lexicales sur la base de donnée VoCoLex (105,000 mots) (Dufour *et al.*, 2002) montrent que les valeurs de densité de voisinage[1] et de fréquence lexicale[2] sont moins importantes pour /psV/ que pour /sVp/ (respectivement, 31 vs. 59 et 114 vs. 371) (pour plus de détails, voir Sato, 2004).

[1] Nombre de mots phonologiquement similaires à une séquence donnée pouvant être générés par le remplacement, l'ajout ou la suppression d'un des phonèmes de la séquence.

[2] Nombre de mots incorporant, quelle que soit sa position dans le mot, une structure syllabique identique à celle d'une séquence donnée.

Stimuli

Pour construire les stimuli, nous avons utilisé les enregistrements des séquences /psə/ et /səp/ qui ont été réalisés dans le cadre de la thèse de M. Sato à l'Institut de la Communication Parlée (actuellement département Parole et Cognition du Gipsa-lab) (Sato, 2004). La technique d'acquisition audio-vidéo utilisée (Lallouache, 1990) consistait en un enregistrement synchrone des deux signaux dans une chambre sourde lors de la production des séquences par un locuteur de langue maternelle française (Jean-Luc Schwartz). Dans cette mise en place, il était prévu que les lèvres du locuteur soient maquillées en bleu. Ainsi, les paramètres labiaux du locuteur (le degré d'aperture, de protrusion et l'aire aux lèvres) lors de la production des séquences ont été extraits par une technique de chroma-key filtrant les composantes bleues présentes dans l'image. Le débit de la répétition était d'environ 1 cycle par seconde. Le taux d'échantillonnage de la vidéo était de 25 images par seconde avec une résolution de 720×576 pixels. Le signal audio était échantillonné à 44.1 kHz (codage sur 16 bits).

Nous avons choisi une séquence /psə/ et une séquence /səp/ parmi celles produites par le locuteur de sorte à satisfaire les trois critères suivants. Premièrement, nous avons vérifié que la première et la dernière image de chaque séquence étaient très similaires et qu'elles correspondaient à une position mi-ouverte des lèvres. Deuxièmement, nous avons vérifié que la durée de chaque séquence était de 640 ms (16 images) et que l'onset consonantique de la consonne labiale /p/ dans /psə/ (explosion bilabiale) et celui de la consonne coronale /s/ dans /səp/ (début de friction coronale) était alignés et tous deux aient lieu dans la cinquième image. Enfin, nous avons vérifié que l'intensité des deux séquences était similaire. La figure 4.1 illustre les caractéristiques acoustiques des séquences sélectionnées /psə/ et /səp/ et la variation de l'aire aux lèvres correspondant à la production de ces séquences. Les analyses acoustiques ont été réalisées à l'aide du logiciel Praat (Boersma et Weenink, 2001). L'aire aux lèvres a été calculée grâce à la technique d'acquisition décrite ci-dessus (Lallouache, 1990).

À l'aide du logiciel Adobe Premiere (www.adobe.com), nous avons construit quatre stimuli par séquence auditive correspondant aux différentes modalités de présentation utilisées dans cette expérience : audio pure (A), vidéo pure (V), audio-visuelle congruente (AV) et audio-visuelle incongruente (AVi). La durée de chaque stimulus était de 96 secondes (150 répétitions de /psə/ ou /səp/). Dans la modalité V, nous avons ajouté au début de chaque stimulus trois répétitions de la séquence /psə/ ou /səp/ d'une manière synchrone avec les séquences visuelles afin de permettre une catégorisation initiale du stimulus par les sujets. En ce qui concerne la modalité AVi, nous avons monté la séquence auditive /psə/ sur la séquence visuelle /səp/ et vice-versa. Le critère de sélection de ces séquences décrit ci-dessus (voir figure 4.1) a permis la synchronisation de l'onset auditif de /psə/ (ou /səp/) avec l'onset visuel de /səp/ (ou /psə/).

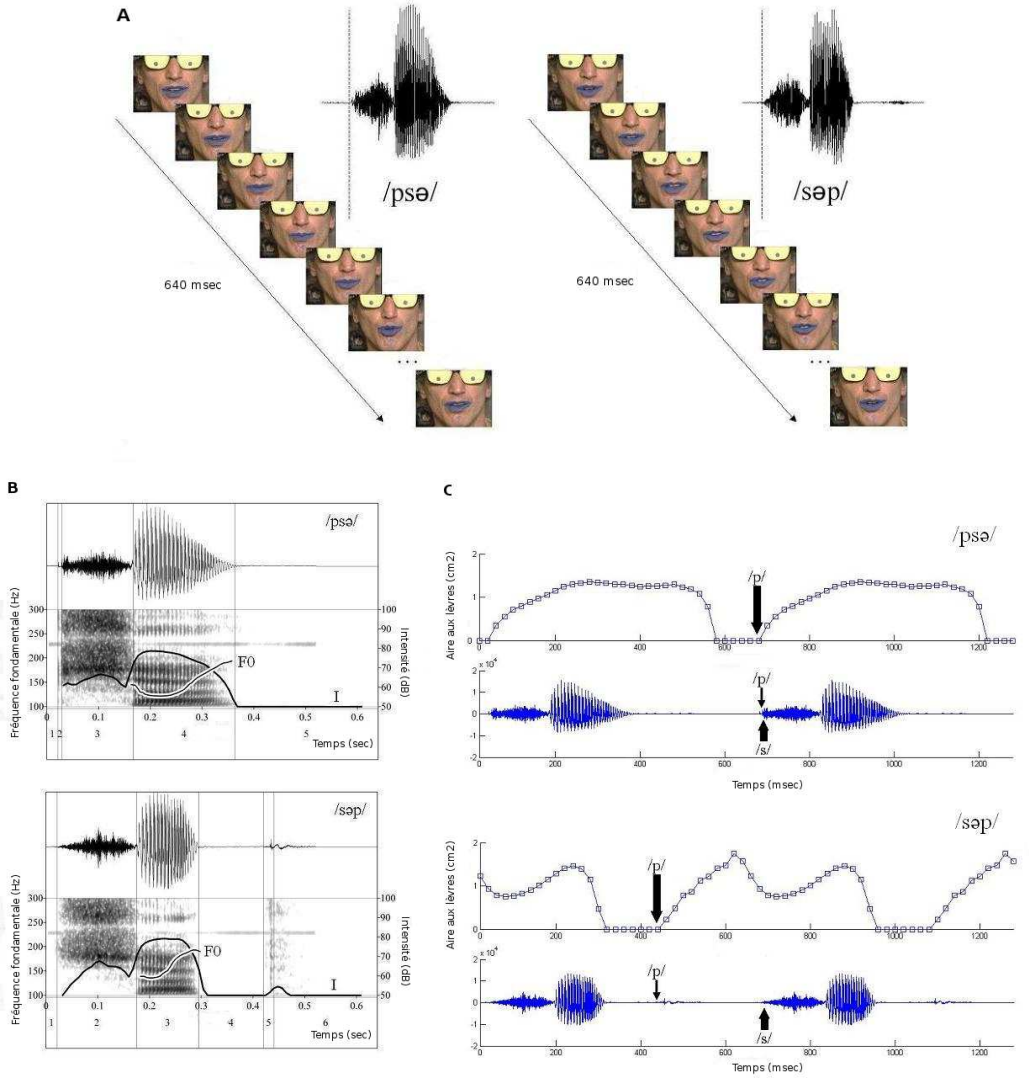


FIGURE 4.1 : Expérience 1 : caractéristiques des stimuli. A : l'onset consonantique de /p/ dans /psə/ et celui de /s/ dans /səp/ sont alignés et tous les deux ont lieu dans la cinquième image. B : le signal acoustique et le spectrogramme des séquences /psə/ et /səp/. F0 : fréquence fondamentale, I : intensité. C : la variation de l'aire aux lèvres correspondant à la production des séquences /psə/ et /səp/. Les flèches représentent les événements acoustiques ou articulatoires correspondant aux consonnes /p/ et /s/. Les flèches plus larges représentent les événements ayant tendance à se lier l'un à l'autre dans un flux audio-visuel (voir texte).

Procédure

L'expérience s'est déroulée dans une chambre sourde. Les sujets ont reçu une brève description de l'effet de transformation verbale. La consigne donnée aux sujets consistait à écouter et/ou regarder chaque stimulus et à signaler au début de chaque stimulus ce qu'ils percevaient et à signaler ensuite oralement les éventuelles transformations perçues tout au long de la présentation. Il était également indiqué que les changements perceptifs pouvaient être subtils ou importants, et correspondre aussi bien à des mots qu'à des non mots (« pseudo-mots »). Les sujets étaient enfin informés qu'il n'y avait pas de réponse correcte ou incorrecte à cette tâche. Les stimuli ont été présentés aux sujets à l'aide d'un ordinateur muni d'un écran 19-pouce. La distance entre les sujets et l'écran était d'environ 60 cm. Les signaux auditifs ont été présentés à l'oreille gauche et droite des sujets à l'aide d'un casque (stimulation binaurale). Les réponses ont été enregistrées directement sous format numérique en utilisant un microphone. Huit stimuli (2 séquences auditives × 4 modalités) ont été présentés à chaque sujet avec un intervalle de 8 secondes entre deux stimuli successifs. L'ordre de présentation des stimuli était contrebalancé entre les sujets.

Analyse des données

Nous avons étiqueté les réponses enregistrées de chaque sujet à l'aide du logiciel Praat (Boersma et Weenink, 2001). Nous avons ensuite calculé pour chaque séquence auditive (/psə/ et /səp) et chaque modalité (A, V, AV et AVi), le nombre de transformations et la durée de la stabilité des percepts. Les transformations signalées ont été classées en trois catégories : percept /psə/, percept /səp/ et autres percepts (« autres »). Nous avons ensuite calculé la durée globale de la stabilité de chaque classe divisée par 96 secondes, la durée totale des stimuli. La durée globale de stabilité d'un percept correspondait à la somme des intervalles pendant lesquels le percept était stable (avant une transformation). La différence entre la durée globale relative de percept /psə/ et percept /səp/ a été également calculée pour chaque condition, ce que nous appelons la valeur *delta*. *Delta* est défini comme la différence entre le temps passé dans le percept théorique (définissant la séquence) et le temps passé dans le percept concurrent (équation 4.1). Ainsi, lorsque la séquence auditive était /psə/, la valeur *delta* correspondait à la durée globale relative du percept /psə/ moins la durée globale relative du percept /səp/. Inversement, pour les stimuli dont la séquence auditive était /səp/, la valeur *delta* était calculée en soustrayant la durée globale relative du percept /psə/ à celle du percept /səp/.

$$delta = \begin{cases} \frac{T_{\text{psə}} - T_{\text{səp}}}{\text{durée totale}} & \text{si la séquence auditive est /psə/} \\ \\ \frac{T_{\text{səp}} - T_{\text{psə}}}{\text{durée totale}} & \text{si la séquence auditive est /səp/} \end{cases} \tag{4.1}$$

Afin de tester l'éventuelle différence entre les conditions de présentation de chaque séquence, nous avons procédé à deux ANOVA à mesure répétée et à deux facteurs (2 séquences auditives × 4 modalités), l'une sur le nombre de transformations et l'autre sur la valeur *delta*. Le seuil de signification était fixé à $p < .05$.

4.1.2 Résultats

Nombre de transformations

Le tableau 4.1 présente les moyennes du nombre de transformations rapportées par les sujets pour les deux séquences auditives et pendant les quatre modalités de présentation. L'ANOVA montre un effet significatif de la modalité [$F(3, 42) = 4.5$, $p < .03$], un effet non-significatif de la séquence auditive [$F(1, 14) = .09$] et de l'interaction entre la séquences auditive et la modalité [$F(3, 42) = 1.46$]. Les analyses post-hoc de Newman-Keuls montrent que le nombre de transformations dans la modalité AVi est significativement plus important que dans la modalité V ($p < .005$).

TABLE 4.1 : Expérience 1 : moyennes du nombre de transformations signalées (M) et écart-types correspondants (ET) pour les différentes modalités de présentation (A, V, AV, AVi) et les deux séquences auditives /psə/ et /səp/.

	/psə/		/səp/	
Modalité	M	ET	M	ET
A	7.00	2.02	9.47	3.23
V	2.80	1.03	3.07	0.90
AV	6.67	1.84	7.73	2.44
AVi	13.67	5.22	8.07	2.20

Durées de stabilité relative et valeurs de *delta*

La figure 4.2 présente la durée moyenne de stabilité relative des percepts pour les deux séquences auditives (/psə/ et /səp/) et les quatre modalités de présentation (A, V, AV et AVi). Les percepts classés en catégorie « autres » correspondaient à différents types de transformations : la substitution ou l'insertion d'un ou plusieurs phonèmes (ex. /tsə/ et /potp/ pour la séquence auditive /psə/ et /sopo/ pour la séquence auditive /səp/), le streaming auditif (ex. /sə/ pour la séquence auditive /psə/), la transformation lexicale (ex. stop pour la séquence auditive /səp/).

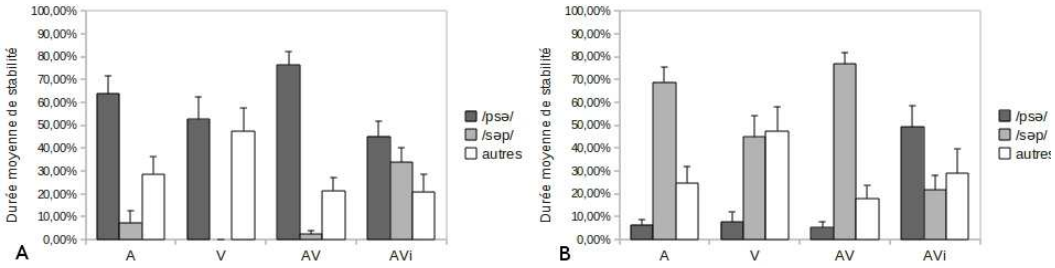


FIGURE 4.2 : Expérience 1 : moyenne de durée globale relative de stabilité des transformations perçues (/psə/, /səp/ et autres transformations) pendant les quatre modalités de présentations et pour la séquence auditive /psə/ (A) et /səp/ (B). Les barres d'erreur représentent les écart-types des moyennes.

La figure 4.3 illustre les valeurs de *delta*. L'ANOVA effectuée sur ces valeurs montre un effet significatif de la modalité de présentation [$F(3, 42) = 22.26$, $p < .0001$], de la séquence auditive [$F(1, 14) = 6.27$, $p < .03$] et de l'interaction entre ces deux facteurs [$F(3, 42) = 3.26$, $p < .04$]. Ainsi, la valeur *delta* est plus importante pour la séquence auditive /psə/ que pour /səp/. Les analyses post-hoc de Newman-Keuls montrent que la valeur *delta* dans la modalité AVi est significativement inférieure à celles des modalités A, V et AV ($p < .001$ pour les trois comparaisons). De plus, la valeur *delta* dans la modalité V était inférieure à celle de la modalité AV ($p < .04$).

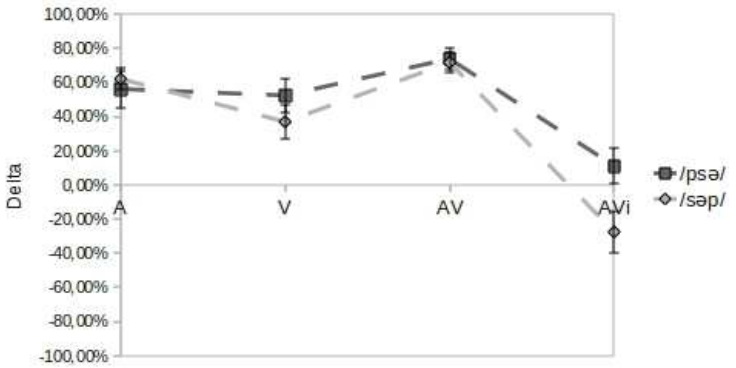


FIGURE 4.3 : Expérience 1 : moyenne des valeurs de *delta* pendant les quatre modalités de présentations et pour les séquences auditives /psə/ et /səp/. Les barres d'erreur représentent les écart-types des moyennes.

L'interaction significative entre l'effet de la séquence auditive et l'effet de la modalité de présentation est principalement due à la différence entre ces deux séquences dans la modalité AVi par rapport aux autres modalités de présentation. En effet, la séquence /psə/ s'est montré relativement stable dans la modalité AVi avec une valeur *delta* de 11% tandis que la valeur *delta* pour la séquence auditive /səp/ est de -28% ($p < .002$) (voir figure 4.3).

4.1.3 Discussion

Transformations verbales dans la modalité vidéo pure

L'expérience présentée dans cette section est la première étude qui utilise des stimuli en modalité visuelle pure dans une tâche de transformation verbale. Il est donc intéressant de souligner que l'effet de transformation verbale est également présent lors d'une présentation purement visuelle des séquences de parole. Cependant, les données sur le nombre de transformations et la durée de stabilité des percepts montrent certaines différences entre la modalité visuelle et les autres modalités de présentation utilisées. Par exemple, les transformations « autres » occupent une durée relative plus longue dans la modalité visuelle (voir figure 4.2). L'ambiguïté de la modalité visuelle pourrait être responsable de ce résultat : différents phonèmes peuvent être associés à la même forme labiale, ce qui a conduit à la perception des

transformations telles que /bdə/, /mnə/ ou /pəsə/ lorsque la séquence visuelle en boucle était /psə/. Il est important de noter que cette ambiguïté n'a pourtant pas entraîné d'augmentation du nombre de transformations, au contraire, ce nombre semble plus faible (voir table 4.1). Ce résultat suggère que les mécanismes d'exploration du nouveau percept dans la modalité visuelle sont plus lents que dans les autres modalités de présentation mais cette exploration s'effectue parmi un nombre plus grand de percepts possibles.

Influence des informations visuelles sur la durée de stabilité des percepts

La baisse significative de *delta* dans la modalité AVi par rapport aux modalités A et AV suggère que l'influence des informations visuelles sur l'effet de transformation verbale se manifesterait par la baisse de stabilité du percept cohérent avec la séquence auditive et/ou la stabilisation du percept cohérent avec la séquence visuelle. En effet, l'observation d'une séquence visuelle incongruente avec la séquence auditive entraîne une stabilité plus importante du percept cohérent avec la séquence visuelle et diminue la stabilité du percept cohérent avec la séquence auditive. Cet effet peut être expliqué par les caractéristiques acoustiques et les paramètres labiaux des séquences utilisées dans cette expérience. Nous avons présenté sur la figure 4.1 les événements de l'onset correspondant aux consonnes /p/ (burst de plosion) et /s/ (bruit de friction) qui sont quasiment synchrones dans la séquence /psə/ (/ps/ : attaque syllabique) et largement asynchrone dans la séquence /səp/ (/s/ : attaque, /p/ : coda). Dans la modalité auditive, les transformations verbales sont majoritairement de type streaming, i.e. /p/ se sépare du flux principal (voir section 3.3). Ce mécanisme conduit les sujets à signaler le percept /sə/. C'est cette séquence qui forme la plupart des percepts « autres » dans la modalité auditive illustrés sur la figure 4.2. En revanche, lorsque le geste saillant d'ouverture labiale correspondant à la consonne /p/ est visible, /p/ ne pourrait plus disparaître du flux perceptif principal. Ainsi, le percept cohérent avec la séquence audio-visuelle pourrait devenir plus stable : nous pouvons constater sur la figure 4.3 que les valeurs *delta* sont plus élévées pour la modalité AV que la modalité A. Il est cependant à noter que cette tendance n'est pas significative probablement à cause de l'effet plafond (*ceiling effect*).

En ce qui concerne la modalité AVi, l'évenement acoustique correspondant à la consonne /p/ n'est pas synchrone avec l'ouverture labiale de /p/. Il serait donc plus souvent séparé du flux principal, ce qui entraînerait le liage entre /s/ acoustique et /p/ visuel dans le flux principal (voir les flèches larges sur la figure 4.1). Plus précisément, lorsque la séquence auditive est /səp/ et la séquence visuelle est /psə/, l'onset visuel de /p/ (geste d'ouverture labiale de /p/) est synchrone avec l'onset auditif (onset de /s/), ce qui favoriserait la perception de /psə/ (voir les flèches larges correspondantes à /səp/ auditif et /psə/ visuel sur la figure 4.1). Au contraire, lorsque la séquence auditive est /psə/ et la séquence visuelle est /səp/, le geste visible d'ouverture de /p/ n'est pas proche de l'onset auditif de /s/, ce qui favoriserait la perception de /səp/ (voir les flèches larges correspondantes à /psə/ auditif et /səp/ visuel sur la figure 4.1). Ce mécanisme pourrait expliquer la stabilité plus importante des percepts cohérents avec la séquence visuelle et en conséquence la

baisse des valeurs *delta* dans la modalité AVi.

Asymétrie entre les percepts /psə/ et /səp/ dans la modalité AVi

Nous avons décrit dans la sous-section 3.3.2 les données montrant une asymétrie entre les transformations /psə/ et /səp/ (Sato *et al.*, 2006) : dans une tâche de transformation verbale, la production (ouverte et mentale) de séquence /səp/ converge souvent vers la séquence /psə/ par la synchronisation des gestes de /p/ et /s/ en début de syllabe tandis que l'effet inverse (/psə/ → /səp/) était extrêmement faible. Cette tendance, due à la cohérence articulatoire, est également présente dans la modalité AVi (voir figure 4.2). Ce résultat pourrait être dû à l'implication des processus de production mentale des gestes du locuteur dans l'intégration audio-visuelle de la parole, notamment lorsque les informations auditives sont absentes ou dégradées (Callan *et al.*, 2003). Dans ce sens, une condition conflictuelle telle que notre modalité audio-visuelle incongruente nécessiterait également l'implication des mécanismes de la production mentale. Ainsi, /p/ et /s/ seraient mentalement organisés dans un seul geste d'ouverture de la mâchoire, ce qui favoriserait une intégration audio-visuelle de type /psə/ lorsque la séquence auditive est /səp/ et la séquence visuelle est /psə/.

4.2 Expérience 2 : induction par la modalité visuelle

Ayant observé l'intervention de la modalité visuelle dans l'effet de transformation verbale, notre objectif dans la présente expérience était de tester si les informations visuelles pouvaient influencer les transformations verbales d'une manière dynamique au cours du temps, et donc en quelque sorte « contrôler » dynamiquement l'état perceptif du sujet. Pour cela, nous avons introduit des bascules visuelles dans les stimuli de sorte qu'ils basculaient de la modalité audio-visuelle congruente à la modalité audio-visuelle incongruente et vice versa. Notre hypothèse était que cette influence devrait se manifester à travers des changements perceptifs se produisant d'une manière synchrone avec les bascules dans la vidéo. L'effet de la modalité visuelle que nous avons observé dans l'expérience 1 serait ainsi encore plus grand lors de la présence des alternances visuelles dans le stimulus.

4.2.1 Méthode expérimentale

Sujets

Les sujets étaient les mêmes sujets que dans l'expérience 1.

Matériel phonétique

Nous avons utilisé le même matériel phonétique que dans l'expérience 1.

Stimuli

À partir des séquences auditives et visuelles /psə/ et /səp/, nous avons construit quatre stimuli audio-visuels différents d'une durée de 96 secondes. Le signal auditif était stable et correspondait à une mise en boucle de la séquence /psə/ ou /səp/. Le flux visuel correspondait à une alternance entre des séquences congruentes (identiques et synchrones au flux audio) et incongruentes avec la séquence auditive (voir figure 4.4). La synchronisation entre stimuli auditifs et visuels incongruents était la même que dans l'expérience 1. Il est à noter que, grâce à la similarité de la première et la dernière image des séquence /psə/ et /səp/ (voir figure 4.1), les sujets ne détectaient rien de non naturel dans ces stimuli. Ainsi, chaque stimulus était composé de plusieurs intervalles en modalité AV et AVi qui se succèdent sans interruption. La durée de ces intervalles a été variée d'une façon aléatoire afin de réduire l'effet d'apprentissage. Les durées des intervalles ont été choisies à partir d'une distribution normale entre 6 s et 12 s ou entre 11 s et 17 s. Nous avons ainsi obtenu quatre stimuli (2 séquences auditives × 2 durées d'intervalles).

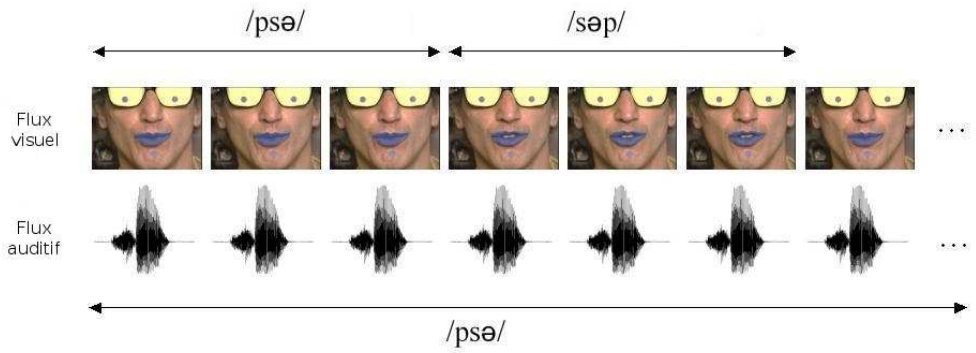


FIGURE 4.4 : Expérience 2 : schéma des stimuli utilisés. Le signal auditif est stable et correspond ici à la séquence /psə/ mise en boucle. Le flux vidéo correspond à des séquences congruentes de /psə/ suivies de séquences incongruentes de /səp/ et vice versa. Le même principe est adopté pour les séquences auditives /səp/.

Procédure

L'expérience 2 a succédé à l'expérience 1. Nous avons utilisé la même procédure que dans l'expérience 1.

Analyse des données

De la même manière que dans l'expérience 1, nous avons étiqueté les réponses des sujets en utilisant le logiciel Praat (Boersma et Weenink, 2001). Les percepts ont été catégorisés en percept /psə/, percept /səp/ et autres percepts (« autres »). Nous avons séparé, au sein de chaque stimulus, les intervalles congruents AV et incongruents AVi. Nous avons ensuite calculé la durée de stabilité globale relative des

percepts. Les valeurs *delta* ont été calculées de la même façon que dans l'expérience 1. Une ANOVA à mesure répétée à été réalisée sur les valeurs *delta*. Les différents facteurs de l'ANOVA étaient la séquence auditive (/psə/ et /səp/), la modalité (AV et AVi) et la durée des intervalles visuels (6-12 secondes et 11-17 secondes).

Afin de pouvoir tester l'effet des bascules visuelles, nous avons procédé à la comparaison de l'expérience 1 avec l'expérience 2. Notre prédiction était que l'influence de la modalité visuelle serait encore plus importante lorsqu'il existe des bascules dans le flux visuel. Autrement dit, les percepts cohérents avec la séquence visuelle seraient plus favorisés dans l'expérience 2 que dans l'expérience 1. Pour les intervalles pendant lesquels le flux auditif et le flux visuel sont congruents, la valeur de delta devrait augmenter et en revanche, pendant les intervalles incongruents, la valeur *delta* devrait baisser. Nous rappelons que la valeur *delta* est calculée par rapport à la séquence auditive et correspond à la durée globale relative du percept cohérent avec la séquence auditive moins celle du percept cohérent avec la séquence concurrente qui est dans la modalité AVi, la séquence visuelle. Ainsi, notre hypothèse était que les valeurs *delta* pendant les intervalles AV de l'expérience 2 seraient plus élévées que celles de l'expérience 1 dans la modalité AV. Au contraire, les valeurs *delta* pendant les intervalles AVi de l'expérience 2 seraient plus basses que celles de l'expérience 1 dans la modalité AVi. Pour tester cette hypothèse, nous avons effectué une ANOVA à mesure répétée à trois facteurs (2 séances expérimentales $\times$ 2 séquences auditives $\times$ 2 modalités) sur les valeurs *delta*. Le seuil de signification était fixé à $p < .05$.

4.2.2 Résultats

Synchronisation entre les bascules perceptives et les bascules visuelles

Nous avons examiné le délai entre la première transformation verbale signalée après chaque bascule visuelle et cohérente avec celle-ci. Nous avons observé une synchronie importante entre les réponses des sujets et les bascules visuelles pour les deux séquences auditives /psə/ et /səp/. En effet, 85% de ces transformations avaient lieu avec un délai de 2 secondes maximum après les bascules visuelles. Ce délai correspond au temps nécessaire pour prendre une décision (bascule perceptive) et pour préparer la réponse et la signaler oralement. Les expériences de type « speech shadowing » pendant lesquelles les sujets répétaient immédiatement ce qu'ils entendaient ont montré que la durée nécessaire pour préparer et signaler la réponse est très courte de l'ordre de 200 ms (par exemple, Marslen-Wilson, 1973; Porter et Castellanos, 1980). Prenant en compte la durée des séquences utilisées dans cette expérience (640 ms), ce résultat suggère que deux ou trois répétitions seraient nécessaires afin que la nouvelle séquence visuelle puisse induire une bascule perceptive.

Durée de la stabilité perceptive et la valeur *delta*

La figure 4.5 présente la moyenne de la durée globale relative de stabilité des percepts /psə/, /səp/ et « autres » pour les deux séquences auditives /psə/ et /səp/ et les deux modalités de présentation AV et AVi utilisées dans cette expérience. Les valeurs illustrées sur cette figure représentent la moyenne des valeurs corres-

pondantes pour les deux stimuli ayant des intervalles de bascules visuelles différents (6-12 secondes et 11-18 secondes).

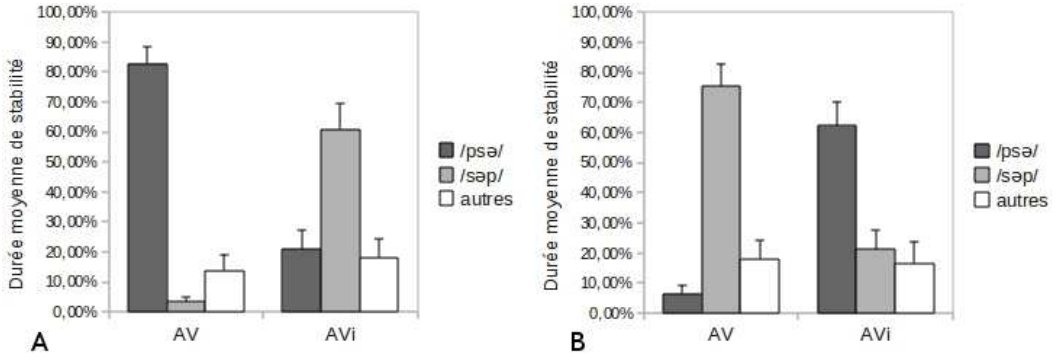


FIGURE 4.5 : Expérience 2 : moyenne de durée globale relative de stabilité des transformations perçues par les sujets (/psə/, /səp/ et autres transformations) pendant les deux modalités de présentations (audio-visuelle congruente et audio-visuelle incongruente) et pour la séquence auditive /psə/ (A) et /səp/ (B). Les barres d'erreur représentent les écart-types des moyennes.

La figure 4.6 illustre les valeurs de *delta*. L'ANOVA effectuée sur ces valeurs *delta* montre un effet significatif de la modalité de présentation [$F(1,14) = 81.44$, $p < .0001$] : la valeur *delta* est plus importante lors d'une présentation audio-visuelle congruente. L'effet de la séquence auditive et celui de la durée des intervalles visuels ne sont pas significatifs, respectivement [$F(1,14) = 1.30$] et [$F(1,14 = 0.46$]. Les interactions modalité × séquence auditive, modalité × durée d'intervalle et séquence auditive × durée d'intervalle ne sont pas non plus significatives, respectivement [$F(1,14) = 0.41$], [$F(1,14) = 2.53$] et [$F(1,14) = 2.05$]. L'interaction entre ces trois facteurs est significative [$F(1,14) = 4.82$, $p < .05$]. Cet effet semble être dû au fait que, dans la modalité AVi, la valeur *delta* pour la séquence /psə/ est très similaire pendant la vidéo 1 (intervalles de bascule de 6-12 secondes) et la vidéo 2 (intervalles de bascules de 11-18 secondes) (respectivement, -38% et -42%) tandis que pour la séquence /səp/ cette différence est assez importante (-57% pendant la vidéo 1 et -28% pendant la vidéo 2).

Expérience 1 vs. Expérience 2

L'ANOVA effectuée sur la valeur *delta* montre un effet significatif de la séance expérimentale [$F(1,14) = 11.24$, $p < .005$] : la valeur *delta* est plus élevée dans l'expérience 1 que dans l'expérience 2. L'effet de séquence auditive et la modalité de présentation étaient également significatifs, respectivement [$F(1,14) = 5.67$, $p < .04$] et [$F(1,14) = 150.20$, $p < .0001$] : la valeur *delta* était plus élevée pour la séquence auditive /psə/ que /səp/ et elle était plus élevée pendant une présentation AV par rapport à une présentation AVi. Enfin, l'interaction entre la séance expérimentale et la modalité était significative [$F(1,14 = 4.66)$, $p < .05$]. Les analyses post-hoc Newman-Keuls montrent que cette interaction significative est due

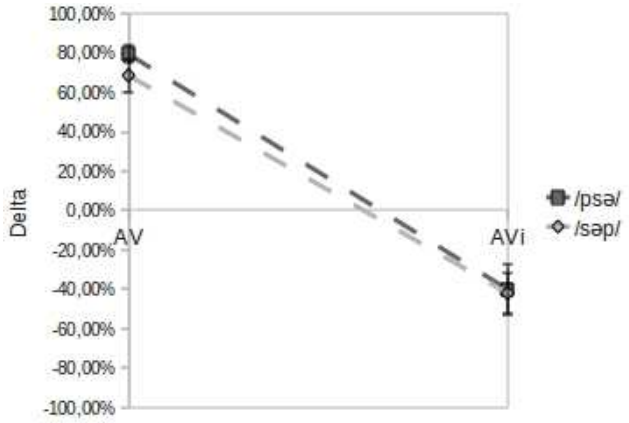


FIGURE 4.6 : Expérience 2 : moyenne des valeurs de *delta* pendant les modalités AV et AVi pour les séquences auditives /psə/ et /səp/. Les barres d'erreur représentent les écart-types des moyennes.

au fait que la valeur *delta* de la modalité AVi est moins élevée dans l'expérience 2 par rapport à l'expérience 1 (-41% vs. -8%, $p < .02$).

4.2.3 Discussion

Induction visuelle plus forte lors des alternances visuelles

La comparaison entre l'expérience 1 et l'expérience 2 a montré que la valeur *delta* est moins élevée dans l'expérience 2. Ce résultat suggère que l'influence de la modalité visuelle sur les transformations perçues est plus importante lorsqu'il y a des bascules dans la vidéo. Pour mieux comprendre cet effet, nous avons examiné les réponses des sujets en fonction de la séquence visuelle (nous rappelons que le calcul des valeurs *delta* est basé sur la séquence auditive). Nous avons constaté qu'en moyenne pendant 71% du temps de passage des stimuli les percepts des sujets étaient cohérents avec la séquence visuelle dans l'expérience 2 par rapport à 59% dans l'expérience 1. La différence est plus flagrante encore si l'on se concentre sur les stimuli incongruents. Dans l'expérience 1, les percepts des sujets étaient cohérents avec la séquence visuelle pendant 42% du temps de passage des stimuli AVi, et ils étaient cohérents avec la séquence auditive pendant 33% du temps tandis que dans l'expérience 2, ces valeurs s'élevaient à 62% contre 21% (voir figures 4.2 et 4.5). Cet effet d'induction se manifeste également par le fait que 85% des premières transformations après une bascule vidéo ont lieu avec un délai de moins de 2 secondes après la bascule.

Saillance visuelle de l'onset

Nos résultats montrent que la baisse de la valeur de *delta* dans l'expérience 2 par rapport à l'expérience 1 est particulièrement importante dans la modalité incongruente. Ce résultat pourrait être expliqué par le fait que le geste de /p/ devient

de plus en plus saillant lors des alternances visuelles. Ainsi, les mêmes arguments que ceux décrits dans l'expérience 1 en se basant sur la figure 4.1 pourraient s'appliquer à ce résultat (voir sous-section 4.1.3). Concernant cet effet de saillance, il est intéressant de rappeler que l'interaction entre la séquence auditive, la modalité de présentation et l'intervalle des bascules (6-12 ms ou 11-17 ms) dans l'expérience 2 était significative. Ce résultat suggère que cet effet de saillance pourrait dépendre de la structure temporelle de la séquence et des intervalles entre les bascules dans la vidéo. Enfin, il est à noter que les bascules visuelles dans l'expérience 2 ont éliminé la préférence pour les percepts /psə/ que nous avions observée dans l'expérience 1 dans la modalité AVi en favorisant les transformations cohérentes avec la séquence visuelle saillante aussi bien pour la séquence auditive /psə/ que /səp/.

4.3 Expérience 3 : rôle de l'onset visuel

Nous avons vu dans l'expérience 2 que la saillance de l'onset visuel pourrait être à la base de l'influence de la modalité visuelle sur les transformations verbales. Dans la présente expérience, notre objectif est d'étudier plus en détail le rôle de l'onset visuel sur l'organisation des percepts. Pour cela, nous avons choisi deux séquences dissyllabiques /pata/ et /tapa/ qui sont produites par deux gestes d'ouverture de la mâchoire avec des degrés de saillance visuelle différents : dans ce contexte, le geste labial de /p/ est plus saillant que le geste associé à /t/. Notre hypothèse est qu'en présentant les séquences auditives /...patapata.../ ou /...tapatapat.../ mais en supprimant, dans le flux vidéo, la syllabe /pa/ ou /ta/, les informations visuelles sur le geste restant conduisent à une préférence perceptive pour la séquence /pata/ (si le geste /ta/ est supprimé et que la syllabe visuelle « pa » peut donner le top de départ structurant le flux) ou pour la séquence /tapa/ (si à l'inverse le geste supprimé est celui de /pa/).

4.3.1 Méthode expérimentale

Sujets

Deux groupes de quinze personnes volontaires ont participé à cette étude présentant deux expériences, 3A et 3B. Le premier groupe a participé à l'expérience 3A et était composé de quatre femmes et onze hommes (moyenne d'âge ± écart-type : 24±4). Le deuxième groupe étaient composé de cinq femmes et dix hommes (moyenne d'âge ± écart-type : 26±4) qui ont participé à l'expérience 3B. Ils étaient tous de langue maternelle française ne présentant pas de troubles auditifs ou articulatoires et ayant une vision normale ou corrigée. Ils se sont présentés individuellement à l'expérience sans avoir été au préalable renseignés sur l'objectif de cette étude.

Matériel phonétique

Deux séquences dissyllabiques, /pata/ et /tapa/, ont été utilisées dans cette étude. Les consonnes /p/ et /t/ ont été choisies car elles sont visuellement distinguables (Summerfield, 1983) et elles sont caractérisées par une amplitude différente

de gestes d'ouverture de la mâchoire. Afin de maximiser la visibilité des gestes d'ouverture associés aux syllabes /pV/ et /tV/, nous avons choisi la voyelle /a/. Selon les analyses utilisant la base lexicale VoCoLex (Dufour *et al.*, 2002), les valeurs de densité de voisinage et de fréquence lexicale correspondant à /pata/ et /tapa/ ne sont pas très différentes (respectivement, 71 vs. 55 et 19 vs. 17).

Stimuli

Plusieurs répétitions de séquence /pata/ et /tapa/ réalisées par un locuteur de langue maternelle française (Jean-Luc Schwartz) ont été enregistrées dans une chambre sourde en utilisant la même technique d'acquisition que celle utilisée dans les expériences 1 et 2 (Lallouache, 1990). Le débit de la répétition était d'environ 520 ms par dissyllabique. Le taux d'échantillonnage de la vidéo était de 25 images par seconde avec une résolution de 720×576 pixels. Le signal audio était échantillonné à 44.1 kHz (codage sur 16 bits).

Pour construire les stimuli de l'expérience 3A, une séquence /pata/ a été sélectionnée de sorte que les deux syllabes /pa/ et /ta/ aient des caractéristiques acoustiques similaires visant à ne fournir aucun indice de segmentation *a priori*. Les caractéristiques contrôlées étaient l'intensité, la fréquence fondamentale, la durée et le degré d'ouverture labiale correspondant aux noyaux vocaliques après /p/ et /t/. La séquence /tapa/ a été construite ensuite en renversant l'ordre des syllabes /pa/ et /ta/. Pour éviter un éventuel biais perceptif en faveur du percept /pata/ dû à cette inversion d'ordre, nous avons construit pour l'expérience 3B des stimuli basés sur une séquence /tapa/. Pour cela, une séquence /tapa/ a été choisie avec les mêmes critères que dans l'expérience 3A. La séquence /pata/ a été ensuite construite par l'inversion de l'ordre des syllabes /pa/ et /ta/. Les analyses acoustiques ont été réalisées avec le logiciel Praat (Boersma et Weenink, 2001). Les analyses sur les mouvements des lèvres ont été réalisées grâce à la technique d'acquisition et au logiciel correspondant mis en place au département Parole et Cognition au Gipsa-lab (Lallouache, 1990).

Pour masquer la syllabe /pa/ ou /ta/ dans les vidéos, nous avons remplacé ces syllabes par les images autour de la voyelle /a/ de la syllabe précédente en les stabilisant tout au long de la syllabe masquée. Ainsi, pour chaque expérience, deux séquences visuelles notées /pa#a/ et /ta#a/ (# : absence de consonne) ont été construites. La figure 4.7 illustre, d'une façon schématique, l'organisation temporelle de degré d'ouverture des lèvres pour les séquences /pa#a/ et /ta#a/ en partant de la séquence /pata/. La durée de la voyelle /a/ stabilisée était de 360 ms dans les deux séquences /pa#a/ et /ta#a/.

Pour chaque séquence auditive (/pata/ et /tapa/), quatre stimuli ont été construits correspondant aux différentes modalités de présentation utilisées dans l'expérience 3A et 3B : audio pure (A), audio-visuelle congruente (AV), audio-visuelle /pa#a/ (AV-pa) et audio-visuelle /ta#a/ (AV-ta). Chaque stimuli consistait en 150 répétitions de la séquence dissyllabique correspondante. Pour les stimuli AV-pa et AV-ta, le geste visuel de /p/ et /t/ était toujours synchrone avec le son correspondant dans le signal auditif (voir figure 4.8).

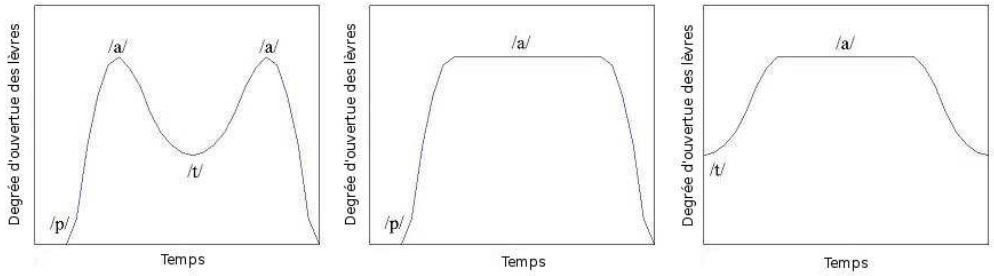


FIGURE 4.7 : Expérience 3 : organisation temporelle de degré d'ouverture des lèvres pour les séquences /pata/ (A), /pa#a/ (B) et /ta#a/ (C).

Modalité AV-pa		Modalité AV-ta	
Flux auditif	p a t a p a t a ...	Flux auditif	p a t a p a t a ...
Flux visuel	p a # a p a # a ...	Flux visuel	# a t a # a t a ...

FIGURE 4.8 : Expérience 3 : schéma des stimuli en modalité AV-pa et AV-ta pour le flux auditif /pata/.

Procédure

La procédure expérimentale était la même que dans l'expérience 1 et 2.

Analyse des données

Comme dans l'expérience 1 et 2, nous avons étiqueté les réponses des sujets en utilisant le logiciel Praat (Boersma et Weenink, 2001). Nous avons ensuite calculé la durée globale relative de stabilité des percepts. Les percepts ont été catégorisés en percept /pata/, percept /tapa/ et autres percepts (« autre »). L'objectif de cette expérience étant l'examen de l'éventuelle préférence pour le percept /pata/ ou /tapa/ en fonction de la séquence visuelle, nous avons utilisé un indice représentant directement la différence entre la durée globale relative de stabilité de /pata/ et celle de /tapa/, ce que nous appelons la valeur *delta*[3] :

$$delta = \frac{T_{\text{pata}} - T_{\text{tapa}}}{\text{durée totale}} \tag{4.2}$$

Notre hypothèse était qu'en modalité AV-pa, la valeur *delta* est plus élevée qu'en modalité AV-ta, avec une valeur intermédiaire de *delta* pour la modalité AV. Afin de tester cette hypothèse, nous avons effectué sur les valeurs de *delta* de l'expérience 3A et 3B une ANOVA à mesure répétée à deux facteurs : 2 séquences auditives (/pata/

[3]La valeur *delta* ainsi calculée est différente de celle utilisée dans les expériences 1 et 2 où la valeur *delta* correspondait à la durée globale relative de la stabilité du percept cohérent avec la séquence auditive moins celle du percept incohérent. Dans l'expérience 3A et 3B, indépendamment de la séquence auditive, la valeur *delta* est la durée de la stabilité globale relative du percept /pata/ moins celle du percept /tapa/.

ou /tapa/) × 4 modalités de présentation (A, AV, AV-pa et AV-ta). Le seuil de signification était fixé à $p < .05$.

4.3.2 Résultats

Durée de la stabilité perceptive et la valeur *delta*

La figure 4.9 illustre la moyenne de la durée de stabilité globale relative des percepts /pata/, /tapa/ et « autres » pour les deux séquences auditives (/pata/ et /tapa/) et pour les quatre modalités de présentation (A, AV, AV-pa et AV-ta) pendant les expériences 3A et 3B.

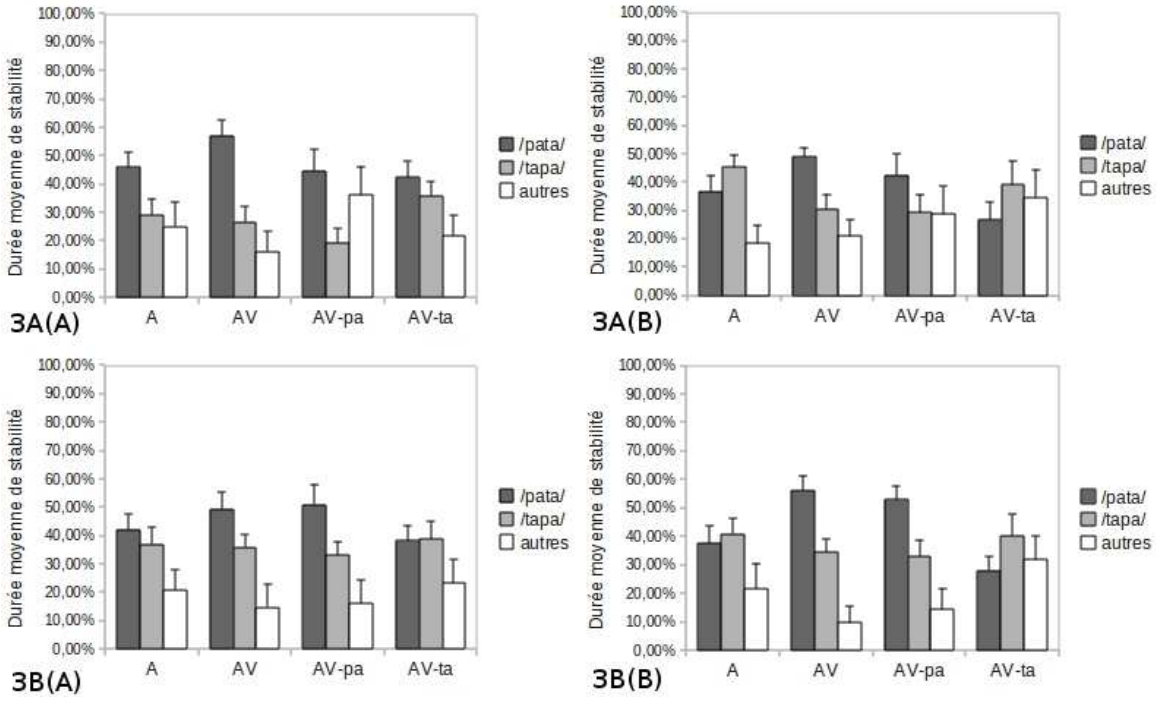


FIGURE 4.9 : Expérience 3 : moyenne de durée globale relative de stabilité des transformations perçues par les sujets (/pata/, /tapa/ et « autres ») pendant les quatre modalités de présentation (A, AV, AV-pa et AV-ta) et pour les séquences auditives /pata/ (A) et /tapa/ (B). En haut : expérience 3A et en bas : expérience 3B. Les barres d'erreur représentent les écart-types des moyennes.

La figure 4.10 illustre les valeurs de *delta*. L'ANOVA effectuée sur les valeurs *delta* de l'expérience 3A montre un effet significatif de la modalité de présentation [$F(3, 42) = 4.32$, $p < .01$] et de la séquence auditive [$F(1, 14) = 8.14$, $p < .05$] : la valeur *delta* est plus grande pour la séquence auditive /pata/ que /tapa/. L'interaction entre ces deux facteur n'était pas significative [$F(3, 42) = 0.35$]. Les analyses post-hoc de Newman-Keuls ont montré que la valeur *delta* est moins élevée dans la modalité AV-ta (-0.3%) que dans la modalité AV (24%) et AV-pa (19%) (respectivement, $p < .02$ et $p < .04$). En ce qui concerne l'expérience 3B, l'ANOVA montre un effet significatif de la modalité [$F(3, 42) = 4.53$, $p < .03$]. L'effet de la séquence

auditive et l'interaction entre les deux facteurs n'étaient pas significatifs (respectivement, $F(1,14) = 0.19$ et $F(3,24) = 1.05$). Les analyses post-hoc montrent la même tendance que dans l'expérience 3A : la valeur *delta* est moins importante dans la modalité AV-ta (-0.6%) que dans la modalité AV (18%) et AV-pa (19%) (respectivement, $p < .02$ et $p < .03$).

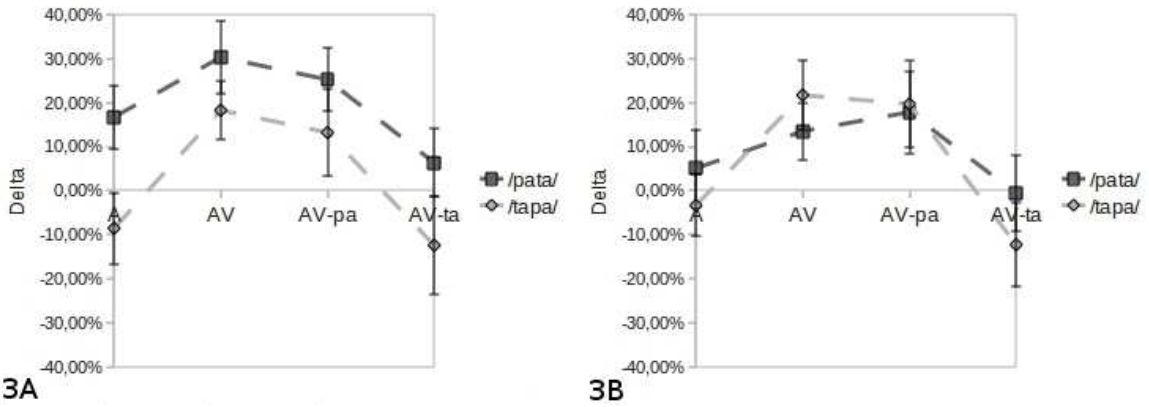


FIGURE 4.10 : Expérience 3 : moyenne des valeurs de *delta* pendant les quatre modalités de présentation pour les séquences auditives /pata/ et /tapa/. À gauche : expérience 3A et à droite : expérience 3B. Les barres d'erreur représentent les écart-types des moyennes.

Afin d'identifier l'éventuelle différence entre l'expérience 3A et 3B, nous avons effectué une ANOVA sur les valeurs *delta* avec la séquence auditive et la modalité de présentation comme facteur intra-sujet et la séance expérimentale (3A ou 3B) comme facteur inter-sujet. Cette analyse montre un effet significatif de la séquence auditive [$F(1,28) = 5.51$, $p < .03$] et de la modalité de présentation [$F(3,84) = 8.75$, $p < .0001$], mais pas d'effet de la séance expérimentale. Les analyses post-hoc montrent que la valeur *delta* est significativement moins importante pour les modalité A et AV-ta que dans la modalité AV et AV-pa ($p < .01$ pour toutes les comparaisons).

4.3.3 Discussion

La différence entre les valeurs *delta* des modalités AV-pa et AV-ta dans cette étude confirme notre hypothèse concernant le rôle de l'onset visuel sur l'émergence des transformations verbales : lorsque les sujets observaient le geste d'ouverture correspondant à /pa/, les transformations cohérentes avec la dissyllabique commençant par cette syllabe, i.e. /pata/, sont plus longtemps perçues. De la même manière, en présentant le geste de /ta/ aux sujets, les transformations de type /tapa/ sont favorisées. Ainsi, il semble se confirmer que l'onset visuel correspondant à un geste d'ouverture de la mâchoire contribue à structurer le percept dans le processus de liage audio-visuel.

Si l'induction visuelle se réalise par l'onset visuel, la valeur *delta* correspondant à la modalité AV, caractérisée par deux gestes d'ouverture CV, devrait être intermé-

diaire entre celles des modalités AV-pa et AV-ta, i.e. AV-pa > AV > AV-ta. Cette tendance a été validée pour la modalité AV et AV-ta mais pas pour la modalité AV et AV-pa. Autrement dit, les analyses présentées ci-dessus ne montrent pas de différence significative entre la valeur *delta* de AV et AV-pa. Ce résultat peut provenir du fait que le geste d'ouverte de /pa/ dans la modalité AV est beaucoup plus important que celui de /ta/. Ainsi, l'effet de l'onset visuel de /pa/ au sein de la séquence /pata/ ou /tapa/ en modalité AV serait très similaire à celui de l'onset visuel de /pa/ en modalité AV-pa (voir figure 4.7). Il est également important de noter que les valeurs *delta* dans les modalités AV-pa et AV-ta sont respectivement sous-estimées et surestimées dans notre calcul. En effet, une grande partie des transformations « autres » en modalité AV-pa et AV-ta sont respectivement de la forme /paCa/ (C : consonne autre que /t/) et /taCa/ (consonne autre que /p/). Cette substitution de /p/ et /t/ par une autre consonne (souvent par /k/) due à l'incongruence entre la séquence auditive (dissyllabique) et la vidéo (monosyllabique) a conduit d'une part à l'augmentation du nombre de transformation « autres » par rapport à la modalité AV et de l'autre, à l'exclusion des transformations qui sont cohérentes avec notre hypothèse dans notre calcul de valeur *delta* : dans ces cas, les informations visuelles conduisent à des transformations dissyllabiques commençant par /p/ en modalité AV-pa et par /t/ en modalité AV-ta, ce qui est cohérent avec notre prédiction sur le rôle de l'onset visuel.

4.4 Expérience 4 : onset visuel non-parole

Les expériences précédentes ont montré que les informations visuelles sur le geste d'ouverture de la mâchoire jouent un rôle dans l'effet de transformation verbale en conduisant à la préférence d'une forme par rapport à d'autres. Notre objectif dans cette dernière expérience était de mieux comprendre la nature de ces informations visuelles. Nous avons voulu principalement tester si l'effet de structuration des percepts par l'onset visuel était spécifique à la vision d'un geste articulatoire (en l'occurrence, un geste d'ouverture de la mâchoire et des lèvres), ou s'il s'agissait d'un mécanisme psychophysique d'alerte qui aurait pu être induit par n'importe quel signal visible. Pour cela, nous avons comparé dans cette expérience les gestes /pa/ et /ta/ de l'expérience 3 et des stimuli vidéo définis par une barre verticale qui s'ouvrait et se fermait en respectant la dynamique temporelle des gestes /pa/ et /ta/.

Nous pouvons poser *a priori* deux hypothèses antagonistes. La première est que l'effet est psychophysique, et peut être induit par n'importe quel stimulus vidéo comportant un signal d'alerte. Dans ce cas, l'effet devrait être maintenu au moins aussi efficacement en remplaçant le geste phonétique par les barres verticales. La seconde est que l'effet est spécifique à la parole, et il devrait alors disparaître ou du moins se réduire avec les barres. On peut aussi imaginer qu'il existe une combinaison de ces deux hypothèses, et dans ce cas, la corrélation entre les effets induits respectivement par les lèvres et par les barres est porteuse d'information : une corrélation forte entre les sujets serait plutôt le signe d'un mécanisme unique, une corrélation

faible, le signe de deux mécanismes complémentaires, inégalement distribués d'un sujet à l'autre.

4.4.1 Méthode expérimentale

Sujets

Dix-sept personnes volontaires, huit femmes et neuf hommes (moyenne d'âge ± écart-type : 29±7) ont participé à cette expérience. Ils étaient tous de langue maternelle française ne présentant pas de troubles auditifs ou articulatoire et ayant une vision normale ou corrigée. Ils se sont présentés individuellement à cette expérience sans avoir été au préalable renseignés sur l'objectif de cette étude.

Matériel phonétique

Nous avons utilisé le même matériel phonétique que dans l'expérience 3.

Stimuli

Nous avons ré-utilisé dans cette expérience les stimuli utilisés en modalité A, AV-pa et AV-ta dans l'expérience précédente. Pour les séquences auditives /pata/ et /tapa/, nous avons utilisé respectivement les stimuli de l'expérience 3A et 3B. Nous avons superposé à ces stimuli 4 flux vidéo. Les deux premiers étaient les mêmes que dans l'expérience 3, nous les appellerons « lèvres-pa » et « lèvres-ta » par la suite. Nous avons également construit deux nouveaux flux vidéo, en remplaçant le mouvement des lèvres par l'expansion d'une barre verticale simulant le geste d'ouverture de la mâchoire correspondant à /pa#a/ et /ta#a/ dans les stimuli AV-pa et AV-ta. Pour construire les séquences vidéos « barre », nous avons procédé de la manière suivante. Nous avons repéré les instants d'ouverture des lèvres, de l'instant correspondant à la fermeture consonantique à l'instant correspondant au noyau vocalique (ouverture maximale) au sein du /pata/ et /tapa/. Nous avons alors schématisé les gestes d'ouverture par le mouvement d'une barre rouge (sur fond noir) symétrique, synchrone avec l'ouverture labiale. Seule la synchronie des événements de début et de fin de geste d'ouverture a été prise en compte : la dynamique du mouvement de la barre a été choisie linéaire entre une valeur minimale et une valeur maximale. Les valeurs minimales (pour /p/ et /t/) et maximale (pour /a/) ont été choisies arbitrairement, de façon à maintenir une bonne visibilité des dynamiques des barres. Ainsi, les 4 images illustrés sur la figure 4.11 ont été construites.

La figure 4.12 représente la trajectoire du mouvement des barres /pa#a/ et /ta#a/ et le début de l'ouverture labiale pour /p/ et /t/ en fonction du mouvement des lèvres pour la séquence /pata/. Nous rappelons que l'objectif de cette expérience est de vérifier si les barres ayant une dynamique synchrone avec le mouvement d'ouverture des lèvres peuvent influencer les transformations verbales de la même manière que les lèvres. Nous avons donc contrôlé la hauteur des barres en choisissant les mêmes barres pour /pa#a/ et /ta#a/ de sorte que seul l'effet de la synchronisation temporelle soit testé et pas un mélange de l'effet de synchronisation

FIGURE 4.11 : Expérience 4 : images des barres verticales utilisées pour simuler la variation de degré d'ouverture des lèvres.

temporelle et le degré d'ouverture des lèvres (voir figure 4.7 pour la différence entre les trajectoires d'ouverture des lèvres pour /pa/ et /ta/). Nous reviendrons sur ce point dans la sous-section 4.4.3.

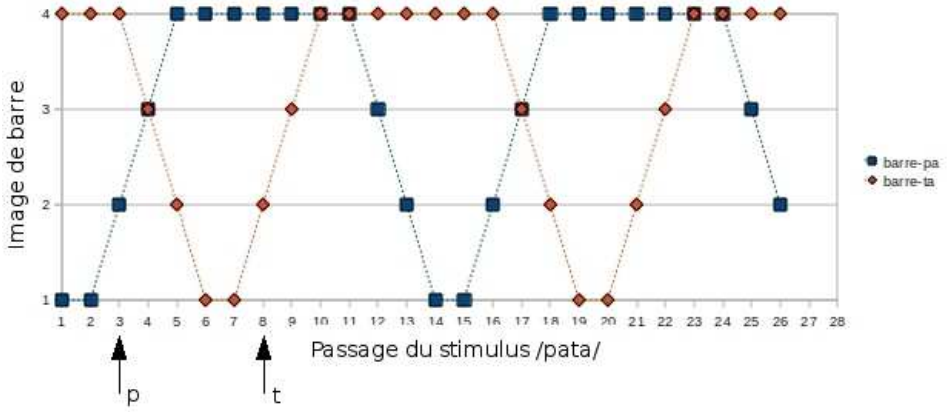


FIGURE 4.12 : Expérience 4 : trajectoires de mouvements des barres en modalité « barre-pa » et « barre-ta » pour la séquence auditive /pata/. En abscisse : le passage du stimulus /pata/ en fonction de ses images ; deux répétitions sont illustrées, chaque séquence est de 13 images et chaque image dure 40 ms. En ordonnée : les images de barres illustrées sur la figure 4.11. Les flèches représentent les images correspondant au début de l'ouverture labiale pour /p/ et /t/.

Procédure

La procédure expérimentale était la même que celle utilisée dans l'expérience 1, 2 et 3. La seule différence concernait l'acquisition des réponses des sujets. Dans cette expérience, nous avons demandé aux sujets de catégoriser eux-mêmes les transformations verbales qu'ils percevaient en /pata/, /tapa/ ou « autres » en appuyant sur les touches Q, S et D du clavier. Pour la moitié des sujets, Q correspondait à la transformation /pata/, S à « autres » et D à /tapa/. Pour l'autre moitié, les catégories correspondant aux touches Q et D étaient inversées, i.e. Q correspondait à /tapa/ et D à /pata/. Les appuis sur les touches du clavier ont été automatiquement enregistrés sur un fichier à l'aide d'un programme Visual Basic. Cette procédure d'enregistrement a permis une estimation plus simple de la valeur *delta*

car la catégorisation a été directement réalisée par les sujets. Cependant, en utilisant cette procédure d'enregistrement, nous n'avons eu aucune information sur la nature phonétique des transformations catégorisées comme « autres ». Notre objectif étant de tester les valeurs *delta*, cette absence n'a pas nui aux résultats de cette expérience.

Analyse des données

Nous avons calculé la durée globale relative de stabilité des percepts /pata/, /tapa/ et « autres » en fonction des réponses enregistrées des sujets par stimulus. Les valeurs *delta* correspondantes ont été ensuite calculées. De la même manière que dans l'expérience 3, la valeur *delta* correspondait à la différence entre la durée globale relative de stabilité de /pata/ moins celle de /tapa/ :

$$delta = \frac{T_{\text{pata}} - T_{\text{tapa}}}{\text{durée totale}} \tag{4.3}$$

L'objectif de la présente expérience était de tester si l'observation des barres, ayant la même dynamique que les lèvres, pouvait influencer les transformations verbales de la même manière que l'observation des lèvres. Pour cela, nous avons effectué une ANOVA à mesure répétée à deux facteurs : 2 séquences auditives (/pata/ ou /tapa/) $\times$ 5 modalités de présentation (A, lèvres-pa, lèvres-ta, barre-pa et barre-ta). Le seuil de signification était fixé à $p < .05$. Nous avons également étudié la corrélation entre les effets « lèvres » et « barres » pour les sujets, selon un paramètre caractéristique de ces effets, que nous définirons plus loin.

4.4.2 Résultats

Durées de stabilité perceptive et valeurs de *delta*

La figure 4.13 illustre la moyenne de la durée de stabilité globale relative des percepts /pata/, /tapa/ et « autres » pour deux séquences auditives (/pata/ et /tapa/) et cinq modalités de présentation (A, lèvres-pa, lèvres-ta, barre-pa et barre-ta) de l'expérience 4.

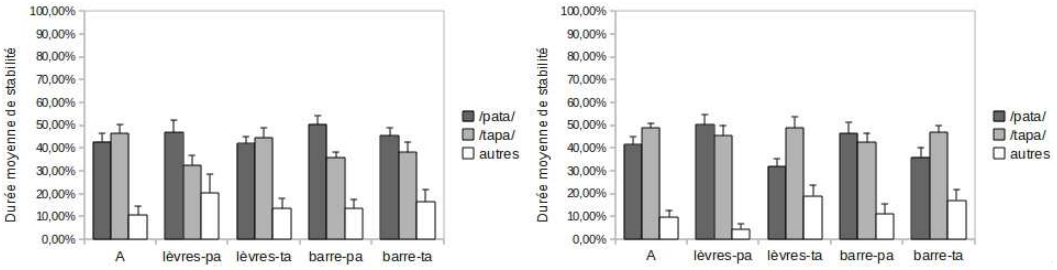


FIGURE 4.13 : Expérience 4 : moyenne de durée globale relative de stabilité des transformations perçues par les sujets (/pata/, /tapa/ et « autres ») dans les cinq modalités de présentation (A, lèvres-pa, lèvres-ta, barre-pa et barre-ta) et pour la séquence auditive /pata/ (A) et /tapa/ (B). Les barres d'erreur représentent les écart-types des moyennes.

La figure 4.14 présente les valeurs de *delta*. L'ANOVA effectuée sur ces valeurs montre un effet significatif de la modalité de présentation [$F(4, 64) = 3.86$, $p < .01$] et de la séquence auditive [$F(1, 16) = 6.20$, $p < .05$] : la valeur *delta* est plus grande pour la séquence auditive /pata/ que /tapa/. L'interaction entre ces deux facteurs n'était pas significative [$F(4, 64) = 0.54$]. Les analyses post-hoc de Newman-Keuls ont montré que les valeurs delta de la modalité lèvres-pa et barre-pa sont significativement plus grandes qu'en modalité A, lèvres-ta et barre-ta.

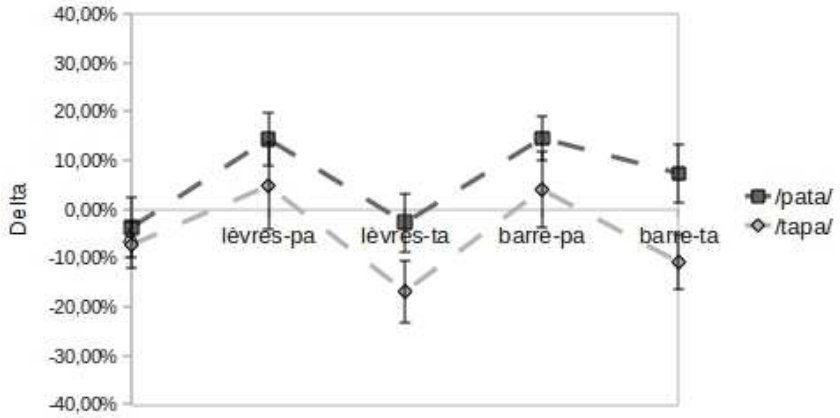


FIGURE 4.14 : Expérience 4 : moyenne des valeurs de *delta* pendant les cinq modalités de présentation pour les séquences auditives /pata/ et /tapa/. Les barres d'erreur représentent les écart-types des moyennes.

Effet de l'induction des lèvres vs. des barres

Afin de pouvoir mieux comparer l'effet de l'induction des lèvres et celui des barres, nous avons introduit un nouveau paramètre intitulé *delta2* centré sur les seuls percepts /pata/ et /tapa/. La valeur *delta2* correspond à la durée globale de stabilité du percept /pata/ moins celle du percept /tapa/, divisée par la durée globale de stabilité des percept /pata/ et /tapa/. Autrement dit, nous n'avons pas pris en compte la durée de stabilité des transformations « autres » dans ce calcul :

$$delta2 = \frac{T_{\text{pata}} - T_{\text{tapa}}}{T_{\text{pata}} + T_{\text{tapa}}} \tag{4.4}$$

La figure 4.15 illustre la moyenne des valeurs *delta2* ainsi calculée pour les modalités A, lèvres-pa, lèvres-ta, barre-pa et barre-ta.

Nous avons alors défini l'effet de l'induction des lèvres pour chaque sujet en soustrayant la valeur *delta2* correspondant de la modalité lèvres-ta à la valeur *delta2* de la modalité lèvres-pa, ce que nous appelons indice-lèvres. De la même façon, l'effet de l'induction des barres (indice-barre) correspondait à la valeur *delta2* de la modalité barre-pa moins la valeur *delta2* de la modalité barre-ta.

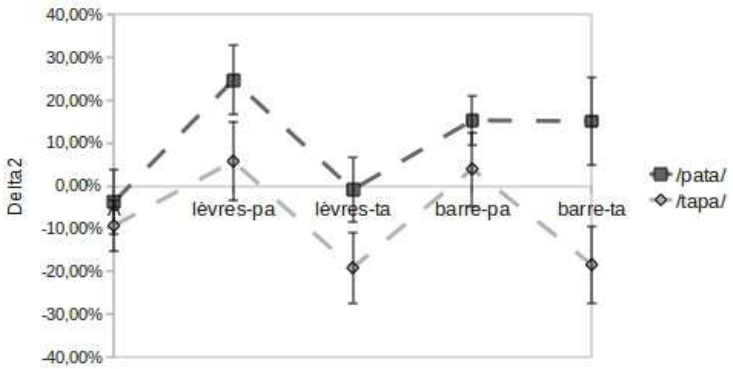


FIGURE 4.15 : Expérience 4 : moyenne des valeurs de *delta2* pour les cinq modalités de présentation pour les séquences auditives /pata/ et /tapa/. Les barres d'erreur représentent les écart-types des moyennes.

$$\text{indice-lèvres} = delta2_{\text{lèvres-pa}} - delta2_{\text{lèvres-ta}} \quad (4.5)$$

$$\text{indice-barre} = delta2_{\text{barre-pa}} - delta2_{\text{barre-ta}} \quad (4.6)$$

Une comparaison entre ces deux indices est destinée à mettre en évidence l'éventuelle différence entre la force d'induction des lèvres vs. celles des barres. Un t-test apparié a donc été réalisé sur la moyenne de l'indice-lèvres des séquences auditives /pata/ et /tapa/ et la moyenne de l'indice-barre correspondant à ces deux séquences auditives. Ce test montre que l'indice-lèvres est plus grand que l'indice-barre [$t(16) = 2.25$, $p < .05$]. Nous présentons sur la figure 4.16 la corrélation entre les deux indices pour les 17 sujets. La corrélation de Spearman entre ces deux indices est significative et assez élevée R=0.55 ($p < .01$). La figure montre bien que l'effet « lèvres » est en général près de deux fois supérieur à l'effet « barres ».

4.4.3 Discussion

Les résultats de cette expérience confirment nos résultats de l'expérience 3 concernant l'influence de l'onset visuel sur les transformations verbales. Ils montrent que la même tendance, mais moins forte, existe si les gestes articulatoires sont remplacés par les barres verticales ayant la même dynamique que les lèvres. Ceci semble plutôt valider notre seconde hypothèse : l'effet d'onset apparaît comme un effet « speech specific » et non un mécanisme psychophysique général. La corrélation entre l'effet des lèvres et l'effet des barres sur les 17 sujets pourrait être basée sur une éventuelle simulation des mouvements des lèvres par les sujets à partir des stimuli non parole (barres verticales) utilisés dans cette expérience : ainsi, le mécanisme de base serait phonétique, et les barres produiraient une émulation de ce mécanisme, avec un « rendement » inférieur à 1, expliquant la plus faible efficacité d'induction des barres comparées aux lèvres. Comme décrit dans la sous-section 2.3.2, les informations visuelles sur la cinématique du visage permettent, dans certaines tâches, une intégration audio-visuelle de la parole. Si les barres fournissent

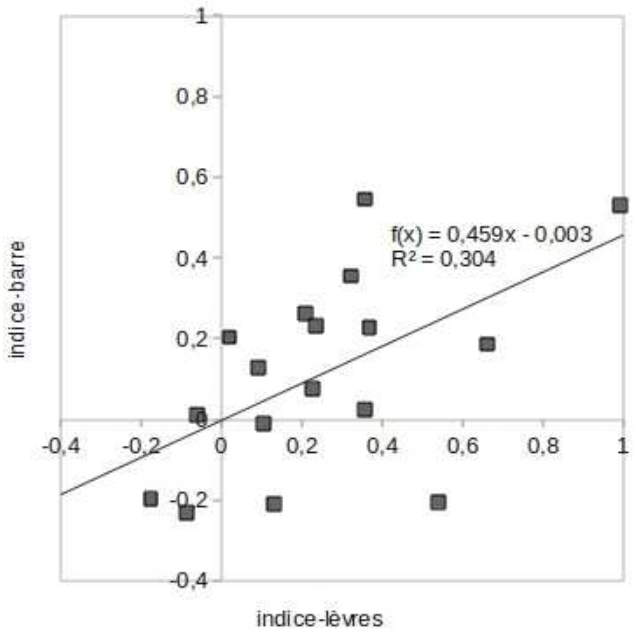


FIGURE 4.16 : Expérience 4 : corrélation entre l'indice-lèvres et l'indice-barre. La corrélation de Pearson est R=0.55 ($p < .05$).

aux sujets des mouvements comparables à ceux des gestes articulatoires, il n'est pas étonnant qu'elles puissent influencer les transformations verbales de même façon que la présentation vidéo des gestes articulatoires.

Il est important de rappeler que la différence entre la modalité audio pure et la modalité lèvres-pa, contrairement à la modalité lèvres-ta, est significative. Cette différence, probablement due à la saillance de /p/ par rapport à /t/, est également présente lors de la présentation des barres : la valeur *delta* de la modalité barre-pa est plus grande que celle de la modalité A, tandis que la valeur *delta* dans la modalité barre-ta n'est pas significativement plus faible. Les images des barres et leur trajectoires schématisant /pa#a/ et /ta#a/ étant identiques, cette différence ne peut donc pas être expliquée en terme d'un effet psychophysique tel que l'effet de la hauteur des barres. Bien qu'une interprétation purement psychophysique ne semble pas être compatible avec nos résultats sur l'apport des lèvres et l'apport moins efficace des barres, nous ne pouvons pas éliminer la possibilité d'une éventuelle interaction entre des effets psychophysiques et des effets spécifiques à la parole. Ainsi, cette interaction pourrait expliquer la corrélation non-parfaite entre l'effet des lèvres et l'effet des barres. D'autres expériences, avec des stimuli plus ciblés, pourraient être envisagées afin de vérifier les éventuels effets psychophysiques et leurs possibles interactions avec les mécanismes spécifiques à la parole. Il serait par exemple intéressant de vérifier si d'autres formes géométriques, moins similaires au patron d'ouverture des lèvres, peuvent également déclencher l'effet d'onset visuel observé dans cette expérience.

4.5 Discussion générale

Ces quatre expériences, visant à tester l'influence des informations visuelles des gestes articulatoires du locuteur sur l'émergence et la stabilité des transformations verbales, nous ont permis d'avancer dans deux directions principales. D'une part, nous avons clairement inscrit la multimodalité de la parole dans le cadre des transformations verbales, qui prennent ainsi un statut de paradigme de multistabilité multimodale extrêmement riche. Nous avons caractérisé le rôle des interactions audio-visuelles dans les transformations verbales, et ceci pourrait ouvrir la voie à des développements expérimentaux nombreux. D'autre part, nous avons poursuivi notre projet principal, qui est l'étude des mécanismes de liage en perception de parole, en caractérisant un mécanisme à notre sens nouveau, que nous avons dénommé « onset visuel ». Nous allons discuter ces résultats dans ce double contexte.

Les résultats de l'expérience 1 démontrent d'une part l'existence des transformations verbales purement visuelles et d'autre part l'influence de la modalité visuelle sur la stabilité des percepts lorsque les flux auditif et visuel sont incongruents, ceci en diminuant la durée du percept cohérent avec la séquence auditive et/ou en augmentant la durée du percept cohérent avec la séquence visuelle. Selon les résultats de l'expérience 2, la modalité visuelle pourrait, dans une certaine mesure, contrôler l'émergence des transformations au cours du passage des stimuli. Les résultats de l'expérience 3 suggèrent que cette influence pourrait être due à l'observation de l'onset visuel correspondant au geste d'ouverture de la mâchoire. L'effet moins important de l'onset visuel dans l'expérience 4, utilisant des stimuli purement géométriques, suggère que les informations visuelles sur le geste d'ouverture de la mâchoire pourraient être récupérées à partir de la dynamique temporelle du signal visuel, et donc que l'effet d'onset est bien « speech specific ».

Nous avons vu dans la section 3.1.2, les études sur la présentation discontinue des stimuli multistables visuelles qui entraînaient, en fonction de la durée de non-présentation, la stabilisation ou la déstabilisation des percepts visuels (ex. Leopold *et al.*, 2002; Kornmeier *et al.*, 2007). En ce qui concerne la parole, les résultats de l'expérience 2 montrent que les manipulations dans la modalité visuelle pourraient jouer un rôle sur la stabilisation ou déstabilisation des percepts parole : en fonction de la congruence ou l'incongruence entre les informations auditives et visuelles, les bascules dans la vidéo permettent une meilleur stabilité ou non du percept cohérent avec la séquence auditive.

Il est également intéressant de souligner que la possibilité de contrôle des transformations avec la modalité visuelle observée dans l'expérience 2 peut fournir un paradigme expérimental intéressant pour suivre les bascules perceptives dans le cerveau. Sachant que les bascules dans le flux visuel conduisent à des changements perceptifs, on pourrait ainsi espérer caractériser en ligne les activités cérébrales associées à l'émergence d'un nouveau percept sans demander une réponse de la part des sujets, ce qui faciliterait l'interprétation des activités observées en lien avec les transformations verbales.

Le rôle de l'onset visuel sur les transformations verbales observée dans l'expérience 3 peut être interprété par rapport aux propositions psycholinguistiques sur

le rôle important de l'onset des mots dans l'accès lexical (Marslen-Wilson et Welsh, 1978) et sur le rôle des syllabes fortes dans la segmentation du flux de parole (Cutler et Norris, 1988). Il est important de noter que ces propositions sont fondées exclusivement sur la modalité acoustique et sur les propriétés auditives des signaux de parole. Dans ce sens, nos résultats apportent de nouveaux arguments en faveur d'une nature multimodale de la parole et proposent que les indices visuels participent également d'une manière active dans la segmentation du flux de parole. Ce résultat est cohérent avec les travaux de Dohen *et al.* (2004) réalisées dans le département Parole et Cognition du Gipsa-lab qui montrent que des indices prosodiques, importants dans la segmentation de la parole (voir Cutler, 1996, pour une revue), pourraient êtres extraits à partir des informations visuelles. Ils sont également intéressants à rapprocher des travaux montrant que les interactions audio-visuelles participent effectivement aux mécanismes d'accès au lexique (Fort *et al.*, 2010).

Enfin, le caractère « speech specific » de l'effet d'onset que nous avons mis en évidence est particulièrement important par rapport à la présentation théorique que nous avons faite dans la partie I concernant les mécanismes de liage perceptif. Le fait que l'onset visuel se rapporte selon nous à une information interprétable articulatoirement renvoie évidemment à la question du rôle du couplage perceptuo-moteur dans le liage. Nous allons aborder cette question par l'outil de la neurophysiologie dans le chapitre suivant. Nous pourrons alors, dans le chapitre 6, essayer de dresser un portrait cohérent de l'effet de transformation verbale, et, à travers lui, du mécanisme de liage audio-visuel des percepts de parole. À la lumière des résultats obtenus dans nos expériences comportementales, des études décrites auparavant sur l'effet de transformation verbale et celles sur l'intégration audio-visuelle de la parole, notamment par Sato (2004); Sato *et al.* (2006, 2007); Kondo et Kashino (2007); Skipper *et al.* (2005, 2007), et en prenant en compte les résultats de l'étude neuro-anatomique que nous présenterons dans le chapitre suivant, un cadre perceptuo-moteur et multimodal sera proposé dans le chapitre 6 afin de tenter de contribuer à expliquer l'effet de transformation verbale et à en déduire des propositions sur l'organisation perceptive de la parole.

Chapitre 5

Le circuit des transformations verbales dans le cerveau

Sommaire

Nos études comportementales ont conclu à l'existence de mécanismes de liage audio-visuel utilisant des événements visuels particuliers, interprétables articulatoirement, les « onsets », associés à des gestes d'ouverture de la mâchoire. C'est à travers ces mécanismes, notamment, que s'organiseraient les interactions audio-visuelles dans l'effet de transformation verbale. Il est donc légitime de chercher à associer à ces interactions audio-visuo-motrices le circuit cortical candidat naturel à organiser ce lien dans le cerveau : la voie dorsale. Nous avons déjà mentionné dans la partie I les études de Sato *et al.* (2004) et Kondo et Kashino (2007) montrant l'activation des aires pariétales (gyrus supramarginal) et frontales (cortex prémoteur, gyrus frontal inférieur, cortex cingulaire antérieur, cortex préfrontal) dans des tâches impliquant l'effet de transformation verbale. Cependant, ces données, obtenues par la technique d'IRMf, ne fournissent pas d'information temporelle et notamment ne peuvent permettre de décider si ce circuit dorsal pariéto-frontal correspond à une activation globalement forte associée par exemple à un renforcement des représentations perceptivo-motrices associé à la modification constante du percept, ou s'il est le lieu spécifique de processus de prise de décision conduisant à ces bascules.

Or, nous avons eu l'occasion pendant ce travail de thèse d'accéder à une technique difficile d'accès mais particulièrement adaptée à notre sujet : l'EEG intracrânienne. En effet, dans le Département de neurologie et psychiatrie du CHU de Grenoble, l'équipe du Professeur Philippe Kahane s'occupe régulièrement de patients atteints d'épilepsies lourdes impliquant une chirurgie corticale, précédée alors de périodes d'hospitalisation longue durée (typiquement deux semaines) pendant lesquelles les patients sont implantés d'électrodes localisées dans des secteurs correspondant à leur étiologie propre. Pendant ces périodes, les patients sont disponibles pour des

tâches expérimentales en dehors des périodes de crise, tâches pendant lesquelles on a ainsi accès à une activité EEG (dire « intracrânienne ») qui assure une double précision, à la fois spatiale (pourvu que l'on ait accès à des électrodes situées dans les régions adéquates) et temporelle inégalable actuellement. Nous avons pu ainsi monter, avec Philippe Kahane et Jean-Philippe Lachaux, un protocole expérimental sur ces patients, visant à déterminer si l'activation pariéto-frontale est confirmée avec cette technique, et surtout si elle correspond précisément aux régions temporelles de bascule perceptive. C'est cette expérience que nous présentons ici.

Nous présentons un résumé de cette étude dans la première section de ce chapitre. Puis, l'étude est présentée en détail dans la deuxième section sous sa forme publiée dans le journal NeuroImage.

5.1 Résumé

5.1.1 Introduction

L'objectif de cette expérience était l'étude des activités cérébrales précédant les bascules perceptives en parole. Pour cela, nous avons enregistré les signaux intracrâniens (iEEG) chez deux patients épileptiques pendant la réalisation d'une tâche de transformation verbale grâce aux électrodes implantées dans différentes régions de leur cortex. Les emplacements des électrodes chez les patients étaient uniquement liés à des raisons thérapeutiques. Nous avons choisi des patients qui avaient simultanément des électrodes dans le gyrus temporal supérieur gauche, le gyrus supramarginal gauche et le gyrus frontal inférieur gauche, trois régions reliées dans le circuit dorsal de la perception de la parole (Hickok et Poeppel, 2007) et mentionnées dans les études IRMf sur l'effet de transformation verbale (Sato *et al.*, 2004; Kondo et Kashino, 2007). Nous avons effectué des analyses temps-fréquence sur les signaux iEEG recueillis dans l'intervalle 1-160 Hz. À la lumière des études montrant le rôle des synchronisations neuronales en bande gamma en lien avec le liage perceptif (voir sous-section 1.1.2) et avec la communication locale et globale entre les neurones (Fries *et al.*, 2007), cette étude avait pour but l'observation des activités dans cette bande fréquentielle précédant les transformations verbales au sein du réseau observé précédemment dans les deux études IRMf. L'objectif était, nous l'avons dit, de déterminer si les bascules sont précédées d'une augmentation significative d'activité dans cette bande et dans ces régions dorsales, signe alors qu'elles seraient impliquées spécifiquement dans le processus de prise de décision et de bascule perceptive.

5.1.2 Méthode expérimentale

Sujet

Deux patients épileptiques droitiers de langue maternelle française ont participé à cette étude (Pt1 : femme, 32 ans et Pt2 : femme, 27 ans). Les sujets ne présentaient aucun trouble auditif, ni articulatoire.

Condition expérimentale

Deux conditions ont été utilisées dans cette expérience : la condition ENDO (pour « modifications endogènes », c'est-à-dire induites par le sujet sur un stimulus non variable) qui était la condition de transformation verbale et la condition EXO (pour « modifications exogènes », c'est-à-dire induites par le stimulus lui-même), la tâche étant toujours détecter des changements auditifs. Pendant la condition ENDO, les sujets écoutaient des séquences /pata/ et /tapa/ en boucle et ils signalaient leurs changements perceptifs spontanés et endogènes en appuyant sur un bouton. Dans la condition EXO, les changements perceptifs étaient dus aux changements exogènes dans les stimuli. Dans cette condition, nous avons demandé aux sujets d'écouter des séquences de type /...papapa...tatata.../. Ils appuyaient sur un bouton s'ils percevaient des changements auditifs de /pa/ à /ta/ ou vice versa. La condition EXO nous a permis de séparer les activités cérébrales dues à l'appui sur le bouton (activités communes dans les deux conditions) et celles dues aux transformations verbales dans la condition ENDO. De plus, les temps de réaction des sujets ont été estimés à l'aide de cette condition. Les deux sujets ont effectué la condition EXO avant la condition ENDO.

Matériel phonétique et stimuli

Un ensemble de séquences /pa/ et /ta/ a été enregistré puis numérisé sur 16 bits à une fréquence d'échantillonnage de 44.1 kHz. Elles ont été contrôlées pour être égalisées au niveau des caractéristiques du signal acoustique (formants de la voyelle, décours temporels de l'intensité et de F0, durées des composantes segmentales). Les séquences /pata/ et /tapa/ ont été construites à l'aide des séquences /pa/ et /ta/ avec un intervalle inter-stimuli [ISI] de 100 ms. La durée des séquences /pa/ et /ta/ était ainsi de 250 ms et celle des séquences /pata/ et /tapa/ était de 500 ms. Pour la condition ENDO, les séquences /pata/ et /tapa/ étaient répétées chacune 300 fois (150 secondes). Pour la condition EXO, nous avons répété 600 fois les séquences /pa/ et /ta/ d'une durée aléatoire entre 4 et 8 secondes pour obtenir des stimuli de la forme /...papapa...tatata...papapa...tatata.../.

Enregistrement

Les enregistrements étaient effectués au moyen, respectivement, de 12 et 14 électrodes semi-rigides pour les patients Pt1 et Pt2. Chaque électrode avait 10 ou 15 contacts distants de 1.5 mm l'un de l'autre (Dixi, Besançon, France. Pour les explications sur cette méthode, voir Kahane *et al.*, 2004). Ainsi, nous avons enregistré 126 sites de l'hémisphère gauche de chaque patient. Les essais montrant des activités épileptiques ont été supprimés de nos analyses.

Analyse des données

Les analyses temps-fréquence ont été effectuées à l'aide du logiciel ELAN-Pack développé au laboratoire INSERM U281, Lyon, France. Pour les deux conditions

ENDO et EXO, les périodes d'une seconde avant l'appui sur le bouton ont été comparées avec les périodes neutres pendant lesquelles les sujets ne signalaient aucun changement perceptif. Ces régions neutres ont été choisies au moins une seconde avant et une seconde après les instants d'appui sur le bouton. Nous avons utilisé la fenêtre temps-fréquence [200 ms $\times$ 8 Hz] pour paver la région -1000 ms à 0 ms avant l'appui sur le bouton en temps et de 1 Hz à 160 Hz en fréquence. Un test Mann-Whitney a été effectué dans cette région pour chaque fenêtre par rapport à la même fréquence dans les régions neutres. Les tests multiples ont été corrigés en utilisant la correction de Bonferroni.

5.1.3 Résultats

Les temps de réaction des sujets ont été calculés dans la condition EXO en estimant la durée entre le changement auditif dans les stimuli et la détection des sujets (l'appui sur le bouton). La moyenne de ces durées était de 334 ($\pm$87) ms pour le patient Pt1 et de 601 ($\pm$371) ms pour le patient Pt2. Dans la condition ENDO, le nombre total des transformations rapportées était de 70 pour Pt1 et de 58 pour Pt2. La moyenne de stabilité de chaque percept était de 3.74 ($\pm$3.54) secondes et de 4.86 ($\pm$2.48) secondes respectivement pour les patients Pt1 et Pt2.

Le tableau 5.1 présente les électrodes montrant les activités les plus importantes dans la bande gamma dans les conditions ENDO et EXO (par rapport aux périodes neutres). Dans la condition ENDO, ces activités étaient 800-300 ms avant l'appui sur le bouton alors que les activités dans la condition EXO étaient principalement dans l'intervalle 200-0 ms par rapport à cette référence.

5.1.4 Discussion

Ayant un nombre limité de sujets, nous concentrons notre discussion sur les régions actives chez les deux patients. Dans la condition de détection de changement auditif « réel » (EXO), nous avons trouvé des activités gamma dans la partie postérieure du gyrus temporal supérieur gauche et dans une moindre mesure, dans le gyrus supramarginal gauche pour les deux patients. Le gyrus temporal supérieur gauche a été observé actif dans différentes tâches de perception de la parole (voir section 2.2) et aussi en lien avec la détection de changement auditif (*oddball*) (Näätänen, 2001), ce qui est cohérent avec la tâche demandée dans cette condition et avec notre résultat. Les activités dans le gyrus supramarginal ont été proposées pour la mémoire phonologique (Honey *et al.*, 2000), le jugement phonologique (Romero *et al.*, 2006) et aussi la préparation motrice (Deiber *et al.*, 1996). Cette étude ne permet pas de préciser parmi ces propositions quel rôle jouent les zones pariétales dans cette condition.

Dans la condition de transformation verbale (ENDO), les activités gamma cohérentes chez les deux patients ont été trouvées dans deux régions, le gyrus frontal inférieur gauche et le gyrus supramarginal gauche, 800 à 300 ms avant l'appui sur le bouton pour les deux patients. Ce résultat est cohérent avec le circuit dorsal de traitement de la parole (Hickok et Poeppel, 2007) reliant les représentations

TABLE 5.1 : Expérience iEEG : les électrodes montrant les activités les plus importantes en bande gamma en condition ENDO et EXO.

Condition ENDO

	Site	Localisation
Pt1	s'9	Lobule pariétal inférieur (BA40)
Pt1	r'8	Gyrus frontal inférieur (BA44/BA6)
Pt1	q'6	Gyrus frontal inférieur (BA44)
Pt1	x'8	Gyrus antérieur brevis de l'insula
Pt2	s'7	Lobule pariétal inférieur (BA40)
Pt2	g'15	Gyrus frontal inférieur (BA45/46)
Pt2	q'5	Gyrus frontal inférieur (BA45)
Pt2	e'5	Gyrus fusiforme (BA36)

Condition EXO

	Site	Localisation
Pt1	s'9	Lobule pariétal inférieur (BA40)
Pt1	r'8	Gyrus frontal inférieur (BA44/BA6)
Pt1	x'8	Gyrus antérieur brevis de l'insula
Pt1	t'7	Partie antérieure du gyrus temporal supérieur (BA41)
Pt1	u'4	Gyrus temporal supérieur (BA41) et Gyrus de Heschl
Pt1	u'10	Partie postérieure du gyrus temporal supérieur (BA42)
Pt2	u'9	Partie postérieure du gyrus temporal supérieur (BA42)
Pt2	s'7	Lobule pariétal inférieur (BA40)

auditivo-phonétiques et les représentations articulatoires. En accord avec ce circuit et conformément aux activités précédemment observées en IRMf (Sato *et al.*, 2004; Kondo et Kashino, 2007), ce couplage pariéto-frontal suggère la présence d'un lien fort entre les mécanismes de la perception et la production de la parole dans l'effet de transformation verbale. La précision temporelle de l'iEEG a notamment permis de mettre en évidence le rôle de ce couplage dans la phase de la prise de décision et en lien avec l'émergence d'un nouveau percept.

Ainsi, cette étude apparaît cohérente avec l'analyse que nous avons faite de nos données expérimentales, impliquant des processus de liage audio-visuel interprétables articulatoirement. Elle précise la connaissance du rôle du circuit dorsal mis en évidence précédemment dans notre équipe par Sato *et al.* (2004), dans deux directions : d'une part en spécifiant le rôle de ce circuit dans le processus de décision, puisque c'est au moment des bascules perceptives que l'activité est renforcée ; d'autre part, et ce point est très important, en remplaçant la tâche perceptuo-motrice utilisée par Sato *et al.* (2004) (étude dans laquelle les sujets devaient produire mentalement des transformations en parole intérieure) par une tâche perceptive pure, puisque nos sujets étaient ici en simple condition d'écoute. Ce point, cohérent avec l'étude de Kondo et Kashino (2007), confirme que c'est bien un processus de cou-

plage perceptuo-moteur, induit par des stimulations auditives n'impliquant par de tâche motrice explicite, qui est en jeu dans les transformations verbales.

5.2 *Parieto-frontal gamma band activity during the perceptual emergence of speech forms*

Dans la suite, notre étude est présentée en détail sous sa forme publiée dans le journal NeuroImage.

ELSEVIER

NeuroImage

www.elsevier.com/locate/ynimg
NeuroImage 42 (2008) 404–413

Parieto-frontal gamma band activity during the perceptual emergence of speech forms

Anahita Basirat,[a] Marc Sato,[a,b] Jean-Luc Schwartz,[a,*] Philippe Kahane,[c,d] and Jean-Philippe Lachaux[e]

[a]*GIPSA-Lab, ICP, CNRS UMR 5216, Institut National Polytechnique de Grenoble, Université Joseph Fourier & Université Stendhal, Grenoble, France*
[b]*Centre for Research on Language, Mind and Brain and School of Communication Sciences and Disorders, McGill University, Montreal, Canada*
[c]*Département de Neurologie et Psychiatrie and INSERM U836-UJF-CEA, Hôpital Michallon, Grenoble, France*
[d]*CTRS-IDEE, Hospices Civils de Lyon, France*
[e]*Unité Dynamique Cérébrale et Cognition, INSERM U821, Lyon, France*

Received 31 October 2007; revised 26 March 2008; accepted 30 March 2008
Available online 16 April 2008

The multistable perception of speech refers to the perceptual changes experienced while listening to a speech form cycled in rapid and continuous repetition, the so-called Verbal Transformation Effect. Because distinct interpretations of the same repeated stimulus alternate spontaneously, this effect provides an invaluable tool to examine how speech percepts are formed in the listener's mind. In order to track the temporal dynamics of brain activity specifically linked to perceptual changes, intracerebral EEG activity was recorded from two implanted epileptic patients while performing a verbal transformation task. To this aim, they were asked to carefully listen to a speech sequence played repeatedly and to press a button whenever they perceived a change in the repeated utterance. For both patients, 300–800 ms prior to the reported perceptual transitions, high frequency activity in the gamma band range (>40 Hz) was observed within the left inferior frontal and supramarginal gyri. An additional auditory decision task was used to rule out the possibility that the increased gamma band activity was due to the patients' motor responses. These results suggest that articulatory-based representations play a key part in the endogenously driven emergence of auditory speech percepts. The findings are interpreted in relation to theories assuming a link between perception and action in the human speech processing system.

Keywords: Multistable perception; Verbal Transformation Effect; Perceptual awareness; Intracerebral EEG; Gamma band activity; Speech perception; Speech production

Introduction

Multistable perception phenomena refer to the perceptual alternation between two or more mutually exclusive conscious interpretations of an unchanging sensory stimulation. Because different interpretations of the same stimulus alternate spontaneously, multistable perception provides a rare opportunity to examine the neural processes underlying endogenously driven conscious perception. In the visual domain, multistable perception has been described for a wide range of stimuli, such as ambiguous figures or binocular rivalry (e.g., Leopold and Logothetis, 1999, for a review). While adaptation or inhibition mechanisms at the sensory level have been suggested as the loci of multistable visual perception, perceptual alternation might also depend on an interplay between bottom–up, sensory and internal, top–down, neural processes (see Leopold and Logothetis, 1999; Kast, 2001; Blake and Logothetis, 2002, for a review). From this view, coordinated neuronal activations among widely distributed visual, parietal and frontal brain regions might be critical for perceptual awareness (Dehaene and Naccache, 2001; Crick and Koch, 2003).

Although multistable perception has been studied mainly in the visual modality, alternating perceptual interpretations of an unchanging auditory speech stimulus has also been reported, the so-called Verbal Transformation Effect (Warren and Gregory, 1958; Warren 1961). This effect refers to the perceptual changes experienced while listening to a speech stimulus cycled in rapid and continuous repetition. Initially, a percept matching the original form is heard, but at some point another percept suddenly arises. For example, the rapid repetition of the word "life" produces a perceptual transform into "fly", "fly" back into "life" and so on. This transformation process persists throughout the repetition procedure, leading to perceptual switches from one speech form to another (or back to the original form). Previous studies have reported that perceptual changes relative

* Corresponding author. Laboratoire GIPSA, Département Parole et Cognition, CNRS UMR 5216, Institut National Polytechnique de Grenoble 961 rue de la Houille Blanche – Domaine universitaire – BP 46, 38402 Saint Martin d'Hères, Cedex, France. Fax: +33 4 76 57 47 10.
E-mail address: jean-luc.schwartz@gipsa-lab.inpg.fr (J.-L. Schwartz).
Available online on ScienceDirect (www.sciencedirect.com).

1053-8119/$ - see front matter © 2008 Elsevier Inc. All right reserved.
doi:10.1016/j.neuroimage.2008.03.063

to the auditory input mainly range from auditory streaming/perceptual grouping (Pitt and Shoaf, 2001, 2002), to phonological (Warren, 1961; Warren and Meyers, 1987) and lexical transformations (Warren, 1961; Shoaf and Pitt, 2002). While verbal transformations have initially been studied as a pure auditory perceptual effect, they appear to occur also when subjects repeatedly utter the speech stimulus in both an overt and a covert mode (Reisberg et al., 1989; Smith et al., 1995; Sato et al., 2006). With this production procedure, the number of verbal transformations has been shown to gradually decrease from a condition of complete externalization to one of complete internalization, when subarticulation is blocked by a concurrent articulatory task, through a condition of partial externalization (i.e., whispering, mouthing; Reisberg et al., 1989). Sato and colleagues (2006) further showed that verbal transformations are specifically influenced by articulatory synergies. They analysed the stability of sequences i.e. the number of times a given sequence was not transformed and the attractivity, i.e. the number of times a sequence was selected as a transformation. They observed that both perceptual stability and attractivity of a sequence depend on articulatory constraints. In addition, the fact that verbal transformations may depend on articulatory constraints appears also consistent with the finding that a concurrent silent articulation decreases the number of reported perceptual changes while listening to a repeating word (MacKay et al., 1993).

The fact that, besides auditory, phonological and lexical influences, articulatory constraints may act on the emergence and stabilization of verbal transformations suggests that they partly rely on motor neural processes. This hypothesis is in keeping with two recent functional magnetic resonance imaging (fMRI) studies examining neural activities during a verbal transformation task (Sato et al., 2004; Kondo and Kashino, 2007). Using a block-design paradigm, Sato and colleagues (2004) contrasted a verbal transformation condition involving the mental repetition of speech sequences with an active search for verbal transformation, with a baseline condition involving the simple mental repetition of the same items. When compared to the baseline, the verbal transformation condition showed a predominantly left-lateralized network of brain activations within the inferior frontal gyrus, extending into the anterior part of the insular cortex, the supramarginal gyrus and the superior temporal gyrus. The authors suggest that this temporo–parieto–frontal neural network likely reflects the online analysis of the rehearsed speech sequence and the temporary storage of the recently built representation. Additional activations were also observed within the right anterior cingulate cortex and the cerebellum bilaterally and were assumed to reflect attentional control and/or comparison of speech forms during the active search for verbal transformations. In an event-related fMRI study, Kondo and Kashino (2007) attempted to identify brain regions activated at the time of perceptual transitions during a purely auditory verbal transformation condition and during a tone detection condition. Both conditions involved bilateral activations within the primary auditory area, the posterior part of the superior temporal gyrus, the supramarginal gyrus and within the left insular cortex. However, the anterior cingulate cortex, the prefrontal cortex and the left inferior frontal gyrus were found to be activated only in the verbal transformation condition.

Despite methodological differences, similar frontal, parietal and temporal areas were found to be activated in these two fMRI studies, including the anterior cingulate cortex, the inferior frontal gyrus extending into the insular cortex, the supramarginal gyrus and the superior temporal gyrus. Besides the activation of the anterior cingulate cortex, which is likely to reflect some competition mechanisms between different possible representations (e.g., Carter et al., 1998), these brain areas strongly resemble those observed during previous brain imaging studies of speech perception in which a left-lateralized network, including the posterior superior temporal gyrus, the inferior parietal lobule and the inferior frontal gyrus, has been consistently identified (see Hickok and Poeppel, 2007, for a review). It has been proposed that this temporo–parieto–frontal "dorsal stream" provides a mechanism for the development and maintenance of parity between sound-based representations in the superior temporal gyrus and articulatory-based representations in the inferior frontal gyrus, via sensorimotor recoding in the inferior parietal lobule (Hickok and Poeppel, 2000, 2004, 2007; Scott and Johnsrude, 2003). Taken together, these results suggest that perceptuo–motor interactions play a key part in the conscious emergence and stabilization of speech percepts.

Despite the rather coherent portrait of brain activations observed during verbal transformation tasks, little is known about the temporal dynamics of brain activity linked to verbal transformations, that is activations occurring prior to the consciously reported perceptual transitions. Brain imaging as well as non-invasive electroencephalographic (EEG) techniques do not provide the combined spatial and temporal resolution necessary to track activity in localized cortical regions in a period of time restricted to basically a few hundred of milliseconds. However, intracerebral EEG (iEEG) recordings, by means of electrodes stereotactically implanted inside the brain of some epileptic patients as part of their presurgical evaluation, enable to track cortical activity with high selectivity both in time and in space. The high spatial resolution in this study (approximately of 5 mm) excludes the problem of source localisation that exists in scalp EEG recordings (Lachaux et al., 2002; Lachaux et al., 2003, 2007). In the present study, iEEG activity was recorded over a wide frequency range (1–160 Hz) from two implanted epileptic patients, while performing a verbal transformation task. Although the implanted sites of the two patients were selected entirely for clinical purposes with no reference to the present study, their implantation sampled a large number of temporal, parietal and frontal regions previously found to be activated during a verbal transformation task (Sato et al., 2004; Kondo and Kashino, 2007). iEEG recordings detect local neural activity related to cognitive processes as transient spectral energy variations in several characteristic frequency bands, including the theta (4–7 Hz), alpha (8–12 Hz), beta (15–30 Hz) and especially gamma (above 30 Hz and up to 200 Hz) bands.

Recent iEEG studies have put a strong emphasis on gamma band activations in particular, in association with a wide range of cognitive processes, including memory (Fell et al., 2001; Howard et al., 2003; Mainy et al., 2007), visual attention and perception (Brovelli et al., 2005; Lachaux et al., 2000, 2005; Tallon-Baudry et al., 2005; Tanji et al., 2005), audition (Bidet-Caulet et al., 2003; Crone et al., 2001a,b; Edwards et al., 2005), somatosensory and motor processes (Crone et al., 1998a,b; Crone et al., 2006; Lachaux et al., 2005; Aoki et al., 1999; Pfurtscheller et al., 2003; Szurhaj et al., 2005), and language (Crone et al., 2001a,b). Importantly, in all those studies, gamma band activity was observed only in very specific brain regions, dependant on the tasks, and in good agreement with the functional networks revealed by fMRI (Lachaux et al., 2007).

The exact function of gamma band activity is still a matter of debate; initial studies have mostly related this phenomenon with visual integration, for instance for the perception of gestalt-like

visual stimuli (Keil et al, 1999; Tallon-Baudry and Bertrand, 1999, for review), but this interpretation seems to be to restrictive: for instance, several studies have found modulation of gamma band activity during speech perception (Palva et al., 2002; Kaiser et al., 2005; Ford et al., 2005). In fact, recent reviews have suggested that gamma band activity may in fact subserve a general mechanism facilitating and channelling local and global communication among neurons (Fries, 2005; Varela et al., 2001; Fries et al., 2007), which would explain why gamma band activity would play a role in processes as different as visual integration and speech perception.

That is, gamma band responses occur precisely in cortical regions associated with major cognitive functions, presumably reflecting local neural communication (Fries et al., 2007); as such, they constitute a candidate of choice to detect and describe the precise timing of short episodes of neural activity associated with verbal transformations. Based on this interpretation, our basic assumption in this study was that neural activity specifically associated with the emergence of a new percept in the verbal transformation task should be observed in the brain regions observed in fMRI studies (Sato et al., 2004; Kondo and Kashino, 2007), the left inferior frontal gyrus and the left supramarginal gyrus, prior to the conscious identification of a change in perception by subjects, and that activity should translate into energy increase in the gamma band in local iEEG recordings of those regions. We found that, for both patients, these two areas showed enhanced gamma band activity 300–800 ms prior to the reported perceptual transitions.

Materials and methods

Participants

Two right-handed patients, suffering from drug-resistant partial epilepsy and candidates for surgery, participated in the study (Pt1: female, 32 year old; Pt2: female, 27 year old; handedness was assessed by means of the Edinburgh Inventory, Oldfield, 1971). Both patients were native speakers of French and reported no hearing or speech disorders. Informed consent was obtained from each participant before the experiment.

Electrodes implantation

Magnetic resonance imaging (MRI) of the brain showed a left hippocampal sclerosis in patient Pt1 and Pt2. Because the location of the epileptic focus could not be identified using non-invasive methods, the two patients underwent iEEG recordings by means of stereotactically implanted multilead depth electrodes (for explanation of this methodology, see Kahane et al., 2004). 12 and 14 semi-rigid electrodes were implanted in patients Pt1 and Pt2, respectively, in various cortical areas depending on the suspected origin of seizures. Each electrode had a diameter of 0.8 mm and comprised 10 or 15 leads of 2 mm length, 1.5 mm apart (Dixi, Besançon, France), depending on the target region. The electrode contacts were identified on each individual stereotactic scheme, and then anatomically localized using the proportional atlas of Talairach and Tournoux (Talairach and Tournoux, 1988). In addition, the computer-assisted matching of post-implantation CT-scan with a pre-implantation 3-D MRI provided a direct visualization of the electrode contacts with respect to the brain anatomy of each patient (IVS Solution, Germany).

Although the implanted sites were chosen entirely for clinical purposes with no reference to the present experimental protocol, the two patients were selected to enter this study because their implantation sampled several cerebral regions previously found to be activated during a verbal transformation task (Sato et al., 2004; Kondo and Kashino, 2007). During the experiment, 126 sites were recorded from each of the two patients in the left hemisphere (see Fig. 1 and Table 1). Both patients performed the experiment four days after the implantation of the electrodes.

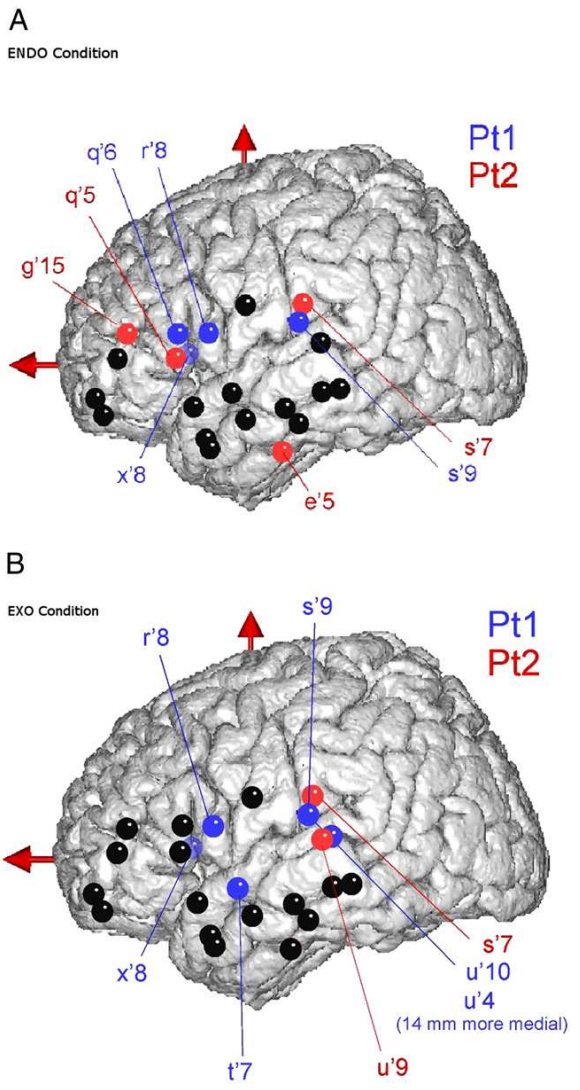


Fig. 1. Entry points of the intracranial electrodes across the two patients, projected onto the lateral view of a 3D reconstruction of the Montreal Neurological Institute (MNI) single-subject MRI (after conversion from Talairach and Tournoux to MNI stereotactic space). Blue and red dots indicate the recorded sites showing enhanced gamma band energy (A) in the verbal transformation task (ENDO condition) and (B) in the auditory decision task (EXO condition) for patients Pt1 and Pt2, respectively. Black dots indicate entry points for depth electrodes with no effect. Talairach coordinates for all the implanted sites are displayed in Table 1.

Table 1
Anatomical locations for each recorded site exhibiting enhanced energy in the gamma band (A) in the verbal transformation condition (ENDO condition) and (B) in the auditory decision task (EXO condition)

Patient ID	Site ID	Anatomical localization	Talairach *x*, *y*, *z*
A)			
Pt1	s′9	Inferior parietal lobule (BA40)	−59, −22, 18
Pt1	r′8	Inferior frontal gyrus (BA44/BA6)	−55, 6, 12
Pt1	q′6	Inferior frontal gyrus (BA44)	−44, 16, 11
Pt1	x′8	Brevis anterior gyrus of the insula	−31, 16, 5
Pt2	s′7	Inferior parietal lobule (BA40)	−61, −24, 21
Pt2	g′15	Inferior frontal gyrus (BA45/46)	−48, 38, 13
Pt2	q′5	Inferior frontal gyrus (BA45)	−39, 19, 4
Pt2	e′5	Fusiform gyrus (BA36)	−37, −18, −26
B)			
Pt1	s′9	Inferior parietal lobule (BA40)	−59, −22, 18
Pt1	r′8	Inferior frontal gyrus (BA44/BA6)	−55, 6, 12
Pt1	x′8	Brevis anterior gyrus of the insula	−31, 16, 5
Pt1	t′7	Anterior part of the superior temporal gyrus (BA41)	−59, −3, −6
Pt1	u′4	Superior temporal gyrus (BA41) and Heschl Gyrus	−36, −25, 10
Pt1	u′10	Posterior part of the superior temporal gyrus (BA42)	−59, −25, 10
Pt2	u′9	Posterior part of the superior temporal gyrus (BA42)	−31, 16, 5
Pt2	s′7	Inferior parietal lobule (BA40)	−61, −24, 21

Anatomical regions were drawn from the individual MRI of the patients, not from standard atlases.

Experimental conditions

Two experimental conditions were considered. In the verbal transformation condition (ENDO condition), patients were asked to listen carefully to a speech sequence (i.e., either /pata/ or /tapa/ syllables being played repeatedly), and to press a button with their left hand whenever they perceived a change in the repeated utterance, even if the utterance changed into one they had heard previously. A verbal transformation experiment using the same material showed that the main organization of the reported transformations for both speech sequences was that of a pairwise coupling between /pata/ and /tapa/ syllables, although other phonological or lexical transformations were also reported (Sato et al., 2007b). In a second condition (EXO condition), patients were asked to listen carefully to randomly alternating repetitions of two syllables (i.e., /...papapa...tatata...papapa.../), and to detect any transition between them by pressing a button with their left hand. The EXO condition was designed to disentangle neural activity specifically linked to endogenously driven perceptual changes in the ENDO condition from those related to motor activities due to the patients' responses. Most specifically, the expectation was that the same motor activity due to button pressing should occur in both conditions and that an increase in neural activity due to the perceptual changes in the EXO condition cannot be expected apart from the temporal interval between the syllable transition and the button-pressing event. This reaction time being estimated in the EXO condition, a modulation of spectral energy appearing in the ENDO condition at latencies prior to those observed in the EXO condition would likely be related to neural activity specifically associated with the emergence of a new verbal percept. For both patients, the EXO condition was performed before the ENDO condition.

Stimuli

Multiple utterances of /pa/ and /ta/ syllables were recorded in a soundproof room by a trained phonetician, native French speaker (J.-L.S.). The speaker pronounced each syllable naturally at a conversational rate, maintaining an even intonation and vocal intensity while producing the sequences. The items were digitized (16 bit resolution) and sampled at 44.1-kHz sampling rate directly to disk on a PC computer. One clearly articulated token was selected for each sequence, /pa/ and /ta/ syllables being matched as closely as possible for acoustic durations (as checked by a spectrogram analysis using the Praat software, Institute of Phonetic Sciences, University of Amsterdam, the Netherlands).

The speech sequences used in the ENDO condition (i.e., /...patapatapata…/) and in the EXO condition (i.e., /...papapa...tatata...papapa.../) were constructed by concatenating the /pa/ and /ta/ syllables, with a 100 ms silent period inserted between the offset of the vowel and the onset of the following consonantal burst. With this procedure, the duration of each consonant–vowel syllable was 250 ms. For each condition, two distinct speech sequences were built, differing in the ordering of the repeated syllables (i.e., /patapata.../ and /tapatapa.../ for the ENDO condition, /papa...tata.../ and /tata...papa.../ for the EXO condition), in order to minimize any possible priming effect. For the EXO condition, the number of repetitive /pa/ or /ta/ syllables was randomly varied from 4 s to 8 s (corresponding to 23 switches between the two syllables). The random distribution was used to reduce the predictability of the switches, and maintain subject's attention. Each of the four speech sequences lasted 150 s. (corresponding to the concatenation of 600 syllables). These samples are available here: http://www.icp.inpg.fr/~basirat/stimuli.html.

Recordings

Intracerebral recordings were conducted using an audio-video-EEG monitoring system (Micromed, Treviso, Italy), which allowed the simultaneous recording of 128 depth-EEG channels sampled at 512 Hz [0.1–200 Hz bandwidth] during the experimental procedure. One of the contact sites in the white matter was chosen as reference. This reference had the same impedance as the other contact sites, and was located in a region with no or little electrical field source. In addition, it was not contaminated by eye-movements artefacts or electromyographic activity from subtle muscle contractions. Furthermore, all signals were re-referenced to their nearest neighbour 3.5 mm away on the same electrode before analysis (bipolar montage). Recording sites showing clear epileptiform activities were excluded from the analysis, and among the remaining sites, monopolar and bipolar data, both raw and high-pass filtered (above 15 Hz), were systematically inspected. Any trial showing epileptic spikes in any of those traces was discarded. Note that the high-pass filtering process was done solely for artefact detection, all analysis presented in this study were performed on raw, unfiltered, signals.

Time-frequency data analysis

EEG signals were evaluated with the software package for electrophysiological analysis (ELAN-Pack) developed in the INSERM U821 laboratory (Lyon, France). For each single trial, bipolar derivations computed between adjacent electrode contacts

were analyzed in the time-frequency (TF) domain by convolution with complex Gaussian Morlet's wavelets (Tallon-Baudry et al., 1997), thus providing a TF power map: $P(t,f)=|w(t,f)*s(t)|^2$ where $w(t,f)$ was for each time t and frequency f a complex Morlet's wavelet: $w(t,f)=A\exp(-t^2/2\sigma_t^2)\times\exp(2i\pi ft)$ with $A=(\sigma_t\sqrt{\pi})^{-1/2}$ and $\sigma_t=1/(2\pi\sigma_f)$, and σ_f a function of the frequency f: $\sigma_f=f/7$.

Task-related spectral energy modulations of iEEG signals were detected using Mann–Whitney non-parametric tests. The objective was to test, for several frequency bands up to 160 Hz, whether the spectral energy before the patients' button press was higher than the energy measured at the same frequencies in neutral episodes during which patients reported no endogenous or exogenous perceptual switches. More precisely, we defined [200 ms × 8 Hz] TF windows covering a TF region from −1000 ms to 0 ms before button press in the time domain, and from 1 to 160 Hz in the frequency domain. For each such TF window, the total wavelet time-frequency energy was measured in that window for each button press to provide N energy values (N being the number of responses from the patient). Those N energy values were compared statistically with N energy values measured in N [200 ms × 8 Hz] TF windows in the same frequency band and at "neutral" latencies, with no perceptual switch and chosen at least one second before and one second after a button press. One Mann–Whitney test was performed for each [200 ms × 8 Hz] TF window. Tests were performed independently for the ENDO and EXO conditions. To correct for multiple testing, we applied a Bonferroni correction taking into account the number of non-overlapping [200 ms × 8 Hz] TF tiles covering the total [1000 ms × 160 Hz] TF domain which was tested (which yielded a corrected p equal or inferior to 5e−4 (0.05/(20 × 5)).

Results

Behavioural responses

The instances where the subjects pressed the button signalling a perceptual switch was the only available behavioural information. In the EXO condition, the mean reaction time corresponding to the delay between a stimulus switch (from /pa/ to /ta/ or from /ta/ to /pa/) and a button press was of 334 ms (±87) for patient Pt1 and of 601 ms (±371) for patient Pt2. The mean percentage of errors, corresponding to the absence of button press within 2 s after a stimulus switch, was 2% and 4% for patients Pt1 and Pt2, respectively. In the ENDO condition, the total number of transformations was of 70 for patient Pt1 and of 58 for patient Pt2. The number of transformations as a function of time is shown on Fig. 2A. In order to test whether the subjects' response pattern was different in the beginning and at the end of the experiment, we analysed the number of transformations in the first and last half of each trial. The t-test did not show any significant difference ($t(3)=-1.75$, $p=0.179$). Analyses of inter-switch durations, that is the time of stability from one transformation to the next, showed a mean stability duration of 3.74 s (±3.54) for patient Pt1 and of 4.86 s (±2.48) for patient Pt2 (the histograms of inter-switch durations are displayed on Fig. 2B).

iEEG responses

Analysis of iEEG responses were designed to detect possible neural activities specifically associated with perceptual transitions in both the ENDO and EXO conditions.

ENDO condition

In patient Pt1, four sites were found with a stronger spectral energy before the button press than in the reference period (see Figs. 1A, 3 and Table 1A). These activations were observed in the gamma band (above 40 Hz) and were maximal between 800 ms and 100 ms before the button press ([−600: −300] for q′6; [−500: −100] for x′8; [−800: −500] for r′8; [−600: −400] for s′9). For patient Pt2, four sites had stronger spectral energy before the button press than in the reference period (see Figs. 1A, 3 and Table 1A). As for patient Pt1, these activations were observed in the gamma band, between 600 ms and 200 ms before the button press ([−500 ms: −300 ms] for q′5; [−500 ms: −300 ms] for g′15; [−600 ms: −300 ms] for s′7; [−500 ms: −200 ms] for e′5). Therefore, gamma band activations common to both patients were found in the left inferior frontal and supramarginal gyri 300–800 ms prior to the reported perceptual transitions. Notice that these activities occur at a negative latency relative to the button press, which is within the reaction time in the EXO condition for patient Pt2, but generally quite larger for patient Pt1. This rules out, for

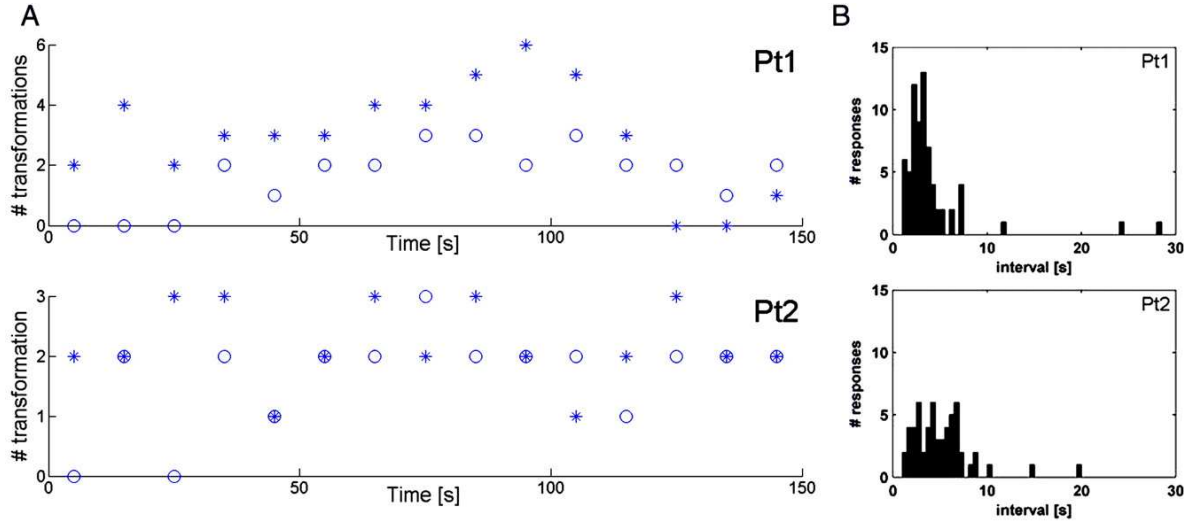


Fig. 2. Behavioural results in the ENDO Condition. A: Number of transformations (O and *, two trials per patient) as a function of time for patient Pt1 and Pt2 (time step = 10 s.). B: Histograms of inter-switch durations for the two patients Pt1 and Pt2.

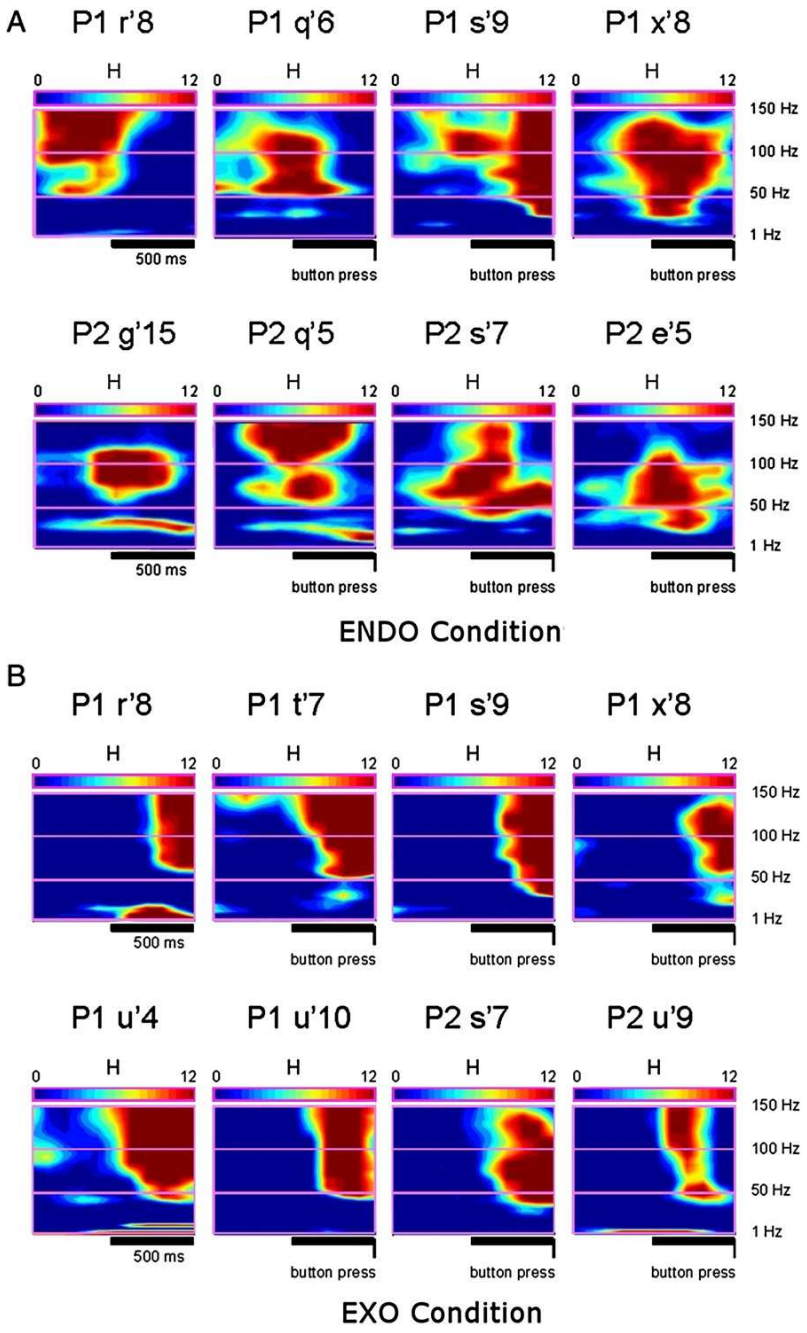


Fig. 3. Time-frequency representations for the sites showing enhanced gamma band energy in the ENDO and EXO conditions for patients Pt1 and Pt2. Maps show H values of the non-parametric comparisons with the neutral periods (see text for details); H values higher than 12 correspond to significant energy increases with $p<0.0005$. Talairach coordinates for the implanted sites are displayed in Table 1.

this patient at least, the possibility that the increase in activity is only due to the preparation of the button pressing action *per se*.

EXO condition

Some of the previous sites mentioned above also demonstrated significant gamma band increases in the EXO condition, but these were found much later always after −200 ms before the button press (Pt1: s′9, r′8, x′8; Pt2: s′7). In addition, the EXO condition was characterized by gamma band increases after −200 ms in the superior temporal gyrus (Pt1: t′7, u′4, u′10; Pt2: u′9 — see Figs. 1B, 3, and Table 1B).

Discussion

The present study was designed to examine the temporal dynamics of brain activity linked to perceptual changes during a verbal

transformation task. To this aim, iEEG activities were recorded from two implanted epileptic patients, while they listened to an auditory speech sequence played repeatedly and reported any perceptual change. For both patients, an increase in gamma band activity was observed 300–800 ms prior to the reported perceptual transitions within the left inferior frontal and supramarginal gyri, while no modulation occurred in temporal areas. In an additional auditory decision task, an increase of gamma band activities was also observed but in a shorter latency relative to the patients' motor responses and mainly in the left superior temporal and supramarginal gyri.

Before discussing these results, it is important to highlight some limitations of the present study. One limitation, inherent to patient studies, is the possible long term effects of epilepsy on cognitive abilities (e.g., Jokeit and Ebner, 2002; Hermann et al., 2006). Although cognitive deficits cannot be excluded, it is worthwhile noting that the two patients performed the tasks correctly and that neural activity found for both patients (who do not have the same epilepsy) were in brain areas previously reported for the verbal transformation effect using fMRI (Sato et al., 2004; Kondo and Kashino, 2007). In addition, the epileptogenic zone of the two patients – as precisely defined by intracerebral recordings – was restricted to antero-mesial temporal lobe structures. A second limitation comes from the fact that intracranial electrodes did not allow sampling of all the cerebral regions previously found to be activated during the verbal transformation effect (Sato et al., 2004; Kondo and Kashino, 2007). Notably, given the patients' implantation and the fact that verbal transformations mainly involve left-lateralized activations (Sato et al., 2004), neural activity was recorded only in the left hemisphere. Finally, a third limitation comes from the fact that no information on the precise nature of the transformations was available. However, a previous verbal transformation experiment using the same material showed that the main organization of the reported transformations for both speech sequences was that of a pairwise coupling between /pata/ and /tapa/ syllables, although various others transformations were reported (Sato et al., 2007b). From these results, the reported transformations in the present study are likely to be strongly linked to a syllabic parsing process, rather than to auditory streaming or lexical competition mechanisms, as also sometimes observed in verbal transformations (see Sato et al., 2006, 2007a,b).

Nevertheless, the present study provides the first direct recordings of neural activity preceding, or simultaneous with, a verbal transformation. A coherent portrait of neural activity was observed for both patients, in the inferior frontal and the supramarginal gyri in the ENDO condition, as well as in the superior temporal and supramarginal gyri in the EXO condition. Given the restricted number of subjects, we will focus on these brain areas found to be activated for both patients. Note also that the modulation of spectral energy observed before the patients' responses is relative to some neutral periods, selected in the same condition but in which no perceptual transitions were reported. Therefore, the absence of modulation in a particular brain region does not necessarily mean that this region is not activated during the task but, possibly, it is not differentially activated than in neutral periods.

Altogether, the observed increase of gamma band activity in the EXO condition appeared around 200 ms prior to the motor responses in the temporal and, to a lesser extent, parietal regions. Activations observed within the left temporal areas likely reflect the auditory identification of the new syllable. The left temporal lobe has long been implicated in the perception of speech sounds (e.g., Zatorre et al., 1992; Binder et al., 2000; Scott et al., 2000; Binder et al., 2004). While processing of the spectrotemporal features of both speech and non-speech sounds has been attributed to the primary auditory cortex and dorsolateral portions of the superior temporal gyrus (Binder et al., 2000), phonetic processing of speech signals involves mainly the left anterior superior temporal gyrus and the adjacent superior temporal sulcus extending both anteriorly and posteriorly (Binder et al., 2000; Scott et al., 2000). Furthermore, a direct relationship between identification accuracy and neural activations has been also observed in the anterolateral aspect of Heschl's gyrus and adjacent lateral superior temporal gyrus during a syllable discrimination task (Binder et al., 2004). Finally, the present neural modulation in temporal areas also appears in line with previous electrophysiological and neuromagnetic studies of speech perception, showing a mismatch negativity response in the auditory cortex elicited by infrequent deviant sounds presented among frequent standard sounds (see Näätänen, 2001, for a review). Indeed, each transition from /pa/ to /ta/ and /ta/ to /pa/ could be considered as an oddball, just as in the mismatch negativity paradigm. The specific signification of increased gamma band activities observed in the supramarginal gyrus is less clear. The supramarginal gyrus has been associated with temporary storage of phonological material (Paulesu et al., 1993; Cohen et al., 1997; Jonides et al., 1998; Honey et al., 2000), phonological judgements (Romero et al., 2006), motor preparation (Deiber et al., 1996), and motor attention (Rushworth et al., 2001). The present study cannot disambiguate the specific contribution of the supramarginal gyrus to these processes.

In the ENDO condition, an increase of gamma band activity was observed for both patients around 300–800 ms prior to the motor responses within the left inferior frontal and supramarginal gyri. Note that these activities occur at a negative latency relative to the motor responses, which for both patients is quite earlier than that observed in the corresponding areas in the EXO condition. This result therefore clearly rules out the possibility that the increase in activity could be due to the preparation of the button pressing action *per se*. The activities observed in the present study in frontal and parietal areas are consistent with our previous study on verbal transformations, which did not involve button pressing at all. This further suggests that button pressing does not modify the multistability phenomenon itself. It confirms the role of this fronto–parietal network in the verbal transformation process. Moreover, the precise temporal localisation allowed by iEEG enables the confirmation of the specific role of this network in perceptual switches and decision making. Furthermore, the fact that most transformations in this kind of sequences are towards /pata/ and /tapa/ sequences rather than related to auditory streaming processes (Sato et al., 2007b) strongly suggests that this network plays a role in speech perception.

Actually, in past studies, the inferior frontal gyrus has been repeatedly found to be activated during phonological processing, in phoneme monitoring, syllable counting and rhyming tasks (e.g., Démonet et al., 1992, 1994; Paulesu et al., 1993; see Poldrack et al., 1999; Démonet et al., 2005; Vigneau et al., 2006, for a review), as well as during auditory speech perception (e.g., Wilson et al., 2004; Watkins and Paus, 2004; Pulvermuller et al., 2006; Wilson and Iacoboni, 2006). This region thus appears well adapted to syllable parsing process in the present verbal transformation task (Sato et al., 2004), and to speech segmentation in general (Burton and Small, 2006). As previously noted, the supramarginal gyrus has been associated with both temporary

storage of phonological material (Paulesu et al., 1993; Cohen et al., 1997; Jonides et al., 1998; Honey et al., 2000) and phonological judgements (Romero et al., 2006). Given that the verbal transformation task appears to require at least minimal verbal storage (Smith et al., 1995; Sato et al., 2004), this region is thus likely to be involved in the temporary storage of the present percept until the emergence of a new one. Altogether, the present fronto–parietal coupling, involved in both syllable parsing processes and temporary storage of the latterly built representation, could therefore provide a well-adapted platform for phonological comparison and decision-making processes before the conscious emergence of a new speech form.

In summary, the observed increase of gamma band activity within the left inferior frontal and supramarginal gyri, 300–800 ms prior to the reported perceptual transitions, suggests that articulatory-based representations may play a key part in the endogenously driven emergence of auditory speech percepts. This result appears consistent with recent neurobiological models of speech perception and language understanding that claim for a tight connection between speech perception and production systems (Hickok and Poeppel, 2000, 2004, 2007; Scott and Johnsrude, 2003; Callan et al., 2004; Wilson and Iacoboni, 2006; Skipper et al., 2007).

Acknowledgments

This work was supported by CNRS (Centre National de la Recherche Scientifique) and INSERM (Institut National de la Santé et de la Recherche Médicale). M.S. was supported by a Richard H. Tomlinson Postdoctoral Research Fellowship. We thank the two patients for their contribution to this study and Christian Abry, Vincent Gracco and Pascale Tremblay for their useful comments on this study.

References

Aoki, F., Fetz, E.E., Shupe, L., Lettich, E., Ojemann, G.A., 1999. Increased gamma-range activity in human sensorimotor cortex during performance of visuomotor tasks. Clin. Neurophysiol. 110, 524–537.

Bidet-Caulet, A., Fischer, C., Bauchet, F., Bertrand, O., 2003. Multiple sources of induced gamma oscillations in the auditory cortex observed from human intracranial EEG. Neuroimage 19, 1554.

Binder, J.R., Frost, J.A., Hammeke, T.A., Bellgowan, P.S., Springer, J.A., Kaufman, J.N., Possing, E.T., 2000. Human temporal lobe activation by speech and nonspeech sounds. Cereb. Cortex 10, 512–528.

Binder, J.R., Liebenthal, E., Possing, E.T., Medler, D.A., DouglasWard, B., 2004. Neural correlates of sensory and decision processes in auditory object identification. Nat. Neurosci. 7, 295–301.

Blake, R., Logothetis, N.K., 2002. Visual competition. Nat. Rev. Neurosci. 3 (1), 13–21.

Brovelli, A., Lachaux, J.-P., Kahane, P., Boussaoud, D., 2005. High gamma frequency oscillatory activity dissociates attention from intention in the human premotor cortex. Neuroimage 28, 154–164.

Burton, M.W., Small, S.L., 2006. Functional neuroanatomy of segmenting speech and nonspeech. Cortex 42 (4), 644–651.

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

Carter, C.S., Braver, T.S., Bach, D.M., Botvinick, M.M., Noll, D., Cohen, J.D., 1998. Anterior cingulate cortex, error detection, and the online monitoring of performance. Science 280, 747–749.

Cohen, J.D., Perlstein, W.M., Braver, T.S., Nystrom, L.E., Noll, D.C., Jonides, J., Smith, E.E., 1997. Temporal dynamics of brain activation during a working memory task. Nature 386, 604–608.

Crick, F., Koch, C., 2003. A framework for consciousness. Nat. Neurosci. 6 (2), 119–126.

Crone, N.E., Miglioretti, D.L., Gordon, B., Sieracki, J.M., Wilson, M.T., Uematsu, S., Lesser, R.P., 1998a. Functional mapping of human sensorimotor cortex with electrocorticographic spectral analysis. I. Alpha and beta event-related desynchronization. Brain 121 (12), 2271–2299.

Crone, N.E., Miglioretti, D.L., Gordon, B., Lesser, R.P., 1998b. Functional mapping of human sensorimotor cortex with electrocorticographic spectral analysis. II. Event-related synchronization in the gamma band. Brain 121 (12), 2301–2315.

Crone, N.E., Boatman, D., Gordon, B., Hao, L., 2001a. Induced electrocorticographic gamma activity during auditory perception. Clin. Neurophysiol. 112, 565–582.

Crone, N.E., Hao, L., Hart Jr., J., Boatman, D., Lesser, R.P., Irizzary, R., Gordon, B., 2001b. Electrocorticographic gamma activity during word production in spoken and sign language. Neurology 57 (11), 2045–2053.

Crone, N.E., Sinai, A., Korzeniewska, A., 2006. High-frequency gamma oscillations and human brain mapping with electrocorticography. Prog. Brain Res. 159, 275–295.

Dehaene, S., Naccache, L., 2001. Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework. Cognition 79 (1/2), 1–37.

Deiber, M.P., Ibanez, V., Sadato, N., Hallett, M., 1996. Cerebral structures participating in motor preparation in humans: a positron emission tomography study. J. Neurophysiol. 75, 233–247.

Démonet, J.-F., Chollet, F., Ramsay, S., Cardebat, D., Nespoulous, J.-L., Wise, R., Rascol, A., Frackowiak, R.S.J., 1992. The anatomy of phonological and semantic processing in normal subjects. Brain 115, 1753–1768.

Démonet, J.-F., Price, C., Wise, R., Frackowiak, R.S.J., 1994. A pet study of cognitive strategies in normal subjects during language tasks: Influence on phonetic ambiguity and sequence processing on phoneme monitoring. Brain 117 (4), 671–682.

Démonet, J.F., Thierry, G., Cardebat, D., 2005. Renewal of the neurophysiology of language: functional neuroimaging. Physiol. Rev. 85 (1), 49–95.

Edwards, E., Soltani, M., Deouell, L.Y., Berger, M.S., Knight, R.T., 2005. High gamma activity in response to deviant auditory stimuli recorded directly from human cortex. J. Neurophysiol. 94, 4269–4280.

Fell, J., Klaver, P., Lehnertz, K., Grunwald, T., Schaller, C., Elger, C.E., Fernandez, G., 2001. Human memory formation is accompanied by rhinal–hippocampal coupling and decoupling. Nat. Neurosci. 4, 1259–1264.

Ford, J.M., Gray, M., Faustman, W.O., Heinks, T.H., Mathalon, D.H., 2005. Reduced gamma-band coherence to distorted feedback during speech when what you say is not what you hear. Inter. J. Psychophysiol. 57 (2), 143–150.

Fries, P., 2005. A mechanism for cognitive dynamics: neuronal communication through neuronal coherence. Trends Cogn. Sci. 9 (10), 474–480.

Fries, P., Nikolić, D., Singer, W., 2007. The gamma cycle. Trends Neurosci. 30 (7), 309–316.

Hermann, B.P., Seidenberg, M., Dow, C., Jones, J., Rutecki, P., Bhattacharya, A., Bell, B., 2006. Cognitive prognosis in chronic temporal lobe epilepsy. Ann. Neurol. 60 (1), 80–87.

Hickok, G., Poeppel, D., 2000. Towards a functional neuroanatomy of speech perception. Trends Cogn. Sci. 4 (4), 131–138.

Hickok, G., Poeppel, D., 2004. Dorsal and ventral streams: A framework for understanding aspects of the functional anatomy of language. Cognition 92, 67–99.

Hickok, G., Poeppel, D., 2007. The cortical organization of speech processing. Nat. Rev. Neurosci. 8, 393–402.

Honey, G.D., Bullmore, E., Sharma, T., 2000. Prolonged reaction time to a verbal working memory task predicts increased power of posterior parietal cortical activation. NeuroImage 12, 495–503.

Howard, M.W., Rizzuto, D.S., Caplan, J.B., Madsen, J.R., Lisman, J., Aschenbrenner-Scheibe, R., Schulze-Bonhage, A., Kahana, M.J., 2003. Gamma oscillations correlate with working memory load in humans. Cereb. Cortex 13, 1369–1374.

Jokeit, H., Ebner, A., 2002. Effects of chronic epilepsy on intellectual functions. Prog. Brain Res. 135, 455–463.

Jonides, J., Schumacher, E.H., Smith, E.E., Koeppe, R.A., Awh, E., Reuter-Lorentz, P.A., Marshuetz, C., Willis, C.R., 1998. The role of parietal cortex in verbal working memory. J. Neurosci. 18, 5026–5034.

Kahane, P., Minotti, L., Hoffmann, D., Lachaux, J., Ryvlin, P., 2004. Invasive EEG in the definition of the seizure onset zone: depth electrodes. In: Rosenow, F., Luders, H.O. (Eds.), Handbook of Clinical Neurophysiology. Pre-Surgical Assessment of the Epilepsies With Clinical Neurophysiology and Functional Neuroimaging. Elsevier Science, Amsterdam.

Kaiser, J., Hertrich, I., Ackermann, H., Mathiak, K., Lutzenberger, W., 2005. Hearing lips: gamma-band activity during audiovisual speech perception. Cereb. Cortex 15 (5), 646–653.

Kast, B., 2001. Decisions, decisions... Nature 411 (6834), 126–128.

Keil, A., Müller, M.M., Ray, W.J., Gruber, T., Elbert, T., 1999. Human gamma band activity and perception of a gestalt. J. Neurosci. 19 (16), 7152–7161.

Kondo, H.M., Kashino, M., 2007. Neural mechanisms of auditory awareness underlying verbal transformations. NeuroImage 36, 123–130.

Lachaux, J.-P., Rodriguez, E., Martinerie, J., Adam, C., Hasboun, D., Varela, F.J., 2000. A quantitative study of gamma-band activity in human intracranial recordings triggered by visual stimuli. Eur. J. Neurosci. 12, 2608–2622.

Lachaux, J.-P., Lutz, A., Rudrauf, D., Cosmelli, D., Le Van Quyen, M., Martinerie, J., Varela, F., 2002. Estimating the time-course of coherence between single-trial brain signals: an introduction to wavelet coherence. Neurophysiologie clinique (Paris) 32 (3), 157–174.

Lachaux, J.P., Rudrauf, D., Kahane, P., 2003. Intracranial EEG and human brain mapping. J. of Physiol. (Paris) 97, 613–628.

Lachaux, J.-P., George, N., Tallon-Baudry, C., Martinerie, J., Hugueville, L., Minotti, L., Kahane, P., Renault, B., 2005. The many faces of the gamma band response to complex visual stimuli. Neuroimage 25, 491–501.

Lachaux, J.-P., Jerbi, K., Bertrand, O., Minotti, L., Hoffmann, D., Schoendorff, B., Kahane, P., 2007. A blueprint for real-time functional mapping via human intracranial recordings. PLoS ONE 2 (10), e1094.

Leopold, D.A., Logothetis, N.K., 1999. Multistable phenomena: changing views in perception. Trends Cogn. Sci. 3 (7), 254–264.

MacKay, D.G., Wulf, G., Yin, C., Abrams, L., 1993. Relations between word perception and production: new theory and data on the verbal transformation effect. J. Mem. Lang. 32, 624–646.

Mainy, N., Kahane, P., Minotti, L., Hoffmann, D., Bertrand, O., Lachaux, J.-P., 2007. Neural correlates of consolidation in verbal working memory. Hum. Brain Mapp. 28 (3), 183–193.

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

Oldfield, R.C., 1971. The assessment and analysis of handedness: the Edinburgh inventory. Neuropsychologia 9, 97–114.

Palva, S., Palva, J.M., Shtyrov, Y., Kujala, T., Ilmoniemi, R.J., Kaila, K., Näätänen, R., 2002. Distinct gamma-band evoked responses to speech and nonspeech sounds in humans. The J. Neurosci. 22 (4), RC211.

Paulesu, E., Frith, C.D., Frackowiak, R.S.J., 1993. The neural correlates of the verbal components of working memory. Nature 362, 342–344.

Pfurtscheller, G., Graimann, B., Huggins, J.E., Levine, S.P., Schuh, L.A., 2003. Spatiotemporal patterns of beta desynchronization and gamma synchronization in corticographic data during self-paced movement. Clin. Neurophysiol. 114, 1226–1236.

Pitt, M., Shoaf, L., 2001. The source of a lexical bias in the verbal transformation effect. Lang. Cogn. Processes 16 (5/6), 715–721.

Pitt, M., Shoaf, L., 2002. Linking verbal transformations to their causes. J. Exper. Psychol.: Hum., Percept. Perfor. 28 (1), 150–162.

Poldrack, R.A., Wagner, A.D., Prull, M.W., Desmond, J.E., Glover, G.H., Gabrieli, J.D.E., 1999. Functional specialization for semantic and phonological processing in the left inferior prefrontal cortex. NeuroImage 10, 15–35.

Pulvermuller, F., Huss, M., Kherif, F., Moscoso del Prado Martin, F., Hauk, O., Shtyrov, Y., 2006. Motor cortex maps articulatory features of speech sounds. Proc. Natl. Acad. Sci. U. S. A. 103 (20), 7865–7870.

Reisberg, D., Smith, J.D., Baxter, A.D., Sonenshine, M., 1989. "Enacted" auditory images are ambiguous; "pure" auditory images are not. Quart. J. Exper. Psychol. 41A, 619–641.

Romero, L., Walsh, V., Papagno, C., 2006. The neural correlates of phonological short-term memory: a repetitive transcranial magnetic stimulation study. J. Cogn. Neurosci. 18 (7), 1147–1155.

Rushworth, M.F.S., Krams, M., Passingham, R.E., 2001. The attentional role of the left parietal cortex: the distinct lateralization and localization of motor attention in the human brain. J. Cogn. Neurosci. 13, 698–710.

Sato, M., Baciu, M., Lœvenbruck, H., Schwartz, J.-L., Cathiard, M.-A., Segebarth, C., Abry, C., 2004. Multistable representation of speech forms: an fMRI study of verbal transformations. NeuroImage 23 (3), 1143–1151.

Sato, M., Schwartz, J.-L., Abry, C., Cathiard, M.-A., Lœvenbruck, H., 2006. Multistable syllables as enacted percept: a source of an asymmetric bias in the verbal transformation effect. Percept. Psychophys. 68 (3), 458–474.

Sato, M., Basirat, A., Schwartz, J.-L., 2007a. Visual contribution to the multistable perception of speech. Percept. Psychophys. 69 (8), 1360–1372.

Sato, M., Vallée, N., Schwartz, J.-L., Rousset, I., 2007b. A perceptual correlate of the labial-coronal effect. J. Speech Lang. Hear. Res. 50 (6), 1466–1480.

Scott, S.K., Johnsrude, I.S., 2003. The neuroanatomical and functional organization of speech perception. Trends Neurosci. 26 (2), 100–107.

Scott, S.K., Blank, C.C., Rosen, S., Wise, R.J., 2000. Identification of a pathway for intelligible speech in the left temporal lobe. Brain 123, 2400–2406.

Shoaf, L., Pitt, M., 2002. Does node stability underlie the verbal transformation effect? A test of node structure theory. Percept. Psychophys. 64 (5), 795–803.

Skipper, J.I., Van Wassenhove, V., Nusbaum, H.C., Small, S.L., 2007. Hearing lips and seeing voices: how cortical areas supporting speech production mediate audiovisual speech perception. Cereb. Cortex.

Smith, J.D., Reisberg, D., Wilson, M., 1995. The role of subvocalization in auditory imagery. Neuropsychologia 11, 1433–1454.

Szurhaj, W., Bourriez, J.L., Kahane, P., Chauvel, P., Mauguiere, F., Derambure, P., 2005. Intracerebral study of gamma rhythm reactivity in the sensorimotor cortex. Eur. J. Neurosci. 21, 1223–1235.

Talairach, J., Tournoux, P., 1988. A co-planar stereo-taxic atlas of human brain. Stuttg. Thieme.

Tallon-Baudry, C., Bertrand, O., 1999. Oscillatory gamma activity in humans and its role in object representation. Trends Cogn. Sci. 3 (4), 151–162.

Tallon-Baudry, C., Bertrand, O., Delpuech, C., Permier, J., 1997. Oscillatory gamma-band (30–70 Hz) activity induced by a visual search task in humans. J. Neurosci. 17, 722–734.

Tallon-Baudry, C., Bertrand, O., Henaff, M.A., Isnard, J., Fischer, C., 2005. Attention modulates gamma-band oscillations differently in the human lateral occipital cortex and fusiform gyrus. Cereb. Cortex 15, 654–662.

Tanji, K., Suzuki, K., Delorme, A., Shamoto, H., Nakasato, N., 2005. High-frequency gamma-band activity in the basal temporal cortex during picture-naming and lexical-decision tasks. J. Neurosci. 25, 3287–3293.

Varela, F., Lachaux, J.P., Rodriguez, E., Martinerie, J., 2001. The brainweb: phase synchronization and large-scale integration. Nat. Rev. Neurosci. 2 (4), 229–239.

Vigneau, M., Beaucousin, V., Hervé, P.Y., Duffau, H., Crivello, F., Houdé, O., Mazoyer, B., Tzourio-Mazoyer, N., 2006. Meta-analyzing left hemisphere language areas: phonology, semantics, and sentence processing. NeuroImage 30 (4), 1414–1432.

Warren, M.R., 1961. Illusory changes of distinct speech upon repetition — The verbal transformation effect. British J. Psychol. 52, 249–258.
Warren, M.R., Gregory, R.L., 1958. An auditory analogue of the visual reversible figure. Am. J. Psychol. 71, 612–613.
Warren, M.R., Meyers, D.M., 1987. Effects of listening to repeated syllables: category boundary shifts versus verbal transformation. J. Phon. 15, 169–181.
Watkins, K.E., Paus, T., 2004. Modulation of motor excitability during speech perception: the role of Broca's area. J. Cogn. Neurosci. 16 (6), 978–987.
Wilson, S.M., Iacoboni, M., 2006. Neural responses to non-native phonemes varying in producibility: evidence for the sensorimotor nature of speech perception. NeuroImage 33 (1), 316–325.
Wilson, S.M., Saygin, A.P., Sereno, M.I., Iacoboni, M., 2004. Listening to speech activates motor areas involved in speech production. Nat. Neurosci. 7, 701–702.
Zatorre, R.J., Evans, A.C., Meyer, E., Gjedde, A., 1992. Lateralization of phonetic and pitch discrimination in speech processing. Science 256, 846–849.

Chapitre 6

Discussion générale

À travers le paradigme de la multistabilité perceptive en parole, les expériences comportementales présentées dans cette partie mettent en évidence l'influence des informations visuelles sur l'organisation perceptive de la parole et suggèrent quelques caractéristiques des mécanismes à la base de cette influence. En ce qui concerne les activités cérébrales en lien avec les transformations verbales, notre expérience iEEG montre une activation plus ample du gyrus frontal inférieur gauche et du gyrus supramarginal gauche 300-800 ms avant les bascules perceptives signalées par les sujets. Ces résultats, reliés à ceux obtenus par Sato et collègues montrant l'effet des contraintes articulatoires sur les transformations verbales (Sato *et al.*, 2006, 2007) et en accord avec les études IRMf utilisant le paradigme de transformation verbale (Sato *et al.*, 2004; Kondo et Kashino, 2007), esquissent un cadre perceptuo-moteur très cohérent pour expliquer l'effet de transformation verbale que nous détaillerons dans la suite.

En premier lieu, nos résultats peuvent être interprétés en lien avec la proposition de Skipper et collègues sur le rôle des liens perceptuo-moteurs dans la perception audio-visuelle de la parole au sein d'un réseau temporo-pariéto-frontal similaire au circuit dorsal du traitement de la parole (Skipper *et al.*, 2005, 2007, voir sous-section 2.3.5). Nous rappelons que selon cette proposition les liens perceptuo-moteurs interviendraient dans la perception audio-visuelle de la parole en effectuant une prédiction sur l'entrée multimodale par une méthode « d'analyse-par-synthèse ». Cette intervention du système moteur serait particulièrement importante en l'absence d'informations suffisantes par exemple en langue étrangère (Callan *et al.*, 2004) ou en présence du bruit (Binder *et al.*, 2004). Ainsi, les informations visuelles et les liens perceptuo-moteurs viennent à l'aide de la perception en indiquant où les informations manquantes peuvent être trouvées (Schwartz *et al.*, sous presse). En accord avec ce scénario, Schwartz *et al.* (2004) montrent que dans la présentation audio-visuelle des séquences /tu/ et /du/ en présence de bruit, la modalité visuelle permet une meilleure distinction entre séquences probablement en fournissant à l'auditeur des indices temporels permettant la détection du prévoisement dans le flux auditif. Dans ce cadre, l'effet de transformation verbale, par sa nature « ambiguë », serait un paradigme qui demanderait une implication importante des liens perceptuo-moteurs afin de permettre une organisation cohérente et robuste de la scène multimodale de la parole.

L'implication des processus moteurs via le circuit dorsal dans l'effet de transformation verbale serait également cohérente avec les mécanismes proposés par Leopold et Logothetis (1999) pour expliquer la multistabilité perceptive. Comme décrit dans la section 3.1, ces auteurs suggèrent que les bascules perceptives sont pro-

duites par des mécanismes actifs de réorganisation perceptive initiés dans les aires pariéto-frontales de l'intégration sensori-motrice. Ces mécanismes interviendraient aussi bien dans la perception « normale » (non multistable) que dans la perception multistable, notamment en lien avec l'expression d'un comportement. Selon les auteurs, cette coordination sensorimotrice serait nécessaire pour la conscience perceptive de l'environnement. Ainsi, l'émergence des transformations verbales sous certaines contraintes articulatoires au sein du circuit dorsal et l'influence des gestes visibles sur l'organisation de ces transformations pourraient refléter les mécanismes généraux de l'organisation perceptive.

Un autre lien possible entre nos résultats et les mécanismes généraux de l'organisation perceptive concerne le couplage pariéto-frontal impliqué dans l'émergence des nouveaux percepts de parole et dans la perception multistable en vision. Comme décrit dans la sous-section 3.1.2, ce couplage a été associé à l'intégration haut-niveau entre les représentations distribuées et multisensorielles impliquées dans la conscience visuelle qui fait partie intégrante du phénomène de multistabilité (Naghavi et Nyberg, 2005). Le rôle de ce couplage a été également souligné en lien avec la redirection de l'attention vers l'entrée sensorielle, ce qui permettrait une réévaluation du percept conduisant à une possible bascule perceptive (Sterzer *et al.*, 2009). Il est important de noter qu'en fonction de la modalité sensorielle, ce couplage s'effectuerait entre différentes régions pariéto-frontales. Dans une tâche de reconnaissance des objets, visuels et auditifs impliquant les mécanismes sous-jacent de la conscience perceptive, Eriksson *et al.* (2007) ont observé que les régions frontales interagissent avec les régions spécifiques postérieures en fonction de la modalité des stimuli : un couplage pariéto-frontal a été observé en lien avec l'émergence des percepts visuels et un couplage temporo-frontal a été observé en lien avec l'émergence des percepts auditifs. Dans ce sens, l'activation du gyrus supramarginal gauche en lien avec l'émergence des nouveaux percepts parole peut refléter le caractère spécifique à la modalité du couplage pariéto-frontal, en l'occurrence, dans notre cas, la parole.

Il est important de rappeler que les mécanismes généraux d'analyse de scènes auditives sont également présents dans l'effet de transformation verbale. Comme caractérisé par (Pitt et Shoaf, 2002), les transformations de type streaming révèlent l'implication des processus de regroupement auditivo-perceptif dans l'effet de transformation verbale (voir sous-section 3.3.1). Ce type de transformations était également présent dans nos expériences surtout pour la séquence auditive /psə/. Cependant, en accord avec la proposition de Remez *et al.* (1994), nos résultats sur les transformations verbales audio-visuelles soulignent que l'organisation de la scène de parole ne peut pas être entièrement expliquée par les principes de l'analyse de scènes auditives. En effet, la présentation audio-visuelle de la séquence /psə/ a réduit, voire éliminé, les transformations de type streaming qui étaient présentes lors d'une présentation purement auditive de la séquence /psə/.

Face à ces résultats, il nous semble que les principes de l'organisation de la scène de parole sont une combinaison entre des principes auditivo-phonétiques (comme proposé par Bregman, 1990) et des principes sensori-moteurs spécifiques à la parole. Les études présentées dans cette partie, en lien avec celles réalisées auparavant au

département Parole et Cognitition du Gipsa-lab (Schwartz *et al.*, 2004; Sato *et al.*, 2004, 2006, 2007) suggèrent quelques caractéristiques de ces principes mulitmodaux et perceptuo-moteurs. En se basant sur ces études, nous proposons que l'organisation préférentielle d'une scène de parole est celle des formes ayant une cohérence articulatoire importante, ce qui peut être caractérisé par un geste d'ouverture de la mâchoire. Cette proposition expliquerait la préférence de /psə/ sur /səp/ et la préférence de /pata/ sur /tapa/ (mais aussi, plus basiquement, la préférence d'une organisation CV sur une organisation VC, classique dans les expériences de transformations verbales). Les informations guidant cette organisation pourraient être récupérées aussi bien à partir d'un flux de parole auditif que visuel. Plus particulièrement, l'onset visuel associé au geste d'ouverture de la mâchoire jouerait un rôle dans la segmentation du flux de parole. Ces éléments fournissent une première base pour un modèle computationnel de la perception de la parole prenant en compte l'effet de transformation verbale.

Troisième partie

Éléments de modélisation

Projet de modélisation

Les résultats expérimentaux présentés dans les chapitres 4 et 5 s'inscrivent dans un cadre cohérent, multimodal et perceptuo-moteur pour expliquer l'effet de transformation verbale, et par extension, l'organisation perceptive de la parole. Si l'effet de transformation verbale a beaucoup été étudié expérimentalement, aucun modèle computationnel réaliste n'a été proposé et confronté à l'ensemble de la phénoménologie. En nous appuyant sur ces résultats expérimentaux, nous avons travaillé dans la suite de cette thèse sur une modélisation computationnelle cognitive de l'effet de transformation verbale. Pour cela, nous nous sommes concentrés sur les stimuli de type /pata/ et /tapa/. Ces stimuli ont permis de démontrer à la fois les effets perceptuo-moteurs et audio-visuels sur les transformations verbales. Les résultats des expériences utilisant ce type de stimuli nous ont fourni deux éléments sur lesquels nos travaux de modélisation sont bâtis. Nous décrivons ces deux éléments dans les paragraphes suivantes.

Rappelons d'abord que, comme présenté dans la sous-section 3.3.2, il existe une tendance en faveur des séquences Labial-Coronal (LC) comme /pata/ par rapport à celles de type Coronal-Labial (CL) comme /tapa/ dans les langues du monde et lors de l'acquisition du langage. Cet effet, appelé l'effet LC, semble avoir une explication articulatoire : les séquences LC peuvent être réalisées en un seul cycle de mâchoire grâce à l'anticipation du geste de la consonne coronale, ce qui n'est pas possible pour les séquences CL. Il existe également une asymétrie perceptive entre ces deux formes : dans une tâche de transformation verbale, le percept /pata/ est un attracteur perceptif plus fort que /tapa/. Cette asymétrie a été expliquée par la cohésion articulatoire présentée ci-dessus (pour les détails et les références, voir sous-section 3.3.2). En se basant sur ces études, une hypothèse mise en place dans notre modèle sera la suivante : le liage perceptif de parole se fait par la cohésion articulatoire. Le liage préférentiel est ainsi celui qui regroupe les matériaux sensoriels contenus dans un geste d'ouverture de la mâchoire, c'est-à-dire passant d'une configuration fermée à une configuration plus ouverte.

Notre étude sur l'effet de l'onset audio-visuel présentée dans la sous-section 4.3 est cohérente avec cette proposition. Nous avons vu que si l'on superpose au stimulus auditif /pata/ ou /tapa/ un geste visuel qui ne porte que sur une des deux syllabes (/pa#a/ ou /ta#a/), ce geste renforce la stabilité du percept commençant par cette syllabe (respectivement /pata/ ou /tapa/), ce que nous avons interprété comme un effet de liage de l'information déclenché par le geste d'ouverture de la mâchoire, visible dans le stimulus visuel. Ainsi, le deuxième élément central de ce travail sera le suivant : les informations sur le geste d'ouverture de la mâchoire peuvent être récupérées aussi bien à partir du signal auditif que visuel.

L'objectif de ce projet de modélisation est de mettre en place les deux éléments présentés ci-dessus (liage articulatoire porté par le geste d'ouverture de la mâchoire et nature audio-visuelle du liage perceptif) dans un cadre computationnel cognitif. Le phénomène de l'effet de transformation verbale étant produit par les mécanismes

généraux de la perception de la parole, il nous semble qu'une plateforme adaptée à la modélisation de ce phénomène est celle d'un modèle de la perception « normale » (non multistable) de la parole. Autrement dit, nos composantes computationnelles doivent être intégrées à un modèle général de la perception de la parole, ce qui fait l'objet de la suite des travaux de cette thèse.

Cette partie comprend deux chapitres. Le premier chapitre consiste en un bref état de l'art sur la modélisation computationnelle cognitive notamment dans le domaine de la perception de la parole. Le deuxième chapitre présente nos travaux de modélisation.

Chapitre 7

Modélisation de la perception multistable de la parole

Sommaire

La première section de ce chapitre porte sur une brève présentation de la modélisation computationnelle cognitive, ses objectifs, les défis qu'elle rencontre et ses méthodes. La deuxième et la troisième section présentent une revue de la littérature sur la modélisation psycholinguistique de la perception de la parole et sur les modèles de l'effet de transformation verbale.

7.1 Modélisation computationnelle cognitive

7.1.1 Rôles et objectifs

Explorer les conséquences des idées théoriques

Selon McClelland (2009), l'objectif de la modélisation cognitive est d'étudier et d'explorer les conséquences des idées au-delà des limites de la capacité de la pensée humaine[1]. Dans une théorie verbalement exprimée, les conséquences des mécanismes et leurs interactions ne sont pas toujours connues tandis qu'une fois les idées mises

[1] "The essential purpose of congitive modeling is to allow the investigation of the implications of ideas, beyond the limits of human thinking." (McClelland, 2009, p. 16)

en place sous forme d'un modèle computationnel, nous pouvons facilement suivre ces mécanismes et leurs interactions malgré leur complexité.

La modélisation révèle si un ensemble d'idées théoriques est incohérent ou incomplet : un modèle implémenté sous forme d'un programme informatique ne fonctionne pas s'il rencontre des conflits. Dans ce cas, la modélisation permet de corriger la théorie (Mareschal et Thomas, 2007).

Réfléchir aux détails

Selon Norris (2005), le premier avantage de la modélisation computationnelle est le fait que la modélisation nous oblige à réfléchir aux détails qui ne sont pas toujours considérés dans une théorie[2]. Lorsqu'un modélisateur formule les idées théoriques sous forme d'un modèle et d'un programme informatique, il se rend souvent compte qu'à cet ensemble d'idées, il manque des éléments supplémentaires ou qu'un mécanisme ne fonctionne pas de la façon prévue dans la théorie. Ce défi rencontré pendant la modélisation contribue ainsi aux progrès dans la compréhension des processus cognitifs.

Prédiction

Un modèle computationnel permet de prédire des résultats précis dans les conditions qui ne sont pas testées de façon expérimentale. Ainsi, le modèle nous fournit de nouvelles hypothèses concernant les processus cognitifs (Plaut, 2000) qui sont à leur tour vérifiables en réalisant des expériences ciblées.

7.1.2 Quelques problèmes pendant la modélisation

Simplification

L'objectif de la modélisation computationnelle d'un processus cognitif est de contribuer à comprendre les mécanismes impliqués et pas seulement de reproduire le phénomène étudié. C'est pourquoi il faut éviter des modèles complexes qui rendent le modèle peu compréhensible et qui ont besoin de paramètres adhocs (McClelland, 2009). Pour modéliser un processus cognitif, plusieurs approximations/simplifications sont souvent possibles. Le choix de l'approximation/simplification dépend de ce que nous voulons modéliser et de la connaissance déjà acquise dans le domaine étudié (Mareschal et Thomas, 2007).

Différents types d'hypothèses

Lorsqu'on fait une modélisation, il faut considérer et séparer deux types d'hypothèses. Le premier type concerne les idées fondamentales du modèle, autrement dit, les hypothèses théoriques. Par exemple, dans le cadre du modèle Shortlist de la perception de la parole proposé par Norris (1994), il n'y a pas de feedback du niveau lexical vers le niveau pré-lexical (voir sous-section 7.2.3). Cette hypothèse

[2] "Modeling makes you think." (Norris, 2005, p. 334)

est une idée fondamentale du modèle selon l'auteur (Norris, 2005). Le deuxième type regroupe les hypothèses non-cruciales, additionnelles, faites pour que le modèle fonctionne lorsqu'il n'y pas de contraintes théoriques ou afin de simplifier le modèle. Par exemple dans le modèle Shortlist, l'entrée du modèle peut être aussi bien une suite de phonèmes qu'une suite de traits phonétiques. Dans ce modèle, il n'existe pas d'hypothèse théorique sur le choix d'entrée mais pour simplifier, l'entrée du modèle est décrite sous forme de phonèmes. Ce choix peut être modifié sans toucher aux principes du modèle (Norris, 2005). Dans un modèle, il faut séparer ces deux types d'hypothèses de sorte que le résultat du modèle soit la conséquence directe des hypothèses théoriques et non des hypothèses « périphériques », ce qui est parfois difficile à vérifier (Plaut, 2000).

Ajustement aux données expérimentales

Une des critiques envers la modélisation computationnelle concerne un critère souvent utilisé pour évaluer un modèle : la qualité d'ajustement des résultats du modèle aux données expérimentales (Roberts et Pashler, 2000). En effet, un modèle peut produire des résultats ajustables à n'importe quelles données (ou presque) grâce à ses paramètres sans qu'il explique nécessairement les processus cognitifs en jeu. L'objectif de la modélisation cognitive n'est pas de proposer un modèle qui produit des résultats compatibles avec les données expérimentales mais de comprendre pourquoi le modèle fonctionne et peut produire ces résultats. Face à ce problème, outre la qualité de l'ajustement (*goodness of fit*), les modélisateurs utilisent des méthodes pour évaluer leurs modèles (par exemple, la généralisabilité, voir Pitt et Myung, 2002).

7.1.3 Différents types de modèles computationnels cognitifs

Les modèles computationnels cognitifs peuvent être classés en cinq catégories (Sun, 2008; McClelland, 2009). Ces différentes catégories sont brièvement présentées dans la suite.

Modèles connexionnistes

Les modèles connexionnistes consistent en des unités simples interconnectées entre elles par des connexions affectées de poids. Chaque unité a une activité qui peut changer au cours du temps et qui se propage entre les unités inter-connectées. Dans un modèle connexionniste, il faut préciser les éléments suivants :

- Structure : un modèle connexionniste peut être sans ou avec des connexions créant des cycles. Une structure sans cycle est appelée *feedforward* et celle avec des cycles est appelée récurrente (figure 7.1).

- Activation : l'activation d'une unité au sein d'un modèle connexionniste est fonction de ses entrées et de sa fonction d'activation. La fonction d'activation peut être linéaire ou non, déterministe ou probabiliste.

- Apprentissage : les poids des connexions d'un modèle se modifient au cours de la phase d'apprentissage par rapport aux données utilisées comme échantillon d'apprentissage. Il existe différents algorithmes d'apprentissage, supervisés ou non.

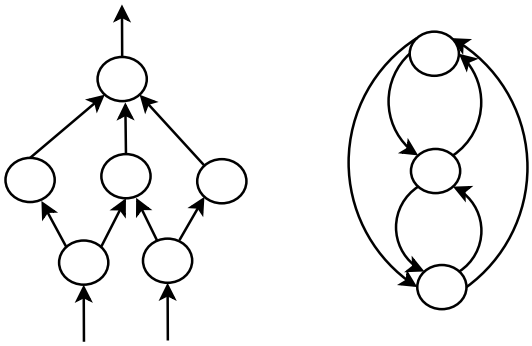


FIGURE 7.1 : Modèles connexionnistes : à gauche une structure *feedforward* et à droite une structure récurrente

Les modèles connexionnistes sont souvent utilisés pour modéliser des tâches cognitives qui comprennent des processus relativement automatiques basés sur des expériences. Un modèle connexionniste peut extraire la régularité dans un ensemble de données et généraliser sans que les règles explicites soient nécessaires (McClelland, 2009).

Modèles Bayésiens

Les modèles Bayésiens utilisent les principes Bayésiens pour comprendre et expliquer comment le cerveau résoudrait le problème de l'inférence (Griffiths *et al.*, 2008). L'inférence Bayésienne se réalise grâce à la règle de Bayes : soient deux variables aléatoires A et B, l'équation 7.1 indique comment calculer la probabilité conditionnelle de $A = a$ sachant $B = b$ en sachant la probabilité conditionnelle de $B = b$ sachant $A = a$.

$$P(A = a|B = b) = \frac{p(B = b|A = a)P(A = a)}{P(B = b)} \tag{7.1}$$

Si on appelle A l'état du monde et B l'entrée sensorielle, la règle de Bayes pourra être appliquée au cerveau de la façon suivante (Wolpert et Ghahramani, 2005) :

$$\overbrace{P(\text{état}|\text{entrée sensorielle})}^{\text{a posteriori}} = \frac{\overbrace{p(\text{entrée sensorielle}|\text{état})}^{\text{vraisemblance}}\overbrace{P(\text{état})}^{\text{a priori}}}{P(\text{entrée sensorielle})} \tag{7.2}$$

L'état du monde peut représenter ce que nous voulons estimer par exemple l'identité d'un objet visuel, la place de nos acteurs moteurs au cours d'une action motrice, etc. $P(\text{état})$ représente notre connaissance *a priori* d'un état avant la réception de l'entrée sensorielle ou indépendante de celle-ci. Ces connaissances peuvent être apprises

au cours de nos expériences. P(entrée sensorielle|état) représente la probabilité de l'entrée sensorielle sachant l'état du monde, ce qui est appelée la vraisemblance (*liklihood*), par exemple : la probabilité de recevoir l'entrée sensorielle e sachant que l'objet visuel est l'objet o. Ainsi, la règle de Bayes, illustrée par l'équation 7.2, peut indiquer la probabilité de chaque état sachant une entrée sensorielle.

Les modèles Bayésiens correspondent au niveau computationnel de la théorie de Marr (1982). Les modèles correspondant au niveau algorithmique ou au niveau d'implémentation matérielle de Marr expliquent la cognition du point de vue des mécanismes impliqués, ce qui rend nécessaire, la plupart du temps, des hypothèses sur des mécanismes inconnus. Au contraire, dans un modèle Bayésien, on n'a pas besoin d'hypothèses supplémentaires car le système met à jour ses probabilités *a posteriori* utilisant les connaissances déjà acquises et des probabilités *a priori*.

Les modèles Bayésiens sont souvent utilisés dans les domaines du raisonnement inductif[3] et de la prise de la décision tandis que la déduction[4], planning ou résolution de problème (*problem solving*) ne sont traditionnellement pas modélisés par les modèles Bayésiens (Griffiths *et al.*, 2008).

Modèles de type système dynamique

Un système dynamique est caractérisé par trois éléments : un espace des états du système S (ensemble des états possibles), un ensemble de temps T et une règle R de l'évolution temporelle du système $R : S \times T \to S$, ce qui conduit l'état du système à $s \in S$. Ainsi, un système dynamique peut être considéré comme un modèle de l'évolution temporelle du système.

Différents types de processus cognitifs sont modélisés par les systèmes dynamiques notamment dans les domaines de la perception et du contrôle moteur. Par exemple le modèle ART (*Adaptive Resonance Theory*) (Carpenter et Grossberg, 1987) est un modèle de la perception générale utilisant l'approche des systèmes dynamiques (une brève présentation du modèle ART sera présentée dans la partie 7.2.3). Il est à noter qu'un modèle connexionniste de type récurrent peut être défini par des équations dynamiques et considéré comme un système dynamique.

Les systèmes dynamiques sont également utilisés pour modéliser les processus cognitifs de plus haut-niveau comme la prise de la décision (Townsend et Busemeyer, 1995). Selon l'approche propre aux systèmes dynamiques, les processus cognitifs de haut-niveau peuvent être considérés comme la conséquence des effets de bas-niveau sensibles au facteur temps (*time-sensitive*), ce qui peut naturellement être modélisé par des équations dynamiques (Port, 2000). Récemment, Lancia et collègues ont proposé un modèle de type système dynamique non-linéaire pour modéliser certaines propriétés de la perception de la parole (Lancia, 2009; Lancia *et al.*, 2008). Un modèle de l'effet de transformation verbale utilisant l'approche système dynamique sera présenté dans la partie 7.3.2.

[3] L'induction consiste à tirer une conclusion générale basée sur les cas particuliers.
[4] La déduction présente les implications des lois en montrant les étapes intermédiaires.

Modèles symboliques et logiques

Un modèle cognitif est appelé symbolique s'il a des propriétés des systèmes physiques symboliques (*physical symbolic system*, *PSS*) définis par Newell et Simon (1976). L'entrée d'un système physique symbolique consiste en des symboles. Le système combine ces symboles en les mettant sous forme de structures appelées expressions. Le système possède des processus qui permettent de créer de nouvelles expressions à partir des symboles d'entrées et des expressions déjà existantes. Par exemple, l'ordinateur est un système physique symbolique qui a le bit (zéro et un) comme symbole. Le processeur central (*CPU*) applique les opérations (processus) aux expressions sauvegardées dans la mémoire pour les modifier/supprimer ou en créer des nouvelles. Selon cette approche, l'intelligence humaine et l'intelligence artificielle sont toutes les deux des systèmes symboliques.

Un exemple connu de modèles symboliques en sciences cognitives est le modèle ACT-R (*Adaptive Control of Thought-Rational*) (Anderson et Lebiere, 1998). Dans ce modèle, il y a deux types de mémoire, l'un sauvegarde les connaissances déclaratives et l'autre les connaissance procédurales. Ces connaissances sont de type si-alors (*if-then*), ce qui est appelé une règle de production[5]. Le modèle ACT-R possède deux niveaux qui peuvent représenter la structure du cerveau :

- Niveau perceptuo-moteur qui est l'interface entre l'environnement et le modèle. Ce niveau contient le module visuel, moteur, auditif, etc.

- Niveau cognitif qui contient la mémoire déclarative et le module procédural.

L'approche symbolique pour modéliser la cognition permet de bonnes simulations des processus cognitifs lorsque les connaissances déclaratives sont fournies au préalable au modèle mais ces systèmes ont des limites quand il s'agit de produire, eux-mêmes, ces connaissances déclaratives (Bringsjord, 2008).

Architectures cognitives et modèles hybrides

Selon Pylyshyn (1991), pour modéliser les processus cognitifs, il faut d'abord définir l'architecture fonctionnelle du système indépendamment des autres aspects du système car les algorithmes utilisés dépendent de l'architecture du modèle. Cette approche amène à définir une architecture fonctionnelle lors la modélisation. Une architecture cognitive peut être classée dans les quatre catégories ci-dessus selon l'approche ou les approches qu'elle utilise. Par exemple, un des modèles récents de type architecture cognitive est le modèle SAL proposé par Jilk *et al.* (2008) qui intègre le modèle symbolique ACT-R et le modèle Leabra qui est une architecture neuronale intégrant les zones du cerveau (O'Reilly et Munakata, 2000). Il est à noter que dans certaines architectures cognitives, l'architecture, elle-même, ne peut pas changer au cours du temps et c'est seulement les informations sauvegardées dans le système qui changent tandis que dans certaines autre, le modèle peut acquérir des

[5] Un exemple d'une règle de production est la suivante : « si le but est de classifier un objet et l'objet possède quatre côtés égaux alors classifie-le comme un carré ».

nouveaux sous-systèmes ou des nouvelles connexions entre les sous-systèmes (pour une revue sur ce sujet, voir Langley *et al.*, 2008).

Un modèle de type architecture cognitive peut être hybride, c'est à dire qu'il peut utiliser différentes approches de modélisation. Deux exemples des architectures cognitive hybrides sont le modèle CLARION (Sun, 2002) et le modèle Dual (Kokinov, 1994). Une architecture peut être hybride au niveau macroscopique, ce qui consiste à avoir des modules séparés par approche (par exemple, des modules symboliques et des modules connexionnistes). Au contraire, un modèle hybride au niveau microscopique possède des agents hybrides mais il n'y a pas de modules séparés pour chaque approche.

7.2 Modèles psycholinguistiques de la perception de la parole

Dans le cadre de notre travail de modélisation, nous nous sommes intéressés aux modèles psycholinguistiques de la perception de la parole. Les modèles psycholinguistiques concernent la modélisation computationnelle du système humain de traitement du langage en utilisant les données fournies par des expériences comportementales. Parmi les modèle psycholinguistiques, nous trouvons les différents types de modèles cognitifs cités ci-dessus tels que les modèles connexionistes, bayésiens, systèmes dynamiques, symboliques ou hybrides. Les modèles psycholinguistiques peuvent être classés en quatre catégories en fonction de leur domaine d'application et la tâche qu'ils effectuent (Lewis, 2000) :

- Modèles de traitement lexical : ces modèles concernent la reconnaissance des mots parlés (*spoken word recognition*) et la reconnaissance des mots visuels (*visual word recognition*).
- Modèles de compréhension : ces modèles concernent la compréhension de phrase. Les modèles de décomposition analytique (*parsing*) et traitement de discours font partie de cette catégorie.
- Modèles de production : ces modèles concernent la production de la parole. La plupart d'entre eux ne s'occupent que des dernières phases de la production, c'est à dire la phase de la production d'une suite de phonèmes correspondant à l'énoncé d'entrée du modèle.
- Modèles d'acquisition : ces modèles concernent l'acquisition du langage par exemple l'acquisition de la syntaxe et de la sémantique, l'acquisition de la langue maternelle et des langues secondes.

Nous nous sommes concentrés dans cette thèse sur les modèles de reconnaissance des mots parlés, aussi appelés modèles de la perception de la parole. Ces modèles essaient de rendre compte de nombreuses caractéristiques de la perception de la parole, ce qui fournit une base intéressante pour la modélisation de l'effet de transformation verbale. Nous reviendrons sur ce point dans la section 7.4 mais avant, nous présenterons une revue de ces modèles.

7.2.1 Questions principales

Les modèles psycholinguistiques de la perception de la parole, malgré leurs différences, proposent un cadre commun pour la perception de la parole. Selon ces modèles, la perception consiste à faire correspondre l'entrée acoustique aux unités représentant ces entrées dans le modèle et ensuite, à faire correspondre ces représentations pré-lexicales de l'entrée aux représentations lexicales. Ce cadre général suscite des débats dans la littérature sur la nature des représentations linguistiques utilisées dans les modèles et sur l'interaction entre les différents niveaux de représentations. Une brève revue sur les questions principales et sur les choix des modèles concernant les représentations et leurs interactions est présentée ci-dessous. Il est à noter que la revue de la littérature sur les résultats expérimentaux en faveur ou contre les choix des modèles va au-delà du cadre de cette thèse, qui se restreint à l'effet de transformation verbale.

Entrée et sa représentation

Les modèles psycholinguistiques de la perception de la parole ne contiennent pas de mécanismes d'analyse pour transformer le signal acoustique d'entrée vers les représentations de nature linguistique utilisées dans le modèle. Ainsi, l'entrée du modèle n'est pas un signal acoustique mais, dans la plupart des modèles, elle est sous forme d'une suite de phonèmes ou de traits phonétiques. Le modèle fait correspondre ces entrées aux représentations linguistiques prévues dans le modèles. En effet, ces modèles, basés sur les expériences psycholinguistiques, n'étudient pas les processus de traitement auditif de la perception de la parole. Cependant, il est possible d'ajouter une composante de traitement du signal auditif à l'entrée de ces modèles, ce qui n'a pas de conséquences sur les problématiques posées, les mécanismes utilisés et les résultats de ces modèles.

Différentes unités linguistiques interagissant avec les représentation lexicales ont été proposées pour servir de représentations pré-lexicales de l'entrée . Par exemple, le modèle TRACE (voir sous-section 7.2.3), le modèle Shortlist (voir sous-section 7.2.3) et le modèle NAM (*Neighborhood Activation Model*, Luce *et al.*, 1990) utilisent des représentations phonémiques avant le niveau lexical. Cependant, des représentations sous-phonémiques et syllabiques ont été également proposées comme des unités de représentations prés-lexicale (par exemple, respectivement, dans Warren et Marslen-Wilson, 1987 et Mehler *et al.*, 1981).

Représentions localisée et distribuée

Dans un modèle psycholinguistique, les représentations peuvent être localisées ou distribuées. Dans un modèle localisé, une seule unité représente un phonème, une syllabe, un mot, etc. L'activation de cette unité représente le degré de cohérence de l'unité avec l'entrée du modèle. Par exemple, les modèles TRACE, Shortlist, ART que nous présenterons dans la suite de ce chapitre sont des modèles localisés. Dans un modèle distribué, une seule unité n'a pas de signification de nature linguistique et c'est l'état entier du système qui correspond à une représentation phonémique, lexi-

cale ou autre. Par exemple, dans le modèle psycholinguistique proposé par Gaskell et Marslen-Wilson (1997), les connaissances lexicales sont modélisées d'une façon distribuée.

Interaction entre le niveau lexical et le niveau pré-lexical

Dans les modèles psycholinguistiques de la perception de la parole, l'activation des représentions lexicales est, bien entendu, basée sur les informations bottom-up fournies par l'entrée. Certains modèles proposent que ce flux d'information de type bottom-up est le seul flux d'information possible dans le système (par exemple le modèle Shortlist) tandis que selon certains autres, le flux d'information peut être à la fois de type bottom-up et top-down. Par exemple dans le modèle TRACE, l'activité des unités lexicales influence l'activité des unités pré-lexicales grâce aux connexions de type feedback. Dans la sous-section 7.2.3, nous présenterons le débat autour de ce sujet.

Activation/sélection des représentations lexicales

Dans la plupart des modèles psycholinguistiques de la perception de la parole, plusieurs candidats lexicaux peuvent être parallèlement actifs. Le degré d'activation de chaque candidat représente son degré de cohérence avec l'entrée. Différents mécanismes ont été proposés concernant la sélection entre les candidats. Cette sélection entraîne la segmentation en ligne de l'entrée. La question de la segmentation étant particulièrement importante dans nos études sur l'effet de transformation verbale, nous présenterons dans la sous-section 7.2.2, une brève revue sur les différentes stratégies de segmentation utilisées dans les modèles psycholinguistiques de la perception de la parole.

Apprentissage

L'apprentissage dans un modèle psycholinguistique de la parole concerne deux aspects :

- Correspondance entre l'entrée du modèle et les représentions linguistiques : l'apprentissage au niveau de la correspondance entre l'entrée et les représentions du modèle consiste à utiliser un algorithme d'apprentissage, souvent la rétro-propagation du gradient d'erreur (*back propagation error*), pour faire correspondre un pattern d'entrée à une représentation en sortie. Un ou plusieurs ensemble d'unités cachées entre le niveau de l'entrée et de la sortie et des connexions récurrentes sont souvent prévues afin de permettre la réalisation de cette correspondance. Le modèle SRN (*Simple Recurrent Network*) (Elman, 1990) utilise ce type d'apprentissage (voir sous-section 7.2.3 dans le modèle Shortlist).

- Acquisition des nouvelles représentions : dans la plupart des modèles psycholinguistiques de la perception de la parole, il n'existe pas de mécanismes d'acquisition des nouvelles représentations. Une exception est le modèle ART

(*Adaptive Resonance Theory*) proposé par Carpenter et Grossberg (1987). Dans le modèle ART, un mécanisme d'apprentissage non-supervisé parallèle à une catégorisation de l'entrée est prévu de sorte que, si l'entrée du modèle et le candidat de la sortie ne sont pas assez cohérents, une nouvelle unité de représentation est créée. La version de base du modèle ART et le modèle ARTWORD de la reconnaissance des mots serons décrits dans la sous-section 7.2.3.

7.2.2 Segmentation du flux de parole

La segmentation du flux de parole est partie intégrante de nos études sur l'effet de transformation verbale car la majorité des transformations verbales observées dans nos expériences comportementales du chapitre 4 correspond aux différentes segmentations de la séquence en boucle (par exemple, les transformations /pata/ et /tapa/ lorsque le stimulus est /patapata.../ ou /psə/ et /səp/ quand le stimulus est /sepsep.../). Dans cette partie, nous présentons brièvement quelques stratégies de segmentation de la parole proposées par les études psycholinguistiques de la perception de la parole (pour une revue plus détaillée, voir Davis, 2000, chap. 2 et Gambell et Yang, 2005)

Segmentation basée sur la probabilité transitionnelle

Selon la stratégie de la segmentation basée sur la probabilité transitionnelle, les corrélations statistiques entre les différentes parties du flux de parole sont utilisées pour effectuer une segmentation. Ainsi, la probabilité transitionnelle entre deux syllabes consécutives indique s'il faut segmenter le flux entre ces deux syllabes. L'équation 7.3 illustre la probabilité transitionnelle (PT) entre les syllabes A et B où $P(AB)$ est la probabilité que la syllabe B suive la syllabe A et $P(A)$ est la probabilité totale de la syllabe A dans une langue.

$$PT(A \rightarrow B) = \frac{P(AB)}{P(A)} \tag{7.3}$$

Les frontières des mots sont situées aux minima locaux où $PT(A \rightarrow B)$ est inférieure à la probabilité transitionnelle de ses voisins. Par exemple, si dans la séquence "joliemaison", $PT(lie \rightarrow mai)$ est inférieur à $PT(jo \rightarrow lie)$ ou $PT(mai \rightarrow son)$, alors, la frontière du mot (minimum local) est identifiée entre *jolie* et *maison*. Saffran et collègues ont montré qu'un enfant de 9 mois aussi bien qu'un adulte, exposé pendant quelques minutes à une langue artificielle composée de pseudo-mots de trisyllabes est capable d'utiliser ces régularités distributionnelles pour segmenter le flux de parole (Saffran *et al.*, 1996a,b).

Segmentation métrique

Les informations sur les patterns rythmiques de la langue ont été également proposées comme indices pour segmenter le flux de parole. Cutler et Norris (1988)

ont proposé qu'en anglais, les auditeurs choisissent les syllabes accentuées (*stressed* ou *strong syllable*) pour début de mot. Cette proposition est basée sur le fait que 90% des mots en anglais commencent avec des syllabes accentuées et 75% des syllabes accentuées apparaissent en début de mot (Cutler et Carter, 1987). Le principe de la segmentation métrique serait universel mais les patterns rythmiques utilisés diffèrent d'une langue à l'autre. Par exemple, il a été proposé que la segmentation est basée sur l'accent en anglais, la syllabe en français (Cutler *et al.*, 1986) et le more en japonais (Otake *et al.*, 1993).

Contraintes phonotactiques

Les contraintes phonotactiques concernent les contraintes des combinaisons de phonèmes et de structures syllabiques pour une langue donnée. Les informations sur ces contraintes peuvent être utilisées par l'auditeur comme un indice de la segmentation. Par exemple, la séquence *vt* ne peut pas être un onset ou un coda en anglais. L'auditeur suppose donc que *v* et *t* appartiennent à deux unités lexicales différentes et ainsi segmente un flux de parole contenant la séquence *vt*. Friederici et Wessels (1993) ont montré qu'un enfant de 9 mois est sensible aux contraintes phonotactiques de sa langue maternelle pour réaliser une segmentation.

Segmentation basée sur les indices articulatoires/acoustiques

Une source d'information pour segmenter le flux de parole peut être les variations allophoniques. Par exemple, l'allophone /t/ en anglais est aspiré au début du mot comme dans *tab* mais il est non aspiré à la fin du mot comme dans *cat*. Le degré de coarticulation entre les phonèmes successifs peut être également utilisé comme un indice reflétant les frontières des syllabes et des mots. Certaines études suggèrent que l'auditeur utiliserait ce genre d'indice pour segmenter le flux de parole (Gambell et Yang, 2005). La durée des syllabes serait également un indice de la segmentation. Par exemple, la syllabe /slip/ devient progressivement de plus en plus courte dans le mot *sleep*, *sleepy* et *sleepiness*, ce qui permettrait à l'auditeur d'effectuer une segmentation cohérente (Lehiste, 1972).

Segmentation basée sur le lexique

La segmentation basée sur le lexique consiste à segmenter le flux de la parole en identifiant les unités lexicales présentes dans le flux. Nous pouvons distinguer deux mécanismes différents dans ce type de segmentation : la segmentation séquentielle et la segmentation basée sur la compétition.

La segmentation séquentielle consiste à identifier un mot lorsque l'entrée correspond uniquement à une seule unité lexicale dans le lexique. Les candidats perdent progressivement leur activité lorsque l'entrée du modèle devient incohérente. Par exemple, si l'entrée est le mot *catalog*, les unités *cat* et *catalog* seront actives jusqu'à ce que le deuxième /a/ du *catalog* soit présenté au modèle. À partir de ce point, *cat* ne reste plus actif. Le modèle Cohort (Marslen-Wilson et Welsh, 1978) utilise la stratégie de la segmentation séquentielle pour la reconnaissance des mots. Les

limites de la segmentation séquentielle ont été démontrées par certaines données expérimentales (pour une revue, voir Davis *et al.*, 2002).

La segmentation par compétition est basée sur le fait que plusieurs candidats lexicaux, plus ou moins cohérents avec l'entrée, peuvent être actifs parallèlement et que l'activité d'un candidat diminue l'activité des autres. La compétition serait une stratégie nécessaire pour la segmentation du flux de parole (McQueen *et al.*, 1995). Dahan et Magnuson (2006) ont distingué trois types de compétition :

- Compétition avec règle de décision : ce type de compétition est une compétition indirecte entre les candidats. Par exemple dans le modèle Cohort révisé (Marslen-Wilson, 1993), afin de sélectionner une unité parmi les candidats, l'activité de chaque unité est calculée par rapport à l'activité de toutes les autres unités. Ainsi, la reconnaissance d'un mot est influencée par l'activité des autres unités cohérentes avec l'entrée, sans qu'il y ait une compétition directe entre les unités.

- Compétition directe : dans le modèle TRACE et Shortlist, les unités lexicales s'inhibent l'une et l'autre en fonction de leur activité qui est, elle-même, en fonction de leur degré de cohérence avec l'entrée. Les prédictions de la compétition directe et la compétition avec la règle de décision sont très proches, ce qui rend difficile le choix entre ces deux mécanismes (pour une revue, voir Dahan et Magnuson, 2006). Une différence entre ces deux types de compétition concerne la dynamique temporelle de la reconnaissance des mots. Dans la compétition directe, au fur et à mesure de la présentation de l'entrée au modèle, le modèle fournit des prédictions à la fois en fonction de la cohérence de chaque candidat avec l'entrée et du degré d'inhibition des autres candidats. Au contraire, dans le mécanisme de compétition avec règle de décision, la dynamique temporelle de la reconnaissance reflète seulement la cohérence des candidats et non pas leurs influences réciproques. Pour pouvoir conclure entre ces deux mécanismes, il faut avoir des données plus claires sur la dynamique réelle de l'activation lexicale chez l'humain.

- Compétition émergente : Gaskell et Marslen-Wilson (1997) ont proposé un modèle de la perception de la parole dans laquelle la connaissance lexicale est modélisée de façon distribuée. Dans leur modèle, les différentes formes de la connaissance lexicale (ex. sémantique, phonologique, etc.) sont représentées en parallèle et sont simultanément accessibles. Puisque les représentations dans le modèle sont distribuées, la sortie, à chaque instant, ne représente pas les unités lexicales cohérentes avec l'entrée mais elle représente un mélange des patterns cohérents avec l'entrée. Ainsi, la compétition entre les candidats est plus complexe et elle émerge en fonction de plusieurs éléments tel que le corpus utilisé dans la phase d'apprentissage du modèle.

Combinaisons de plusieurs stratégies de segmentation

Parmi les stratégies décrites ci-dessus, les stratégies pré-lexicales sont plus acceptées du point de vue développemental car elles peuvent expliquer comment un

enfant peut segmenter le flux de la parole sans connaître les mots qui composent le flux de parole. Du point de vue computationnel, les stratégies pré-lexicales sont moins efficaces que la segmentation lexicale. Cependant, les différentes stratégies de segmentation n'étant pas mutuellement exclusives, une combinaison de deux ou plusieurs d'entre elles peut être utilisée pour segmenter le flux de la parole.

Les modèles computationnels combinant plusieurs stratégies de segmentation sont apparus plus efficaces dans la prédiction des frontières des mots. Par exemple, Christiansen *et al.* (1998) ont proposé un modèle qui utilise la segmentation basée sur la régularité statistique et la segmentation métrique basée sur les syllabes accentuées pour segmenter le flux de la parole. Ils ont montré que la combinaison de ces deux stratégies conduit à un meilleur résultat par rapport à l'utilisation d'une seule stratégie. Norris *et al.* (1995) ont également utilisé la combinaison de la segmentation basée sur la compétition et la segmentation métrique basée sur l'accentuation au sein du modèle Shortlist.

7.2.3 Exemples de modèles

Modèle TRACE

Le modèle TRACE est un modèle de type Activation Interactive et Compétition (AIC) proposé par McClelland et Elman (1986). Les modèles AIC de la perception de la parole ont souvent plusieurs niveaux hiérarchiques de représentation. Chaque niveau est composé des unités représentant la nature de l'entrée (par exemple, les unités lexicales, phonémiques, etc.). Au cours de la présentation de l'entrée, les unités cohérentes avec l'entrée deviennent actives. Les unités à l'intérieur d'un niveau sont connectées l'une à l'autre par des connexions inhibitrices. Les connexions entre les unités des différents niveaux sont excitatrices et souvent bi-directionnelles.

Le modèle TRACE contient trois niveaux de représentations : niveau pseudo-spectral (niveau de trait phonétique), niveau phonémique et niveau lexical. L'entrée du modèle est une suite de lettres représentant des phonèmes qui seront traduits, à l'intérieur du modèle, en une matrice de traits phonétiques. Une unité lexicale représente chaque mot. Chaque unité est répétée au cours du temps afin de représenter à la fois sa valeur linguistique et son début (voir figure 7.2).

La figure 7.2 illustre quelques connexions entre différentes unités et différents niveaux dans TRACE. L'entrée, le mot *abrupt*, est présentée de gauche à droite au niveau de trait, ce qui conduit à l'activation des traits, plus ou moins cohérent avec l'entrée. L'activation des traits entraîne l'activation des phonèmes. Les phonèmes activés excitent, à leur tour, les mots qui contiennent ces phonèmes et inhibent les autres phonèmes qui recouvrent la même proportion de l'entrée. Ils envoient également un feedback excitateur vers les traits qui leur correspondent. De même façon, les mots activés inhibent les autres mots qui correspondent à la même proportion de l'entrée et ils envoient un feedback excitateur vers les phonèmes qui font partie de ces mots. Ainsi, TRACE prend en compte la continuité du signal et réalise la segmentation de la séquence d'entrée.

Dans TRACE, le niveau de trait et le niveau de phonème et celui de phonème

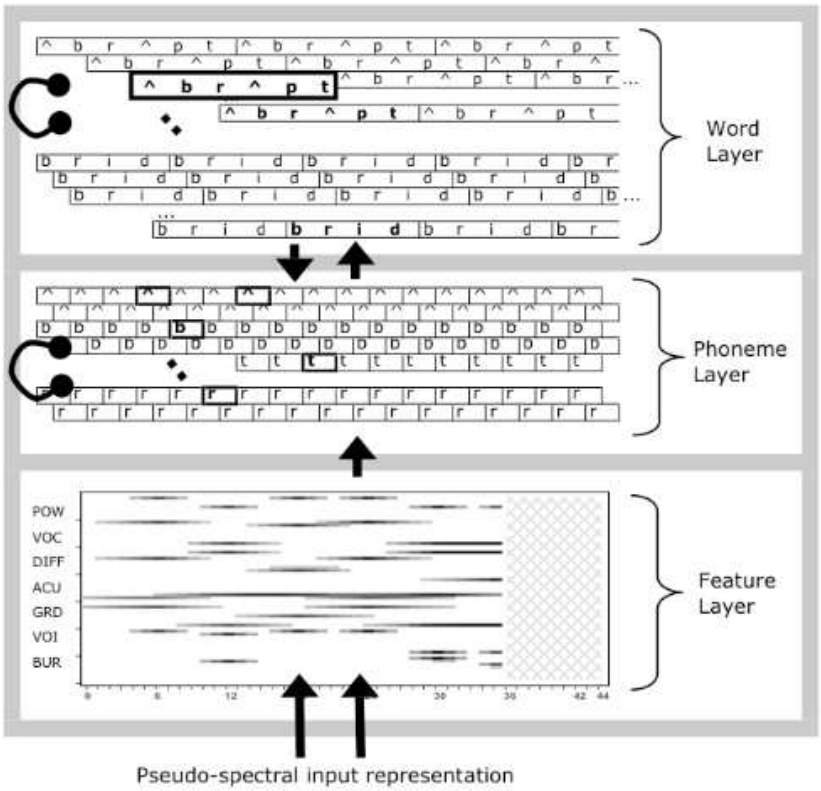


FIGURE 7.2 : Architecture du modèle TRACE. Activation des différentes unités et leur connexions lors de la présention du mot *abrupt*.

et de mot se chevauchent temporellement (voir figure 7.2). Grâce à cette caractéristique, ce modèle peut rendre compte de certaines données expérimentales telles que la perception des phonèmes ambigus ou coarticulés.

Le modèle TRACE est considéré comme un des modèles les plus influents de ces dernières années dans le domaine de la perception de la parole (Protopapas, 1999, p. 417 ; Gaskell, 2007, p. 65). Plus de 20 ans après sa naissance, TRACE continue à se développer (voir par exemple, Mirman *et al.* (2006) : implémentation de l'apprentissage hebbien dans TRACE, Strauss *et al.* (2007) : interface *user-friendly* de TRACE en langage java). Le modèle TRACE a réussi à rendre compte de plusieurs phénomènes perceptifs observés dans les expériences psycholinguistiques. Par exemple, TRACE peut prédire l'effet Ganong, effet du contexte lexical sur l'identification des phonèmes. Ganong (1980) a montré que l'identification du son initial d'une syllabe construit de façon à être ambigu entre /k/ et /g/ est influencée par la connaissance lexicale : le son est identifié comme /k/ si le reste du mot est -iss (*kiss* étant un mot en anglais est favorisé contre *giss*, un non-mot) et il est identifié comme /g/ si le reste du mot est -ift (*gift*, un mot, est favorisé contre *kift*, un non-mot). TRACE peut reproduire ce phénomène grâce aux connexions de type feedback à partir du niveau lexical vers les phonèmes. Cependant, l'existence ou non de l'influence top-down du niveau lexical vers les représentations phonémiques suscite un débat dans

la littérature. En effet, plusieurs phénomènes expliqués par le feedback dans TRACE peuvent également être expliqués par des mécanismes purement bottom-up (pour une revue sur le débat et les références, voir Christiansen et Chater, 2001, p. 21–37).

Une des critiques majeures envers le modèle TRACE concerne la façon dont ce modèle représente le temps. Comme décrit ci-dessus, afin de représenter le temps du passage d'un flux de parole, TRACE répète les unités les unes après les autres (voir figure 7.2) ce qui rend ce modèle peu plausible du point de vue physiologique. Par exemple, cette structure n'est pas adaptée pour l'apprentissage des nouvelles représentations lexicales car, pour chaque nouvelle acquisition, il faut répéter au cours du temps l'unité la représentant et établir de nouvelles connexions inhibitrices entre cette nouvelle unité et les autres unités dans le même niveau avec qui elle partage une proportion de l'entrée.

Modèle Shortlist

Dans le modèle Shortlist proposé par Norris (1994), la reconnaissance de l'entrée s'effectue en deux phases. La première phase consiste en une recherche lexicale pour trouver les candidats cohérents avec l'entrée. Pour cela, il utilise un réseau récurrent SRN proposé par Elman (1990). Ce réseau est présenté sur la figure 7.3. La deuxième phase consiste à sélectionner le meilleur entre ces candidats. Un réseau d'Activation Interactive et Compétition, décrit ci-dessus, est utilisé dans cette deuxième phase.

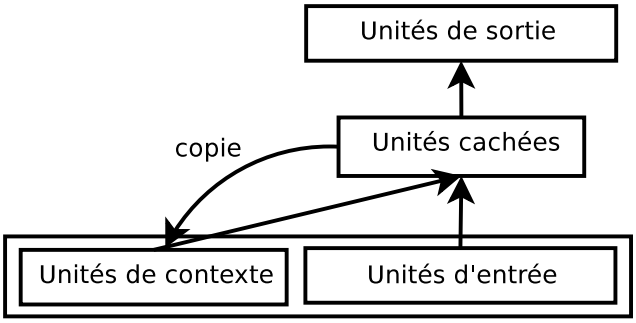


FIGURE 7.3 : Réseau récurrent SRN (Elman, 1990) utilisé dans le modèle Shortlist. Les unités d'entrée et de sortie représentent respectivement l'entrée et la sortie à l'instant t et les unités de contexte représentent l'état du réseau à l'instant $t-1$.

L'entrée du modèle Shortlist est sous forme d'une séquence de traits phonétiques[6]. En fonction de son entrée, le réseau attribue une seule unité lexicale à la sortie, correspondant au mot identifié. La représentation du temps dans ce modèle est totalement différente du modèle TRACE. En effet, comme illustré sur la figure 7.3, l'activation des unités cachées du réseau récurrent est copiée vers les unités appelées unités de contexte. Les activités de ces unités seront ensuite renvoyées vers les unités cachées lors du pas de temps suivant. Ainsi, les unités de contexte représentent l'état du système à l'instant $t-1$. Grâce à ce mécanisme, le réseau garde la

[6]Dans la version implémentée du modèle, les lettres représentant des phonèmes ont été utilisées comme entrée. La performance du modèle est indépendante de la forme d'entrée (Norris, 2005).

trace des entrées précédentes, ce qui permet au modèle d'intégrer les informations de l'instant t et celles de l'instant $t-1$. C'est ainsi que le modèle Shortlist détecte un mot sans garder explicitement l'instant du début de mot.

Le deuxième réseau dans le modèle Shortlist est responsable de la sélection d'une unité lexicale parmi les unités identifiées par le réseau récurrent. Pour cela, un réseau de type Activation Interactive et Compétition est utilisé. Dans ce réseau, les candidats s'inhibent l'un l'autre en fonction du nombre de phonèmes et de la portion de l'entrée qu'ils partagent (voir figure 7.4 pour un exemple).

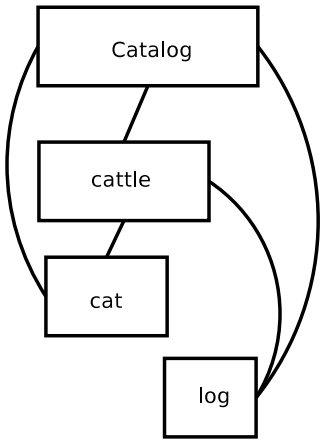


FIGURE 7.4 : Les connexions inhibitrices dans le modèle Shortlist entre quelques candidats identifiés par le réseau SRN pendant la présentation du mot *catalog*.

Le Modèle Shortlist partage certaines caractéristiques avec le modèle TRACE. Leur différence majeure concerne l'absence de feedback à partir des représentations lexicales vers les représentations pré-lexicales dans le modèle Shortlist. En effet, Shortlist n'utilise que les informations bottom-up, du niveau phonème vers le niveau lexical. Il est également à noter que dans le modèle TRACE, l'activation des unités lexicales est uniquement basée sur leur degré de cohérence avec l'entrée tandis que dans le modèle Shrortlist, la cohérence des unités lexicales avec l'entrée augmente l'activité de ces unités (comme dans TRACE) mais leurs incohérences (*mismatch*) influencent également l'activité des unités lexicales en la diminuant. Une autre différence importante entre Shortlist et TRACE concerne la représentation temporelle de ces deux modèles. Comme nous l'avons vu ci-dessus, Shortlist, contrairement à TRACE, ne duplique pas les unités lexicales pour chaque pas de temps. Enfin, dans le modèle Shorlist, toutes les unités lexicales n'entrent pas en compétition et, contrairement à TRACE, seules les unités lexicales recevant assez d'excitation du niveau pré-lexical entrent en compétition. Ainsi, du point de vue computationnel, le modèle Shortlist est moins lourd que le modèle TRACE mais du point de vue de la plausibilité physiologique, Shortlist n'a pas prouvé avoir significativement progressé par rapport à TRACE (Protopapas, 1999). Récemment, Norris et McQueen (2008) ont proposé un modèle Shortlist de type Bayésien où les principes Bayésiens remplacent le réseau d'Activation Interactive et Compétition de Shortlist traditionnel.

Modèle ART

Le modèle ART proposé par Carpenter et Grossberg (1987) est à l'origine un modèle de reconnaissance adaptative des patterns avec apprentissage non-supervisé en temps réel qui peut être appliqué à tous les domaines du traitement adaptatif de l'information (Carpenter et Grossberg, 1988). Le modèle ART essaie de répondre à un dilemme, auquel font face tous les systèmes intelligents, appelé dilemme de stabilité-plasticité : un système qui apprend dans un environnement doit s'adapter lorsqu'il rencontre des événements significatifs et en même temps, il doit rester stable quand il rencontre des événements non significatifs n'ayant pas de rapport avec le système. Notamment, en acquérant de nouvelles connaissances, un système intelligent ne doit pas oublier les connaissances déjà acquises. Le modèle ART a été proposé comme une solution pour ce type de problèmes.

La figure 7.5 illustre l'architecture du modèle ART telle que définie dans Carpenter et Grossberg (1987). Ce système est composé de deux sous-systèmes : le sous-système attentionnel et le sous-système orientationnel. Le sous-système attentionnel contient le niveau F1, responsable de la comparaison de l'entrée et de la réponse du système, le niveau F2, responsable de la reconnaissance de l'entrée et les paramètres de contrôle G1 et G2 qui modulent respectivement l'activation des unités du niveau F1 et F2. Quand au sous-système orientationnel, il redémarre le cycle de la reconnaissance si la réponse du système ne correspond pas à l'entrée.

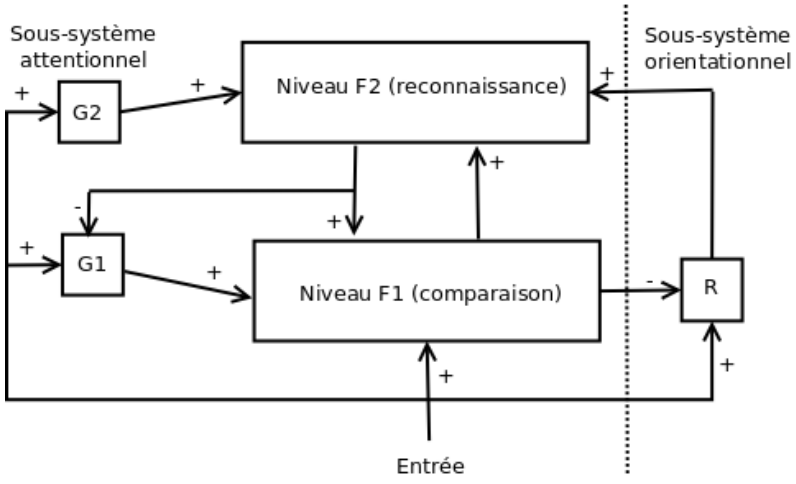


FIGURE 7.5 : Architecture du modèle ART.

Dans le modèle ART, l'entrée est présentée au niveau F1 qui sera ensuite transmise au niveau F2. Les informations bottom-up correspondant à l'entrée sont multipliées par les poids des connexions connectant le niveau F1 au niveau F2. Le niveau F2 catégorise l'entrée et le résultat de cette catégorisation sera ensuite renvoyé vers le niveau F1 afin de comparer la catégorie trouvée et l'entrée du modèle. Ces informations top-down sont multipliées par les poids de connexions connectant le niveau F2 au niveau F1. Le degré de ressemblance entre l'entrée et la catégorie trouvée est comparé avec un paramètre du système appelé paramètre de vigilance. Si le degré

de ressemblance est supérieur à la valeur de vigilance, la catégorie trouvée est acceptée et l'entrée suivante est présentée au système sinon le système orientationnel inhibe la catégorie trouvée dans le niveau F2 afin de permettre à une autre catégorie d'émerger. Dans le modèle ART, les niveaux F1 et F2 représentent la mémoire à courte terme et la matrice des poids de connexions entre les niveaux F1 et F2 représente la mémoire à long terme du système.

En ce qui concerne la perception de la parole, deux versions de modèle ART ont été proposées par Grossberg et collègues : ARTPHONE (Grossberg *et al.*, 1997) qui traite la perception des phonèmes et ARTWORD (Grossberg et Myers, 2000) qui modélise la perception des mots. Le modèle ARTWORD sera brièvement présenté dans la suite.

Dans ce modèle, les processus bottom-up de traitement du signal acoustique de la parole, via une transformation acoustique-phonétique, entraînent l'activation des items phonétiques dans la mémoire de travail (voir figure 7.6, A). L'ordre temporel des items est représenté par le gradient des activités : l'item le plus actif est le plus récent. Les patterns actifs dans la mémoire de travail activent les représentations unifiées (*unitized representations* ou *list chunks*) correspondant aux items actifs et à leur ordre. Ces représentations unifiées contiennent les unités linguistiques telles que les phonèmes, les syllabes et les mots. Les représentations unifiées actives entrent en compétition les une avec les autres (voir figure 7.6, B). Lorsque l'activité d'une représentation unifiée dépasse un seuil, elle envoie un feedback excitateur vers ses items phonétiques dans la mémoire de travail. La boucle excitatrice entre les items et les listes des représentations unifiées crée un processus appelé « résonance » qui excite à la fois les représentations unifiées et leurs items (voir figure 7.6, C). Selon le modèle ARTWORD, lorsque l'auditeur écoute un signal de parole, la résonance à travers la mémoire de travail fait le liage entre les items phonémiques pour créer des unités de langage plus grandes (ex. syllabe, mot) et les faire émerger dans la perception consciente de l'auditeur. Enfin, pour qu'une nouvelle boucle excitatrice puisse s'initialiser, le réseau est remis à nouveau dans un état non-résonant (voir figure 7.6, D).

Le modèle ARTWORD a quelques caractéristiques en commun avec les modèle TRACE et Shortlist. Ainsi, ARTWORD, comme les deux autres modèles, prévoit une inhibition latérale entre les candidats. Le flux des informations dans ARTWORD, comme dans le modèle TRACE et contrairement à Shortlist, est à la fois bottom-up et top-down. Il est à noter que la nature du mécanisme top-down est différente dans ARTWORD de celui de TRACE car dans le modèle ARTWORD, la résonance qui est nécessaire pour l'émergence des percepts, ne peut pas se produire sans le feedback, tandis que dans TRACE le feedback n'est pas nécessaire pour activer les unités. Parmi ces trois modèles, le modèle ARTWORD est le seul qui est capable d'ajouter de nouvelles représentations en ligne pendant la présentation de l'entrée, ce qui rend ce modèle plus plausible que les modèles *hardwired* comme TRACE et Shortlist. Cette capacité de ARTWORD rencontre ses limites quand la nouvelle représentation contient les représentations déjà familières au modèle (ex. syllabes et mots plus courts composant un mot). En effet, dans ce cas, avant que le modèle ne puisse identifier l'entrée comme une nouvelle représentation,

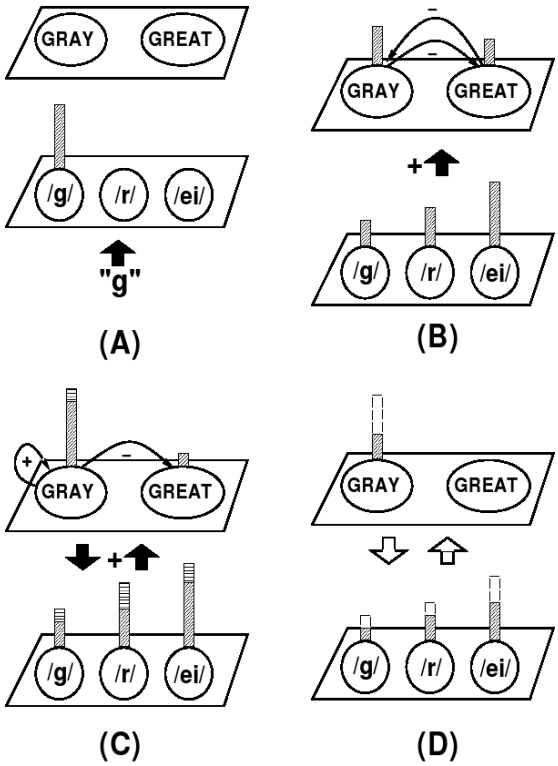


FIGURE 7.6 : Cycle de la perception dans le modèle ARTWORD. (A) Activation bottom-up (B) Compétition entre les représentations unifiées (C) Résonance entre les items phonétiques et les représentations unifiées (D) Remise du réseau dans un état non-résonant. Figure tirée de Grossberg et Myers (2000).

les composantes familières au modèle seront traitées et reconnues. Pour résoudre ce problème, ARTWORD est conçu préalablement de sorte que les représentations les plus longues soient toujours favorisées par rapport aux représentations courtes, ce qui rend une partie du modèle *hardwired*.

7.3 Modèles de l'effet de transformation verbale

Dans cette partie, nous présenterons d'abord une théorie expliquant l'effet de transformation verbale intitulée *Node Structure Theory*. Cette théorie explique l'effet de transformation verbale dans le cadre de la perception/production de la parole et propose, d'une manière qualitative, des mécanismes impliqués dans l'émergence des transformations. Nous présenterons ensuite un modèle de l'effet de transformation verbale de type système dynamique qui rend compte de la dynamique temporelle des transformations. Ce modèle ne traite pas l'effet de transformation verbale en tant que phénomène langagier.

7.3.1 *Node Structure Theory*

MacKay *et al.* (1993) ont proposé une explication de l'effet de transformation verbale dans le cadre de la *Node Structure Theory* (NST) (MacKay, 1982, 1988). À notre connaissance, cette théorie est la seule qui propose des mécanismes communs pour expliquer la perception et production « normale » (non multistable) de la parole et l'effet de transformation verbale. Selon NST, trois systèmes sont impliqués dans la perception et la production de la parole :

- Système phrastique (*sentential system*) qui contient des unités hiérarchisées représentant les concepts sous-jacents des mots et des phrases.
- Système phonologique (*phonological system*) qui contient des unités hiérarchisées représentant les syllabes, clusters consonantiques, voyelles et les traits phonétiques.
- Système de mouvement musculaire (*muscle mouvement system*) qui est impliqué dans la production ouverte de la parole. Ce système contient des sous-systèmes laryngé (*laryngeal*), phonatoire (*airflow*) et articulatoire (*articulatory*).

Les connexions entre le système phrastique et le système phonologique sont bidirectionnelles car les mêmes unités sont utilisées pendant la perception et la production. La figure 7.7 illustre l'activation des unités lors de la répétition du mot *frisbee*.

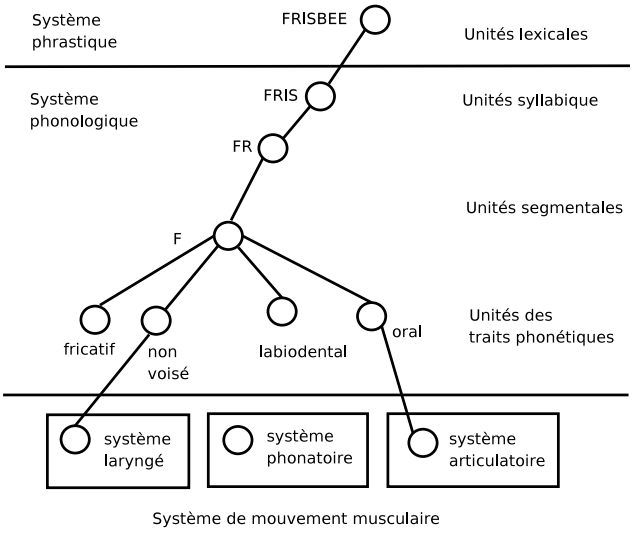


FIGURE 7.7 : *Node Structure Theory* : quelques unités représentant le mot *frisbee*.

NST peut générer les transformations verbales grâce aux trois mécanismes suivants :

- Amorçage des unités (*node priming*) : dans une expérience de transformation verbale, le signal acoustique entraîne l'amorçage des unités. L'amorçage se propage automatiquement à travers des connexions entre les unités, d'une

manière bottom-up, à partir des unités d'analyse du signal vers des unités phonologiques et lexicales. Dans chaque niveau, une valeur d'amorçage est attribuée aux unités. La valeur d'amorçage diminue en fonction du nombre de connexions connectant une unité aux autres.

- Activation des unités (*node activation*) : l'unité lexicale dont la valeur d'amorçage est la plus élevée devient active (*most-primed-wins principle*). L'activation d'une unité se termine par une courte période d'auto-inhibition (environ 150 ms) qui baisse l'amorçage de l'unité sous le seuil normal d'activation.

- Saturation des unités (*node satiation*) : ce mécanisme est la cause de l'émergence de nouvelles représentations verbales. L'activation répétée d'une unité réduit, de façon temporaire, la capacité de l'unité d'accumuler les valeurs d'amorçage. Ainsi, d'autres unités, précédemment moins amorcées, peuvent devenir actives, ce qui entraîne la perception d'une nouvelle représentation verbale. La figure 7.8 illustre la réaction de deux unités, l'une saturée et l'autre non-saturée, au même signal d'amorçage. L'amorçage commence à t_0 et se termine à t_1. L'unité la plus amorcée, ici l'unité non-saturée, devient active pour un intervalle fixe de temps (jusqu'à t_2). La phase d'activation est suivie par une phase de récupération.

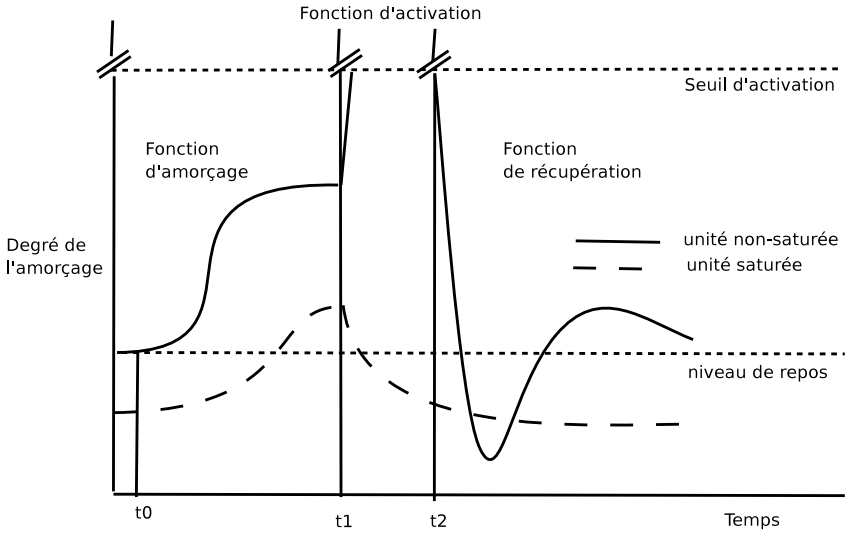


FIGURE 7.8 : La phase d'amorçage, d'activation et de récupération pour deux unités, l'une saturée (courbe en pointillé) et l'autre non-saturée (courbe solide).

La figure 7.9 illustre comment l'effet de transformation verbale se produit lors de la répétition du mot *base*. La présentation du mot *base* entraîne l'activation de l'unité lexicale *base* car, étant en parfaite cohérence avec l'entrée, elle reçoit plus d'amorçage que les autres unités dans le même niveau. Les unités lexicales qui représentent les formes similaires au mot *base*, par exemple les mots *face*, *pace*, *say*, sont également amorcées mais leur degré d'amorçage est inférieur à celui de *base*.

Après un certain nombre de répétitions, l'unité représentant *base* devient saturée et accumule moins d'amorçage, ce qui permet à l'unité dont le degré d'amorçage est supérieur aux autres de s'activer. Ainsi, le percept *base* bascule vers un autre percept, par exemple, *face*. À son tour, *face* devient saturé et une autre unité s'active et ainsi de suite.

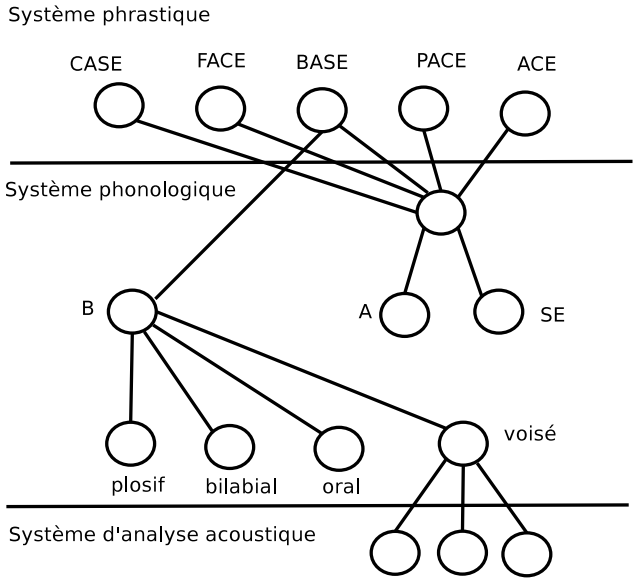


FIGURE 7.9 : L'amorçage et l'activation des unités selon NST lors de la présentation du mot *base* à l'entrée.

NST a réussit à rendre compte de certaines propriétés des transformations verbales telles que l'effet du voisinage phonologique, l'effet de la fréquence lexicale et la différence entre les mots et les pseudo-mots (MacKay *et al.*, 1993; Shoaf et Pitt, 2002). Cependant, cette théorie ne traite pas la dynamique des transformations et elle n'a pas été implémentée sous forme d'un modèle computationnel.

7.3.2 Modèle Synergique pour l'effet de transformation verbale

S'inspirant d'un modèle synergique (dynamique non linéaire) de la perception multistable en vision (Ditzinger et Haken, 1995), Ditzinger *et al.* (1997a) ont suggéré un modèle de l'effet de transformation verbale. Un modèle synergique (ou auto-organisé) contient une mémoire associative qui permet de compléter et identifier un pattern d'entrée comme un pattern déjà mémorisé dans sa mémoire (Haken, 1991). Les patterns d'entrée peuvent être des stimuli visuels, auditifs, etc. Dans la phase de reconnaissance des patterns d'entrée, la ressemblance entre l'entrée et les patterns mémorisés est calculée. Cette ressemblance est appelé le paramètre d'ordre du système ξ_k, k étant l'index du pattern mémorisé.

L'évolution temporelle de ξ_k peut être représentée par le mouvement d'une particule dans le diagramme (*landscape*) d'énergie du système. Chaque minimum local du diagramme est attribué à un pattern mémorisé k. La figure 7.10 illustre un exemple

d'un diagramme d'énergie correspondant à la perception du cube de Necker. Les deux percepts possibles sont les minimums du diagramme (dans ce cas, $k = 2$). Le

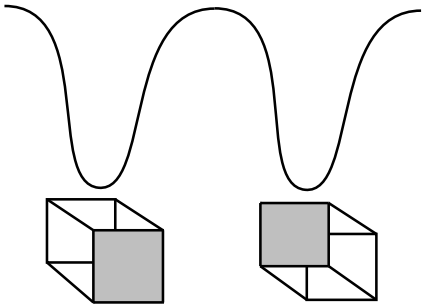


FIGURE 7.10 : Diagramme d'énergie du système correspondant à la perception du cube de Necker.

mouvement d'une particule dans le diagramme d'énergie du système (V) se fait de sorte que la particule soit attirée vers le minimum le plus proche du pattern initial :

$$\dot{\xi}_k = -\frac{\partial V}{\partial \xi_k}, \qquad k = 1, \cdots, M \tag{7.4}$$

La forme explicite de V (diagramme de l'énergie) utilisée dans ce modèle est :

$$V = -\frac{1}{2}\sum_{k'=1}^{M} \lambda_{k'}\xi_{K'}^2 + \frac{B}{4}\sum_{k\neq k'}^{M} \xi_k^2\xi_{k'}^2 \left[1 - 4\alpha_{kk'}\frac{(\xi_k^2 - \xi_{k'}^2)}{(\xi_k^2 + \xi_{k'}^2)}\right] + \frac{C}{4}(\sum_{k'=1}^{M} \xi_{k'}^2)^2 \tag{7.5}$$

B et C sont les paramètres indiquant la vitesse de la discrimination entre les patterns. Les paramètre $\alpha_{kk'}$ et λ_k représentent respectivement le biais entre le pattern k et k' et le degré d'attention au pattern k.

Dans ce modèle, les changements perceptifs ont lieu grâce aux deux mécanismes ci-dessous :

- Saturation de l'attention : dans ce modèle, le système ne se comporte pas comme une particule dans un diagramme d'énergie constant mais la forme du diagramme change en fonction de la position de la particule et de son mouvement. En effet, les paramètres λ_k dans l'équation 7.5, appelés les paramètres de l'attention, déterminent la forme du diagramme. Quand une particule entre dans un minimum local du diagramme de l'énergie, la forme du diagramme change au fur à mesure et le niveau de la vallée augmente et elle devient, éventuellement, un maximum local. Ainsi, les fluctuations dans le système peuvent déplacer la particule vers un autre minimum local, ce qui conduit à une transformation verbale. L'équation ci-dessous représente l'évolution temporelle des paramètres d'attention :

$$\dot{\lambda}_k = \gamma(1 - \lambda_k - \xi_k^2) + F_k(t), \qquad k = 1, \cdots, M \tag{7.6}$$

L'augmentation de ξ_k entraîne la saturation de l'attention au pattern k. Le paramètre γ représente la vitesse de l'oscillation des différents sujets, autrement dit, la vitesse de changement de la forme du diagramme d'énergie pour chaque sujet. $F_k(t)$ représente les fluctuations du système qui provoquent le déplacement soudain de la particule vers un autre minimum local.

- Biais : ce paramètre permet d'expliquer la différence entre différentes formes de transformations et notamment le fait que certaines transformations sont signalées plus souvent que d'autres. Dans l'équation 7.5, $\alpha_{kk'}$ représente le biais, autrement dit, la différence entre le pattern k et k'. Le biais change la pente des bords séparant les minimums locaux pour favoriser la perception d'une transformation à l'autre.

En remplaçant l'équation 7.5 dans 7.4, on obtient l'équation suivante de la dynamique des paramètres d'ordre du système :

$$\dot{\xi}_k = \left[\lambda_k - B\sum_{k'=1}^{M}\xi_{k'}^2(1-\alpha_{kk'}\left(1-\frac{2\xi_{k'}^4}{(\xi_k^2+\xi_{k'}2)^2})\right)+B\xi_k^2 - C\sum_{k'=1}^{M}\xi_{k'}^2\right] \quad (7.7)$$

L'équation 7.5 est identique à celle utilisée par Ditzinger et Haken (1995) pour modéliser la multistabilité perceptive en vision. Dans le modèle de l'effet de transformation verbale, les patterns mémorisés (les minimums du système) sont les différentes formes phonémiques signalées pendant une expérience de transformation verbale. L'extension du modèle de la perception multistable en vision pour modéliser l'effet de transformation verbale concerne principalement le nombre plus important de percepts possibles dans une expérience de transformation verbale, ce qui est limité à deux ou trois en vision. Pour modéliser l'effet de transformation verbale, il faut donc considérer les équations 7.6 et 7.7 dans un cas général avec le nombre M de différents percepts. Ainsi, les bascules perceptives ont lieu entre des points fixes dans un espace de M dimensions. Chaque point fixe $\xi_{0,k}$ est attribué à une forme phonétique, autrement dit au percept k, de sorte que

$$\xi_{0,k,k}^2 = \frac{1}{1+C} \quad (7.8)$$

Les autres coordonnées sont

$$\xi_{0,k,k'}^2 = 0, \qquad k' \neq k \quad (7.9)$$

Les oscillations entre deux percepts k et k' peuvent se produire seulement si

$$-\alpha_{crit} < \alpha_{kk'} < \alpha_{crit}, \qquad k = 1, 2, \cdots, M \quad (7.10)$$

avec

$$\alpha_{crit} = \frac{1-B}{4B} \quad (7.11)$$

Si la valeur de biais $\alpha_{kk'}$ est proche de $-\alpha_{crit}$ ou α_{crit}, le percept k ou k' sera toujours dominant. Dans ce cas, sans fluctuation, le percept le plus faible ne peut jamais émerger.

La figure 7.11 illustre les paramètres d'ordre du système pour $M = 4$ (4 différentes transformations) au cours du temps. À chaque instant, le paramètre d'ordre ayant la valeur la plus élevée est identifié comme la forme perçue. Comme illustré sur la figure 7.11, après une durée de stabilité, un changement soudain se produit, ce qui correspond à une transformation verbale (pour les détails de calculs, voir Ditzinger *et al.*, 1997a).

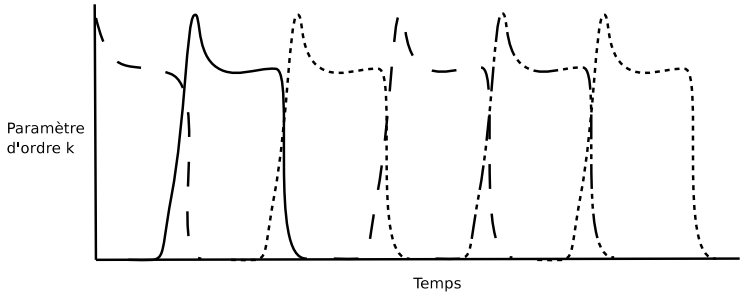


FIGURE 7.11 : Une simulation du modèle proposé par Ditzinger *et al.* (1997a) pour quatre différents patterns k correspondant à quatre transformations verbales (4 courbes différentes). En abscisse : le temps du passage du stimulus, en ordonnée : les valeurs de ξ_k (paramètre d'ordre k).

Ce modèle reproduit les résultats expérimentaux observés par Ditzinger *et al.* (1997b). Selon cette étude expérimentale, l'organisation perceptive des transformations était celle du couplage par paire et l'une des deux transformations majoritairement signalées correspondait au stimulus original. Ce modèle peut rendre compte de ces données notamment grâce au paramètre biais.

Comme nous l'avons vu ci-dessus, ce modèle traite l'effet de transformation verbale comme une occurrence du phénomène général de la multistabilité, d'où l'utilisation du même type de modèle pour la multistabilité perceptive en vision et pour l'effet de transformation verbale. Cette approche de l'effet de transformation verbale en fait un modèle dépourvu de représentations et de prise en compte de mécanismes de nature langagière.

7.4 Conclusion

Nous avons décrit dans ce chapitre le modèle NST proposé par MacKay *et al.* (1993) qui est à notre connaissance le modèle le plus complet expliquant l'effet de transformation verbale. Cependant, à ce modèle, non computationnel, manquent beaucoup de détails afin de pouvoir rendre compte de la dynamique temporelle des transformations verbales. Au contraire, le modèle computationnel proposé par Ditzinger *et al.* (1997a) peut expliquer la dynamique des transformations mais ce modèle n'est pas un modèle de la perception de la parole dans le sens où il manque des représentions et des mécanismes de nature linguistique. Ainsi, nous ne disposons au départ que d'un modèle représentationnel, sans moteur computationnel, et d'un modèle computationnel sans représentations. Notre objectif est de proposer un

modèle, même préliminaire, qui soit à la fois computationnel et représentationnel.

Il nous semble qu'un cadre intéressant, jamais étudié, pour modéliser l'effet de transformation verbale serait celui fourni par les modèle psycholinguistiques de la perception de la parole : d'un côté, ils permettent d'expliquer la dynamique temporelle de l'émergence des percepts et de l'autre, ils fournissent les éléments pour pouvoir intégrer l'effet de transformation verbale dans le cadre général du système humain de la perception de la parole. C'est ce que nous allons présenter dans le chapitre suivant.

Chapitre 8

Transformations verbales dans le modèle TRACE

Sommaire

La première section de ce chapitre donne les principes de base de l'implémentation du phénomène des transformations verbales au sein d'un modèle psycholinguistique général. La seconde section est consacrée au modèle psycholinguistique TRACE. Nous y présentons le modèle et la façon dont il est implémenté par McClelland et Elman (1986). La troisième section concerne les mécanismes que nous avons intégrés au modèle TRACE afin d'expliquer les transformations verbales. Les quatrième et cinquième sections concernent respectivement des simulations et quelques propositions neuro-anatomiques.

8.1 Principes de base de la modélisation

L'effet de transformation verbale étant produit par le système humain de traitement du langage, un modèle psycholinguistique devrait l'expliquer. Pourtant, même les modèles psycholinguistiques les plus influents de la littérature ne sont pas capables de produire les bascules perceptives. Prenons l'exemple du stimulus /patapatapata..../. Lorsque ce stimulus est présenté à un modèle psycholinguistique, comme ceux que nous avons décrits dans la sous-section 7.2.3, si la première unité reconnue est /pata/, la deuxième unité reconnue sera la deuxième séquence de /pata/ et ainsi de suite. Autrement dit, si la première unité reconnue est /pata/, le modèle traite la partie de l'entrée qui succède à la première séquence de /pata/ (/pata$^{\downarrow}$ patapatapata..../) et les segmentations suivantes seront de la forme /pata$^{\downarrow}$pata$^{\downarrow}$pata$^{\downarrow}$ ..../. Ainsi, cette segmentation de l'entrée ne conduit jamais à l'émergence de la séquence /tapa/.

Notre objectif dans la suite de cette thèse est de proposer, à partir de nos résultats expérimentaux, quelques mécanismes pour modéliser l'effet de transformation verbale dans le cadre d'un modèle psycholinguistique de la perception de la parole. Comme décrit dans la sous-section 3.3, différents types de transformations verbales sont observés dans les expériences comportementales. Dans les études réalisées dans cette thèse, nous avons majoritairement rencontré et étudié les transformations de type resegmentation (par exemple, transformation de /pata/ vers /tapa/ ou /psə/ vers /səp/). Dans la partie de modélisation, nous nous limitons à ce type de transformations.

- Adaptation : dans les modèles psycholinguistiques décrits ci-dessus, les représentations lexicales et pré-lexicales peuvent rester actives indépendamment de la durée de leur activations précédentes. Ainsi, un percept peut être actif pendant une longue période de temps. Nous proposons que dans une expérience de transformation verbale, les nouveaux percepts émergent, au moins pour une partie, à cause de l'adaptation des unités représentant le percept actuel. Par exemple, dans le cas du stimulus /patapatapata.../, le percept /pata/ étant actif au début du passage du stimulus, il s'adapte après une certaine durée d'activation et ainsi peut permettre aux autres percepts d'émerger. Cette hypothèse est cohérente avec le modèle NST présenté dans la sous-section 7.3.1 ainsi qu'avec la littérature sur la multistabilité perceptive en vision revue dans la section 3.1.

- Fenêtre de liage/décision : comme illustré dans l'exemple ci-dessus, une segmentation faisant la correspondance temporelle point par point entre l'entrée et la sortie ne peut pas rendre compte de l'effet de transformation verbale. Pour résoudre ce problème, nous proposons que le traitement s'effectue à l'intérieur d'une fenêtre qui peut glisser sur différents points temporels de l'entrée, ce qui permet d'implémenter différentes segmentations de l'entrée (voir figure 8.1).

- Décision : les modèles psycholinguistiques basés sur la compétition fournissent

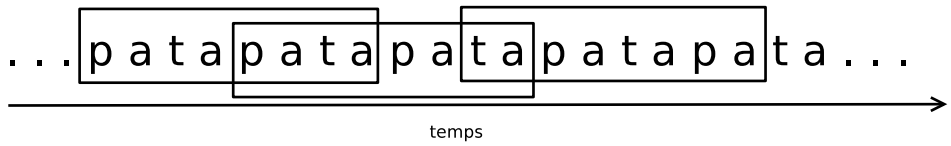


FIGURE 8.1 : Exemple de la fenêtre de liage/décision pour une répétition de la séquence /pata/.

plusieurs candidats pour la sortie du modèle. Le candidat dont l'activité est la plus élevée est considéré comme gagnant. Ce mécanisme de décision entraîne une dynamique déterministe des transformations verbales tandis que, selon les résultats expérimentaux, la dynamique de l'émergence des percepts n'est pas déterministe. Prenons l'exemple du stimulus /patapatapata.../ : en ajoutant la fenêtre de liage et le mécanisme d'adaptation, après un certain temps t_1, le percept initial /pata/ bascule vers le percept /tapa/. Le percept /tapa/ se transforme, à son tour, vers /pata/ après la durée t_2. Les valeurs t_1 et t_2 n'étant pas constantes au cours du stimulus, nous proposons un mécanisme de prise de décision probabiliste pour expliquer la dynamique non déterministe des transformations.

Notre objectif dans la partie de modélisation de cette thèse est d'implémenter les mécanismes ci-dessus dans un modèle psychophysique de la perception de la parole de sorte que le modèle soit capable de produire nos résultats expérimentaux présentés dans le chapitre 4. Ces mécanismes sont définis indépendamment du modèle psycholinguistique que nous allons utiliser. Nous pouvons les implémenter aussi bien dans TRACE que dans Shortlist et ART. Pour la suite de travail, nous avons choisi le modèle TRACE. Outre le fait que TRACE est un modèle influent dans la littérature et qu'il continue à se développer (par exemple, Mirman *et al.*, 2006; Strauss *et al.*, 2007), ce choix est basé sur des raisons pratiques : le code source de TRACE est libre et il est relativement simple à utiliser. Nous présentons dans la section 8.3 les techniques utilisées pour ajouter ces mécanismes dans le modèle TRACE de base.

8.2 Modèle TRACE

Comme décrit dans la sous-section 7.2.3, TRACE est un modèle connexionniste et localisé dans lequel les unités indépendantes représentent les traits phonétiques, les phonèmes et les mots dans leurs niveaux respectifs. Les connexions entre le niveau lexical et le niveau phonémique et entre le niveau phonémique et le niveau de traits phonétiques sont excitatrices et bi-directionnelles et celles à l'intérieur de chaque niveaux sont inhibitrices. Dans la suite, nous décrivons la manière dont le modèle TRACE fonctionne et notamment réalise une segmentation du flux de parole.

8.2.1 Architecture

Entrée et niveau de traits phonétiques

L'entrée de TRACE prend la forme d'une séquence de phonèmes que l'utilisateur présente au modèle pour effectuer une simulation. Un phonème d'entrée est fourni au modèle à chaque cycle de traitement. Les entrées peuvent comprendre /p/, /b/, /t/, /d/, /k/, /g/, /s/, /ʃ/, /r/, /l/, /a/, /i/, /u/, /ʌ/ et le silence (/-/). Il est également possible de présenter des segments spéciaux au modèle par exemple, un segment dont les caractéristiques soient entre celles de /p/ et /b/. Grâce à une fonction de type texte-à-« parole », TRACE traduit des phonèmes d'entrée en des représentations pseudo-spectrales appelées des traits phonétiques. Ainsi, l'entrée du modèle pour chaque cycle est traduite en une matrice 7×9. Cette matrice comprend sept traits phonétiques : « consonantique », « vocalique », « diffus », « aigu », « voisé », « puissance » et « amplitude du burst ». Chaque trait consiste en neuf continuums : par exemple, le trait vocalique varie de « complètement non-voisé » jusqu'au « tout à fait voisé ». Ces traits ont pour but de représenter la structure acoustique du signal de parole. Les cinq premiers traits ont été choisis en lien avec les travaux classiques en phonologie (Jakobson *et al.*, 1969). Le sixième trait, la puissance, augmente la différence entre les consonnes et les voyelles. Enfin, l'amplitude du burst a été choisie afin de renforcer la distinction entre les consonnes plosives qui ne sont pas très différentes les unes des autres sans ce trait.

Bien qu'une seule matrice d'entrée soit fournie au modèle par cycle de traitement, son effet persiste pendant les cycles suivants. Comme illustré sur la figure 8.2, les traits correspondant à chaque phonème d'entrée durent 11 tranches de temps. La valeur de chaque trait augmente jusqu'à la sixième tranche et ensuite baisse d'une manière linéaire.

Le tableau 8.1 présente la corrélation entre les traits phonétiques correspondant aux différents phonèmes d'entrées. C'est cette corrélation qui détermine le comportement du modèle en indiquant à quel point un phonème à l'entrée excite le détecteur d'un autre dans le niveau phonémique.

Niveau phonémique et lexical

Le niveau phonémique contient une série de détecteurs de phonèmes. Ces détecteurs ont une extension temporelle de 6 tranches de temps et ils sont situés sur différentes positions des traits (voir figure 8.2). Les centres des détecteurs sont espacés de 3 tranches de temps. Cette architecture permet de représenter l'alignement temporel des activations par rapport au passage du stimulus d'entrée que nous détaillerons plus loin.

L'organisation du niveau lexical dans TRACE est très similaire à celle du niveau phonémique. Les unités lexicales dans TRACE contiennent deux informations : le degré d'activité des mots qu'elles représentent et leur alignement temporel par rapport au début de la séquence d'entrée. Afin de représenter l'alignement temporel des unités activées par rapport à l'entrée, TRACE duplique toutes les unités spatialement. La figure 8.3 illustre un exemple concernant cet alignement : l'activité

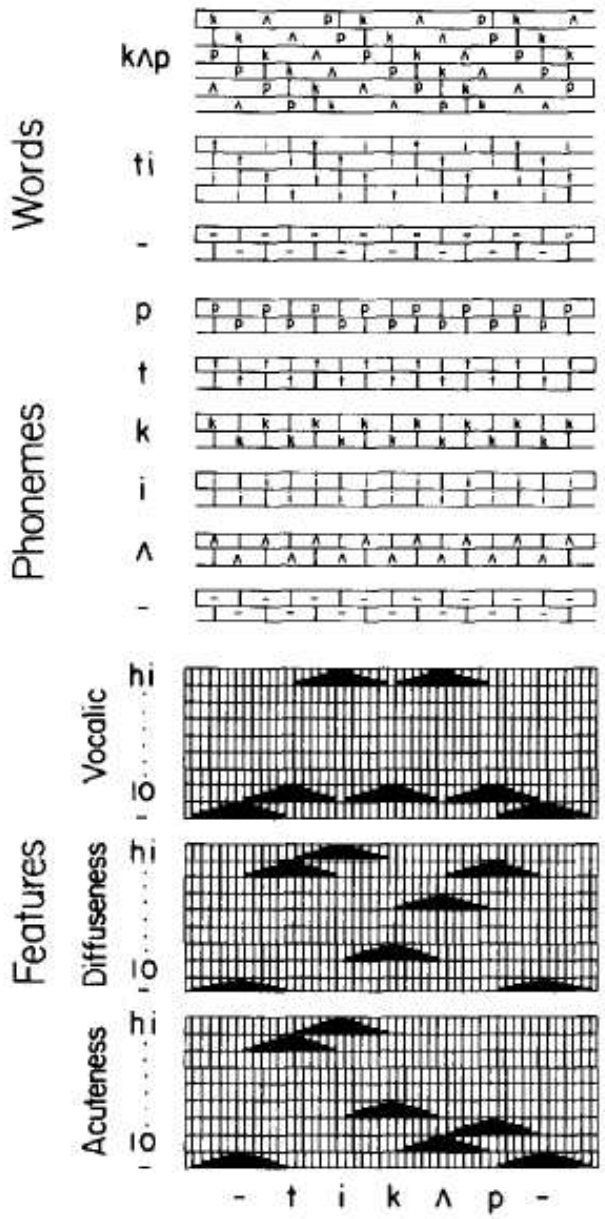


FIGURE 8.2 : Modèle TRACE de la perception de la parole. L'entrée du modèle est la séquence « tea cup ». Chaque rectangle représente une unité différente. Les côtés horizontaux des rectangles représentent l'extension temporelle de chaque unité dans TRACE. Figure tirée de McClelland et Elman (1986).

TABLE 8.1 : Corrélation entre les traits phonétiques correspondant aux différents phonèmes dans TRACE. Les corrélations inférieures à 0.2 ne sont pas présentées.

	p	b	t	d	k	g	s	ʃ	r	l	a	i	u	ʌ
p	–	.76	.71	.56	.60	.46	.30							
b	.76	–	.56	.71	.46	.60								
t	.71	.56	–	.76	.56	.42	.35							
d	.56	.71	.76	–	.42	.56								
k	.60	.46	.56	.42	–	.77		.24						
g	.46	.60	.42	.56	.77	–								
s	.30		.35				–	.65						
ʃ					.24		.65	–					.20	
r									–	.80	.29		.32	.37
l									.80	–	.32			.32
a									.29	.32	–	.65	.75	.67
i											.65	–	.65	.49
u								.20	.32		.75	.65	–	.59
ʌ									.37	.32	.67	.49	.59	–

de l'unité représentant le mot /kʌp/ dans la tranche numéro 24 représente l'hypothèse selon laquelle l'entrée contient le mot /kʌp/ et que son début correspond à la tranche temporelle 24.

8.2.2 Mécanismes de traitement du flux de parole

Traitement par cycle

Le traitement de l'entrée consiste en huit étapes consécutives par cycle :

1. activation des traits par l'entrée
2. inhibition latérale des traits
3. activation des phonèmes par les traits
4. inhibition latérale des phonèmes
5. activation des traits par les phonèmes (si le feedback est activé)
6. activation des mots par les phonèmes
7. activation des phonèmes par les mots (feedback)
8. inhibition latérale des mots

L'activité par défaut des unités est fixée à 0, l'activité minimale est de -3 et l'activité maximale est de 1. Lorsque l'activité d'une unité est supérieure à 0, elle peut être transmise à d'autres unités via des connexions excitatrices et inhibitrices. Les étapes d'interaction se réalisent pendant chaque cycle de traitement, ce qui conduit à une mise à jour des activités de toutes les unités dans TRACE à la fin de chaque

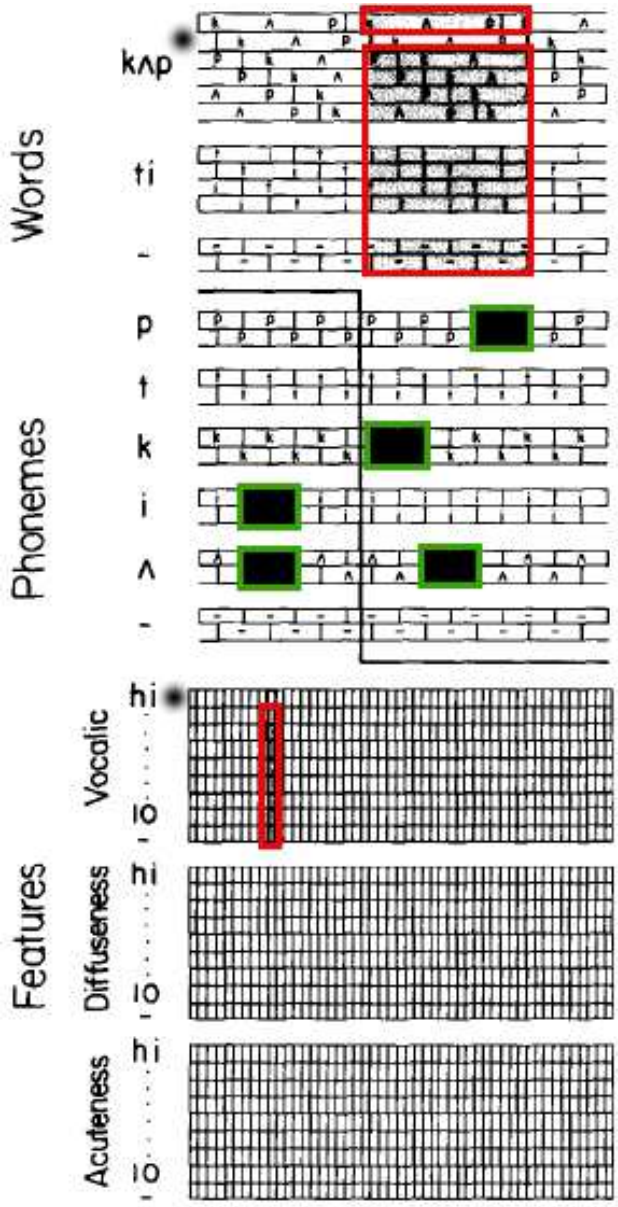


FIGURE 8.3 : Activation des unités et leurs connexions dans le modèle TRACE. Les connexions illustrées sont celles des unités correspondant à la valeur maximale du trait vocalique et au mot « cup » (signalées par un point à gauche) commencant respectivement à la tranche temporelle numéro 9 et 24. Les connexions excitatrices sont illustrées en noir (entourées en vert, niveau phonémique) et les connexions inhibitrices en gris (entourées en rouge, niveau de traits et niveau lexical). Figure tirée de McClelland et Elman (1986).

cycle. Ainsi, lors du premier cycle de traitement, l'entrée est présentée à la première partie des unités représentant les traits phonétiques, les étapes de traitement s'effectuent les unes après les autres et les activités des unités sont mises à jour. Pendant le deuxième cycle, l'entrée est présentée à la deuxième partie de ces unités et ainsi de suite. Nous rappelons que le décours temporel du passage du stimulus dans TRACE est représenté par une extension spatiale des unités. Autrement dit, il existe deux notions de temps dans TRACE : la première représente le temps de passage du stimulus qu'on appelle les cycles de traitement et la deuxième concerne l'alignement temporel des unités par rapport à l'entrée.

Paramètres

Le tableau 8.2 présente les paramètres utilisés dans TRACE. Ces paramètres sont fixés à la main. McClelland et Elman (1986) notent que ces paramètres ne sont pas directement comparables d'un niveau à l'autre. Par exemple, l'inhibition entre les traits phonétiques concerne seulement les traits qui sont dans la même tranche temporelle tandis que l'inhibition entre les phonèmes et entre les mots est en fonction de leur recouvrement temporel (voir figure 8.3 pour les inhibitions entre les mots). Les valeurs des paramètres réglant les connexions excitatrices influencent généralement l'amplitude ou la dynamique d'un effet sans que la nature des effets observés soit touchée. Par exemple, des excitations bottom-up plus fortes peuvent augmenter les effets top-down tel que l'effet lexical dans une tâche d'identification phonémique : l'activation plus importante des unités lexicales entraîne des feedbacks plus forts à partir du niveau lexical vers les phonèmes, ce qui augmente à son tour le degré d'activité des phonèmes constituant ces unités lexicales.

Table 8.2 : Paramètres du modèle TRACE.

Paramètre	Valeur
Excitation trait-phonème	.02
Excitation phonème-mot	.05
Excitation mot-phonème	.03
Excitation phonème-trait	.00
Inhibition entre les traits	.04
Inhibition entre les phonèmes	.04
Inhibition entre les mots	.03
Atténuation des activités des traits	.01
Atténuation des activités des phonèmes	.03
Atténuation des activités des mots	.05

En revanche, les paramètres réglant la valeur d'inhibition entre les unités du même niveau peuvent changer le comportement du modèle. En effet, une inhibition forte entre les unités rend le modèle très sensible à de petites différences entre l'activité des unités au début de la simulation. Cet effet peut empêcher, dans certaines situations, l'influence du contexte en éliminant les unités peu activées mais

pertinentes dès le début de la simulation.

Le paramètre d'atténuation permet aux activités des unités de persister pendant des cycles autres que le cycle où l'entrée correspondante est fournie au modèle. Notamment, une valeur plus petite dans le niveau de traits phonétiques permet une extension temporelle plus large des activités de ces unités. La figure 8.4 illustre ce paramètre dans les trois niveaux de TRACE. Dans le niveau de traits phonétiques, la baisse du niveau de couleur grise représente l'atténuation de l'activité des traits (voir également figure 8.2). Dans les niveaux phonémique et lexical, la couleur verte claire illustre cette atténuation.

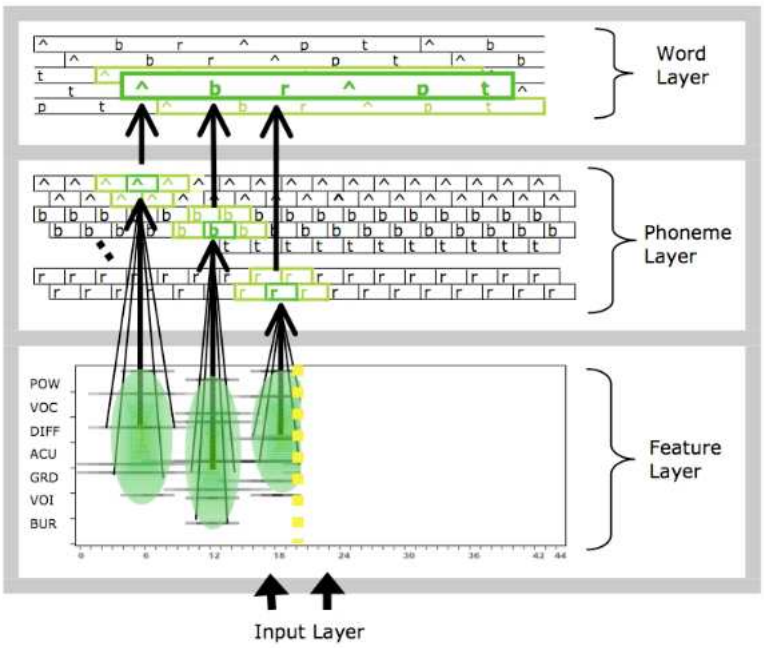


FIGURE 8.4 : Atténuation des activités dans le modèle TRACE. Le changement de couleurs du foncé vers le clair représente l'atténuation linéaire dans chaque niveau. Entrée : « abrupt ». Figure tirée de Strauss *et al.* (2008).

Activités des unités

L'activité des unités dans le niveau de traits phonétiques est la somme des valeurs suivantes :

- entrée bottom-up calculée par une fonction texte-à-« parole » (stimulus).
- entrée top-down du niveau phonémique (si le feedback entre les phonèmes et les traits est actif).
- entrée négative due à l'inhibition : les traits de même dimension correspondant à la même tranche de temps sont mutuellement inhibiteurs.
- activité de l'unité pendant les cycles précédents.

Le degré d'activité des unités phonémiques est la somme des valeurs suivantes :

- entrée bottom-up des traits dans les mêmes tranches de temps.
- entrée top-down du niveau lexical vers le niveau phonémique.
- entrée négative due à l'inhibition des autres unités phonémiques en fonction de leur degré de recouvrement temporel. Par exemple, un détecteur de phonème inhibe les autres détecteurs situés dans la même tranche de temps deux fois plus que ceux situés à trois tranches de temps plus loin.
- activité de l'unité pendant les cycles précédents.

Enfin, le degré d'activité des unités lexicales est la somme des valeurs suivantes :

- entrée bottom-up des phonèmes qu'elle contient (dans l'exemple présenté sur la figure 8.3, l'unité « cup » reçoit l'excitation du phonème /k/ proche de la tranche numéro 24, du /ʌ/ proche de la tranche numéro 30 et du /p/ proche de la tranche numéro 36).
- entrée négative due à l'inhibition des autres unités lexicales. De la même façon que dans le niveau phonémique, cette inhibition dépend du degré de recouvrement temporel des mots : les unités lexicales représentant les différentes hypothèses étendues sur les même tranches temporelles des phonèmes s'inhibent les unes les autres et celles représentant les hypothèses sur des tranches non-recouvrantes n'entrent pas en compétition (voir figure 8.3).
- activité de l'unité pendant les cycles précédents.

Décision

Le choix entre les différents candidats dans le niveau lexical et le niveau phonémique est basé sur la règle de Luce (1959) : la probabilité de choisir l'unité i comme sortie est

$$p(R_i) = \frac{e^{ka_i}}{\sum_j e^{ka_j}} \tag{8.1}$$

où j comprend tous les choix possibles. a_i étant l'activité de l'unité i. La transformation exponentielle de a garantit que toutes les activités sont positives et le dénominateur assure que la somme de toutes les probabilités est égale à 1.

Segmentation

Dans cette thèse, nous nous intéressons particulièrement à la segmentation du flux de parole. Afin de réaliser une segmentation de la séquence d'entrée dans TRACE, il n'est pas nécessaire d'avoir des indices spécifiant les frontières des mots. En effet, la segmentation dans TRACE est le résultat de l'activation bottom-up et de la compétition entre les unités dans le niveau lexical. Nous rappelons que le degré d'inhibition entre ces unités est en fonction de leur recouvrement temporel. Ainsi,

seuls les candidats qui se chevauchent entrent en compétition l'un avec l'autre. Ces mécanismes d'activation et de compétition conduisent généralement à une bonne segmentation du flux de parole. Cependant, lorsque deux mots peuvent être incorporés dans un seul mot long, TRACE favorise le mot le plus long. Par exemple, en présentant /parti/ à l'entrée de TRACE, l'unité représentant « party » sera plus active que « par » et « tea ». Ceci vient du fait que les mots longs reçoivent des excitations bottom-up plus importantes grâce au nombre plus élevé de phonèmes les excitant et que l'entrée est présentée d'une façon séquentielle. Dans cet exemple, lorsque /ti/ est présenté au modèle, l'unité lexicale « party » est déjà activée et inhibe l'unité représentant « tea ».

8.3 Architecture et mécanismes du modèle TRACE-VT

Dans cette section nous présentons les éléments que nous avons ajouté au modèle initial TRACE dans ce travail. Ces éléments ont été déjà brièvement présentés dans la section 8.1. Dans cette section, nous poursuivons cette introduction en expliquant en détail les mécanismes que nous avons intégrés à TRACE afin de rendre compte de l'effet de transformation verbale. Pour pouvoir distinguer facilement ce modèle du modèle TRACE et montrer qu'il en constitue une proposition de développement, il sera désigné dans cette thèse comme TRACE-VT (pour *Verbal Transformation*).

Lors de la modélisation computationnelle d'un processus cognitif, nous sommes amenés à intégrer des éléments qui ne font pas nécessairement partie de nos hypothèses d'ordre théorique sur le processus cognitif en question. De ce point de vue, il est important de séparer au sein d'une implémentation computationnelle les hypothèses centrales (théoriques) et périphériques (non-théoriques) (Norris, 2005, voir sous-section 7.1.2). Nos hypothèses centrales sont basées sur les résultats des expériences réalisées dans le cadre de cette thèse, sur la littérature sur l'effet de transformation verbale et, dans une moindre mesure, sur la littérature sur la multistabilité perceptive en vision. En revanche, nos hypothèses périphériques sont uniquement faites afin de rendre possible ou simplifier l'implémentation computationnelle de nos hypothèses centrales. Avant de détailler les mécanismes intégrés au modèle TRACE, nous présentons la liste de nos hypothèses centrales. Quant aux hypothèses périphériques, elles seront présentées au fur et à mesure dans les sous-sections correspondantes.

Les hypothèses centrales de ce travail sont les suivantes. D'abord, nous posons trois hypothèses nécessaires pour faire émerger des transformations verbales dans TRACE :

- C1 – pré et post-traitement : il existe préalablement au modèle TRACE un processus de liage et postérieurement, un processus de décision.

- C2 – stochasticité : le processus de décision est stochastique, et non pas déterministe. La vraisemblance d'une réponse active ce processus, les réponses les plus vraisemblables étant affectées de probabilités plus élevées.

- C3 – adaptation : il existe au niveau des percepts un processus de saturation (adaptation) permettant à de nouvelles représentations d'émerger.

C'est l'ensemble de ces trois hypothèses qui assure l'existence du phénomène de transformations verbales, quel qu'en soit le contenu. Les hypothèses C4 à C6 sont elles spécifiquement adaptées à nos données expérimentales, et peuvent être considérées comme des résultats à part entière de nos recherches, exprimant les résultats sous forme d'hypothèses computationnelles.

- C4 – liage articulatoire : il existe un biais articulatoire dans le processus de liage favorisant les liages cohérents avec le geste d'ouverture de la mâchoire.
- C5 – multisensorialité : le degré d'ouverture de la mâchoire peut être récupéré à partir de la modalité auditive et visuelle.
- C6 – fusion audio-visuelle : la valeur perceptive correspondant au degré d'ouverture de la mâchoire est une combinaison (à définir) des valeurs auditive et visuelle.

La figure 8.5 illustre la structure du modèle TRACE-VT. Nous commentons dans la suite les différentes composantes illustrées sur cette figure.

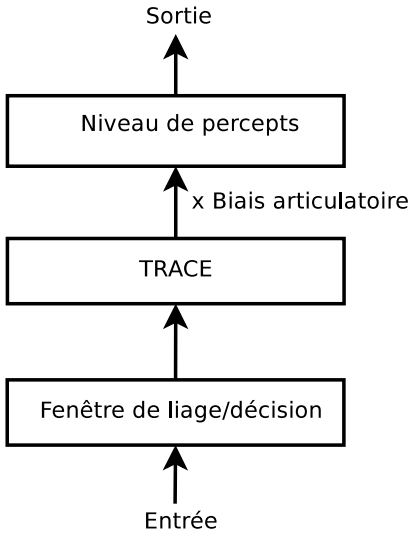


FIGURE 8.5 : Architecture du modèle TRACE-VT.

8.3.1 Fenêtre de liage/décision

Comme décrit dans la section 8.1, un traitement purement séquentiel du flux de parole avec l'attribution point-par-point de l'entrée à la sortie ne peut pas rendre compte des transformations verbales. Supposons que la séquence /patapatapata.../ soit présentée à l'entrée de TRACE (la sortie initiale de TRACE est donc /pata/) et que le percept /tapa/ puisse émerger pendant le passage de cette séquence. Lors

d'une bascule de /pata/ vers /tapa/, il existe deux possibilités concernant l'alignement de la sortie avec l'entrée : soit il doit rester une syllabe /pa/ à l'entrée sans qu'elle soit attribuée ni à /pata/ ni à /tapa/ en sortie (/*patapatapa*.../) soit une syllabe /ta/ de l'entrée doit être utilisée dans les deux percepts /pata/ et /tapa/ (/*patapa*.../). Dans le premier cas, la sortie de TRACE sera /pata/+/pa/+/tapa/, ce qui n'est pas cohérent avec nos données expérimentales. Quant au deuxième cas, il ne pourra pas se produire à cause des compétitions entre les unités et la règle de décision dans TRACE. Face à ce problème, nous proposons les mécanismes suivants (hypothèses périphériques) :

- P1 : le liage temporel entre différentes parties du flux de parole se réalise à l'intérieur d'une fenêtre temporelle.
- P2 : une décision perceptive est prise pour chaque fenêtre temporelle ce qui conduit à la stabilisation du percept précédent ou à une transformation verbale.
- P3 : après chaque décision, la fenêtre glisse à la fin du nouveau percept pour fournir une nouvelle entrée au modèle.

La figure 8.6 illustre les mécanismes proposés ci-dessus pendant la présentation de /pata/ en boucle.

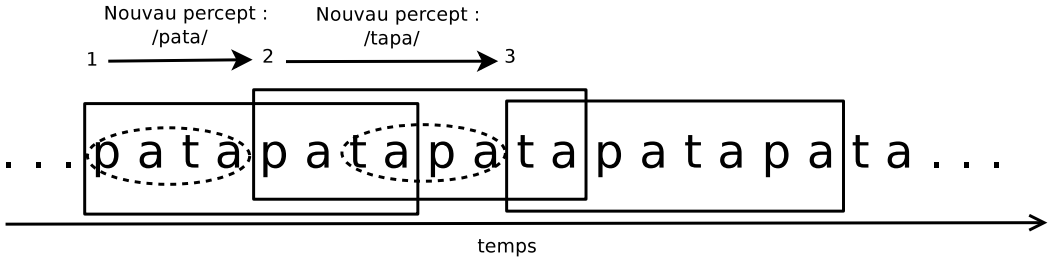


FIGURE 8.6 : Fenêtre de liage/décision d'une taille de 4 syllabes (les rectangles). Les chiffres représentent le numéro des fenêtres. Les ovales en pointillé illustrent le résultat des processus de liage/décision, c'est-à-dire le nouveau percept.

8.3.2 Niveau de percepts

Afin de pouvoir séparer les processus en lien avec les percepts et ceux concernant la reconnaissance des séquences de parole et en accord avec l'hypothèse C1, un niveau de percepts a été ajouté au modèle TRACE. Un des avantages de la séparation entre le niveau lexical de TRACE et le niveau de percepts concerne la représentation temporelle utilisée dans TRACE. Nous avons vu que les unités lexicales dans TRACE sont dupliquées spatialement afin de rendre compte du décours temporel du stimulus et de l'alignement de la sortie avec l'entrée. L'utilisation d'un niveau de percepts avec une fenêtre de liage/décision nous permet d'avoir une seule unité représentant chaque percept. La distinction entre les unités lexicales et pré-lexicales

et les unités décisionnelles a été également suggérée par Norris *et al.* (2000) dans le cadre du modèle psycholinguistique Merge. Dans ce modèle, les unités décisionnelles rassemblent les résultats du traitement effectué dans les niveaux lexical et pré-lexical et fournissent une sortie en fonction des ces entrées.

La figure 8.7 illustre l'architecture du niveau de percepts. Les mécanismes mis en place dans ce niveau sont les suivants :

- P4 : les unités ayant la même valeur phonétique que les unités lexicales représentent les percepts.
- P5 : la sortie des unités lexicales fournit l'entrée du niveau de percepts : le percept cohérent avec une unité lexicale reçoit l'activité de cette unité via une connexion excitatrice (m) et les percepts incohérents reçoivent cette activité via des connexions inhibitrices ($-m$).
- P6 : l'activation des unités et leurs interactions sont similaires à celles d'une machine de Boltzmann que nous présentons ci-dessous.

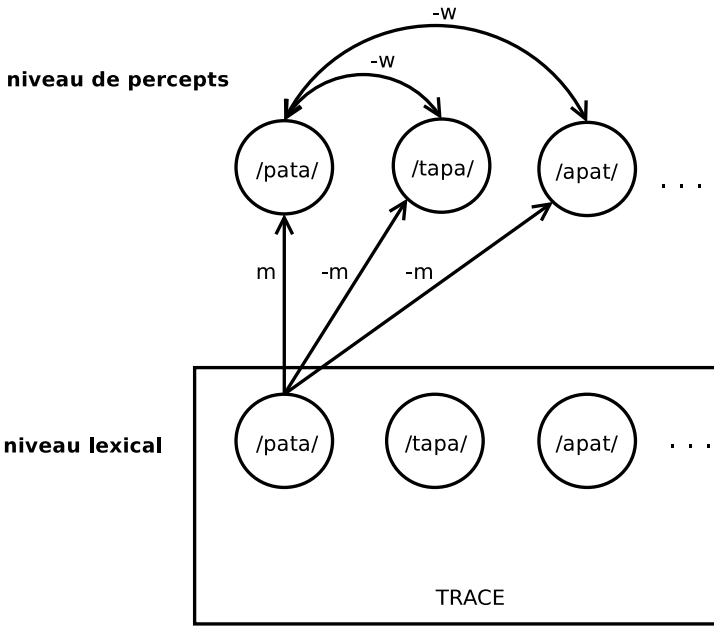


FIGURE 8.7 : Niveau de percepts dans TRACE-VT et ses connexions avec le niveau lexical. m et w sont positifs.

La machine de Boltzmann proposée par (Hinton et Sejnowski, 1986) est un réseau de neurones de type Hopfield (Hopfield, 1982) (ses unités sont symétriquement interconnectées sans feedback à soi-même) dont l'activation ou non dépend d'une fonction stochastique :

$$\text{Sortie de l'unité}_k\ (y_k) = \begin{cases} 1 & \text{avec la probabilité } P_k = \frac{1}{1+e^{\frac{-\Delta E_k}{T}}} \\ -1 & \text{sinon} \end{cases} \tag{8.2}$$

T est équivalent à la température d'un système physique. ΔE_k est le changement dans l'énergie totale du réseau si l'état du neurone$_k$ change ($\Delta E = E_{après} - E_{avant}$). L'énergie du réseau dans l'état α correspond à :

$$E_\alpha = -\frac{1}{2}\sum_{i,j} y_i w_{i,j} y_j + \sum \theta_i y_i \tag{8.3}$$

où y_i est la sortie de l'unité i, $w_{i,j}$ est le poids de connexion de l'unité i vers l'unité j et θ_i est le seuil de l'unité i.

Comme les poids dans le réseau sont symétriques et qu'il n'y pas de feedback d'un neurone à lui-même, le changement de l'énergie totale se calcule localement par ΔE = entrée totale du neurone$_k$ (entrées internes et externes). Si les unités sont mises à jour les unes après les autres dans un ordre quelconque indépendant de leurs entrées, le réseau arrive à la distribution de Boltzmann (état d'équilibre) où la probabilité d'état α est déterminée seulement par l'énergie relative de cet état par rapport à l'énergie de tous les états possibles (β) :

$$P\alpha = \frac{e^{-E_\alpha}}{\sum_\beta e^{-E_\beta}} \qquad \beta = \text{tous les états} \tag{8.4}$$

Lorsque la température est basse (T dans l'équation 8.2), il y une préférence dans le réseau pour les états de basse énergie (minima locaux) mais la durée nécessaire pour arriver à un équilibre est longue. Au contraire, pour les températures élevées, les états de faible énergie ne sont pas favorisés mais le réseau arrive plus rapidement à son équilibre. La façon la plus rapide et la plus efficace pour arriver à l'équilibre dans ce réseau consiste généralement à utiliser l'algorithme de *simulated annealing* (Hinton et Sejnowski, 1986) : la baisse progressive de la température d'une valeur initiale élevée à une valeur faible permet à la fois de bénéficier du fait que l'équilibre est obtenu plus rapidement pour les T élevées et d'arriver à un état final de relativement faible énergie. Lorsque T est égal à zéro, la règle de mise à jour (équation 8.2) devient déterministe et la machine de Boltzmann se comporte comme un réseau de Hopfield. Il est à noter qu'une machine de Boltzmann a généralement des unités cachées (*hidden units*) en plus des unités représentant des états de sortie (*visible units*). Les unités cachées permettent de modéliser les distributions qui ne sont pas possibles à modéliser directement par des interactions par paires entre les unités visibles. Dans ce travail, nous n'utilisons pas d'unités cachées.

La modélisation de la perception par un tel réseau consiste à attribuer les percepts stables aux états qui forment les minima locaux du système qu'on appelle les états d'attracteurs. Ainsi, en présentant le pattern d'un percept à l'entrée, le réseau converge vers l'état le représentant. Dans TRACE-VT, l'état α représente la forme phonétique α, $P\alpha$ est la probabilité de percevoir cette forme lorsque le système est en équilibre et $E\alpha$ est l'énergie du réseau correspondant au percept α. Les connexions inhibitrices entre ces percepts ont été choisies en accord avec un réseau Hopfield *flipflop* (voir figure 8.8) et *multiflop* dont les états stables consistent en une seul unité avec l'activité 1 et le reste avec l'activité -1 (Rojas, 1996). Les valeurs des paramètres utilisés dans le niveau de percepts seront présentées dans la section 8.4.

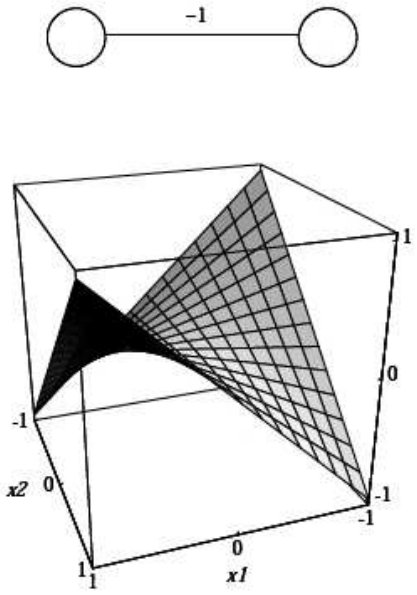


FIGURE 8.8 : Réseau Hopflield *flipflop*. En haut : l'architecture du réseau. Seuil = 0. En bas : la fonction d'énergie du système. $x1$ et $x2$ sont respectivement la sortie de la première et de la deuxième unité. (1,-1) et (-1,1) sont les états stables de ce réseau. Figure tirée de Rojas (1996).

Il est à souligner que nous n'utilisons pas une machine de Boltzmann proprement dite mais certaines notions présentées dans ce cadre. En effet, nous faisons l'hypothèse qu'à la fin de chaque fenêtre de liage/décision, le réseau arrive à son équilibre, autrement dit, la dynamique du réseau entre la phase de présentation de l'entrée au niveau des percepts et la phase de décision pour activer un percept n'est pas prise en compte. De plus, la décision se fait selon l'équation 8.4 seulement sur les états de type (1, -1, -1, ...)[1], (-1, 1, -1, ...)[2], etc. : pour chaque fenêtre de liage/décision, les probabilités des percepts sont calculées, un de ces états sera actif en réglant l'activité d'une unité à 1 et l'activité des autres unités à -1. Ainsi, nous excluons à la main les états parasites.

Le choix d'activation des percepts basé sur la machine de Boltzmann vient du fait que nos travaux expérimentaux sur l'effet de transformation verbale ne visaient pas à étudier la dynamique des transformations mais les tendances générales concernant la stabilité globale des percepts. La notion de convergence du système vers les différents états en fonction de leur énergie permet d'étudier ces tendances générales. Ainsi, les percepts conduisant le système à une énergie plus basse sont plus probables. Bien que certains états soient plus probables que d'autres, l'utilisation des unités stochastiques évite que le réseau s'installe, d'une manière déterministe, dans ces états plus probables, ce qui peut conduire à des transformations verbales.

Il est important de noter ici la difficulté du traitement en temps réel du flux de

[1]Cet état représente le percept /pata/ sur la figure 8.7.
[2]Cet état représente le percept /tapa/ sur la figure 8.7.

parole dans un réseau d'attracteur tel qu'une machine de Boltzmann. En effet, ces réseaux fournissent des sorties stables seulement lorsqu'ils s'installent dans des états stables tandis qu'un stimulus variant dans le temps conduit le réseau à des états intermédiaires transitoires. Face à ce problème, comme décrit ci-dessus, nous avons choisi de ne vérifier les unités représentant des percepts que lors de l'équilibre du système. Autrement dit, les états intermédiaires ne sont pas pris en compte. Une alternative à un réseau de neurones classique pour traiter les entrées variantes avec le temps comme le signal de parole est proposée par Maass *et al.* (2002). Ce modèle computationnel appelé *Liquid State Machine* (LSM) peut transformer en temps réel les états intermédiaires non-stables en sorties stables en utilisant les unités *readout*. Dans LSM, les unités sont inter-connectées d'une façon aléatoire. Les connexions sont récurrentes et transforment l'entrée du modèle en une représentation spatio-temporelle à partir des activités des unités. Les unités *readout* appliquent ensuite une fonction de discrimination sur ces réponses afin de fournir la sortie du modèle. Il nous semble que la combinaison d'un tel réseau avec le modèle psycholinguistique TRACE est peu pertinente, voire impossible, car leur approche de la question du traitement du flux de parole se fait à un niveau computationnel différent.

8.3.3 Adaptation

Afin de mettre en place un mécanisme de saturation des unités perceptives (hypothèse C3), nous avons intégré un mécanisme d'adaptation :

- P7 : l'adaptation agit sur le seuil des percepts. Lorsqu'un percept est actif, son seuil augmente, ce qui entraîne une augmentation de l'énergie et une baisse de sa probabilité (voir équation 8.3). Cette baisse favorise l'activation d'un autre percept et pourrait conduire à une transformation verbale.

- P8 : l'équation 8.5 présente la fonction du changement de seuil utilisée dans ce travail. Elle est basée sur les modèles de type LIF (*Leaky Integrate Fire*) (Stein, 1967) où le seuil des neurones baisse d'une façon exponentielle après une décharge et augmente après un potentiel d'action. Un circuit électrique composé d'une résistance et d'un condensateur montés en parallèle peut modéliser ce comportement. Dans ce cas, la constante du temps du système (τ) correspond à RC. Lorsqu'un percept est activé, le changement de seuil correspond à la phase de chargement du condensateur et lorsqu'elle est inactive, il suit la fonction qui correspond au déchargement du condensateur.

$$\theta(t) = \begin{cases} \theta_0(1 - \exp^{-\frac{t}{\tau}}) & \text{si le percept est actif} \\ \theta_i \exp^{-\frac{t}{\tau}} & \text{sinon} \end{cases} \tag{8.5}$$

8.3.4 Biais articulatoire

La préférence pour certains liages par rapport à d'autres en fonction du geste de la mâchoire récupéré à partir du signal audio-visuel de la parole (hypothèses C4, C5, C6) est modélisée de la façon suivante :

- P9 : les sorties du niveau lexical sont multipliées par des coefficients représentant la dynamique d'ouverture de la mâchoire qui amplifient ou non les entrées des unités perceptives.

- P10 : afin de représenter cette dynamique, le coefficient correspondant au phonème X est proportionnel à l'ouverture de la mâchoire lors la production de la syllabe XV, V étant la voyelle la plus ouverte, i.e. /a/.

- P11 : pour chaque modalité, ces coefficients sont calculés séparément lorsque l'entrée est fournie au modèle par une fonction texte-à-coefficient similaire à celle utilisé pour la traduction de l'entrée en traits phonétiques dans TRACE. Ainsi, on obtient des coefficients J_A et J_V correspondant respectivement au signal auditif et visuel.

- P12 : la capacité perceptive à récupérer les coefficients J est plus importante pour les flux visuels qu'auditifs.

- P13 : lors d'une présentation audio-visuelle, ces biais articulatoires correspondent à une somme pondérée des coefficients en modalité auditive et visuelle :

$$J_{AV} = a \times J_A + b \times J_V \tag{8.6}$$

Concernant l'hypothèse P12, précisons que nous avons observé un effet LC[3] plus fort dans la modalité audio-visuelle qu'en modalité auditive (voir sections 4.3 et 4.4). Si l'effet LC a des origines articulatoires comme nous le défendons dans cette thèse, son augmentation lors d'une présentation audio-visuelle suggère l'implication plus forte de ces éléments articulatoires par rapport à la modalité auditive. Il est important à noter ici que l'effet LC semble être plus fort dans la modalité visuelle pure qu'en modalité auditive et audio-visuelle (Basirat, 2005). En effet, le geste d'ouverture de la mâchoire est naturellement visible. Il serait donc légitime que « sa perception » par les sujets soit plus forte lorsqu'ils reçoivent des informations visuelles que lors d'une présentation auditive, d'où notre hypothèse P12. Nous proposons dans la section 9.2.1 des perspectives expérimentales afin de vérifier cette hypothèse d'une façon plus directe.

La figure 8.9 illustre les mécanismes présentés ci-dessus lors du passage du stimulus /patapata.../. L'activation des unités qui commencent par /pa/ (dynamique importante du geste d'ouverture) est plus amplifiée que celle des unités commençant par /ta/ (présentant une dynamique moins importante du geste d'ouverture) et *a fortiori* encore plus que celles qui commencent par /ap/ ou /at/ (geste de fermeture). Cette préférence conduit d'abord à l'émergence de syllabes CV plutôt que VC, résultat classique et majeur des transformations verbales. Ce résultat est également en accord avec les théories de la sonorité selon lesquelles la structure d'une syllabe est organisée de sorte que la sonorité des phonèmes augmente jusqu'à un

[3] C'est-à-dire la préférence pour les séquences Labial-Coronal comme /pata/ par rapport aux séquences Coronal-Labial comme /tapa/, voir sous-section 3.3.2.

sommet et décroit ensuite (Kent, 1993, pour une revue). Les mécanismes proposés ici généralisent cette préférence à l'effet LC : dans cette implémentation, LC est à CL (ou /pata/ est à /tapa/) ce que CV est à VC (ou /pa/ à /ap/). Ainsi, l'augmentation relative de l'entrée externe du percept /pata/ conduit à l'augmentation de la probabilité du percept /pata/ par rapport à celle de /tapa/, /apat/ ou /atap/. Quant à l'effet de l'onset visuel décrit dans la section 4.3, la présentation visuelle des syllabes /pa#a/ et /ta#a/ conduira respectivement à une probabilité plus forte de /pata/ et /tapa/.

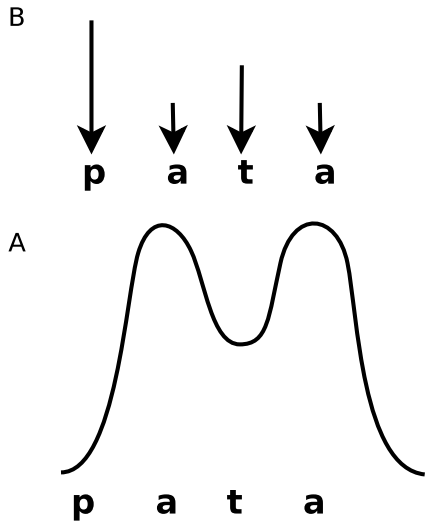


FIGURE 8.9 : Rôle de biais articulatoire dans TRACE-VT. A : la courbe représente l'ouverture labiale correspondant à la séquence /pata/. B : les flèches illustrent les coefficients J, leur longueur représente les valeurs de J et leur position représente la partie du signal amplifiée, i.e. le début d'une segmentation.

8.4 Réglage des paramètres

Le tableau 8.3 représente les valeurs que nous avons utilisées pour les biais articulatoires (matrice J) dans nos simulations. En accord avec nos hypothèses, ces valeurs renforcent l'entrée excitatrice de l'unité perceptive /pata/ plus que /tapa/ et encore plus que /atap/ ou /apat/. Ce pattern de renforcement s'applique également aux entrées inhibitrices des unités perceptives : l'unité lexicale /pata/ inhibe plus fortement les autres percepts que l'unité /tapa/ et encore plus que les unités /atap/ ou /apat/. Bien que ces valeurs avec celles présentées dans le tableau 8.4 conduisent à des simulations cohérentes avec nos données, aucune expérience comportementale n'a été réalisée ni sur la capacité perceptive des sujets à extraire ces coefficients à partir du stimulus ni sur ces valeurs elle-mêmes. Il est cependant possible de vérifier expérimentalement nos hypothèses concernant ces valeurs (voir sous section 9.2.1).

Le tableau 8.4 représente les paramètres du modèle avec leurs valeurs respectives utilisées dans les simulations. Ces valeurs ont été fixées à la main de façon à pouvoir

TABLE 8.3 : Valeurs de la matrice J représentant les biais articulatoires pour les entrées auditives et visuelles.

	p	t	a
Entrée auditive	1	0.95	0.20
Entrée visuelle	1	0.40	0.10

rendre compte de nos données expérimentales.

TABLE 8.4 : Paramètres du modèle TRACE-VT.

Paramètre	Valeur
Taille de la fenêtre de liage/décision en syllables	4
Excitation entre les unités lexicales et les percepts cohérents (m)	8
Inhibition entre les unités lexicales et les percepts incohérent (m)	8
Inhibition entre les unités dans le niveau de percepts (w)	1
Constante de temps dans l'équation de l'adaptation (τ)	20
Seuil maximum (θ_0)	5
Coefficient de pondération des biais articulatoires auditifs en modalité audio-visuelle (a)	1
Coefficient de pondération des biais articulatoires visuels en modalité audio-visuelle (b)	0.1

Il est plus ou moins possible de suivre l'influence des valeurs de ces paramètres sur le comportement du modèle. Les coefficients de pondération a et b peuvent directement influencer la stabilité des percepts et les effets observés. Par exemple, pour les valeurs J du tableau 8.3 et a constante, l'augmentation de b conduit à une augmentation de l'excitation du percept /pata/ par rapport au percept /tapa/ lorsque le flux visuel est /pa#a/ et, au contraire, à une excitation plus importante du percept /tapa/ par rapport au percept /pata/ lorsque le flux visuel est /ta#a/. Ainsi, l'effet de l'onset visuel observé dans nos expériences comportementales sera plus important. Concernant le paramètre τ, son augmentation rend la phase d'adaptation plus lente, ce qui entraîne une diminution du nombre de transformations. La valeur du paramètre m peut refléter le degré d'influence de l'entrée externe à partir des unités lexicales vers les percepts.

Concernant le paramètre w, il est important de préciser que pour que le niveau de percepts représente une machine de Boltzmann ou un réseau de Hopfield stochastique, il faut choisir les valeurs de w de sorte que les percepts (états stables) forment les minima de l'énergie du réseau. Cette analyse n'a pas été effectuée car nous ne considérons que les états stables représentant des percepts dans ce réseau. Autrement dit, comme décrit dans la sous-section 8.3.2, le niveau de percepts n'est pas un vrai réseau de Hopfield.

8.5 Simulations et discussion

Les figures présentées à la fin de cette section (à partir de la page 190) illustrent différentes simulations de TRACE-VT avec les valeurs des paramètres présentées dans la section précédente. Les transformations verbales possibles sont /pata/, /tapa/, /atap/, /apat/, /pa/ et /ta/. Le modèle TRACE utilisé dans les simulations est la version originale implémentée par McClelland et Elman (1986) en langage C. Les mécanismes que nous avons intégrés à cette version sont aussi écrits en langage C. L'environnement de travail est l'environnement Linux. Les figures illustrant les données de simulations ont été réalisées en utilisant le logiciel Octave-3.2.2.

La première série de simulations (figures 8.10 et 8.11) correspond au modèle avec les biais articulatoires et l'adaptation. De sorte qu'on puisse comparer ces résultats et nos données expérimentales, pour chaque série de simulation, nous y présentons la durée de stabilité des percepts /pata/, /tapa/ et « autres », le *delta*[4] et le nombre de transformations en modalité auditive, audio-visuelle, audio-visuelle /pa#a/ et audio-visuelle /ta#a/[5]. Afin de mieux comprendre l'apport de ces deux mécanismes, nous présentons également trois autres séries de simulations correspondant à TRACE-VT sans biais articulatoire (figures 8.12 et 8.13), sans adaptation (figures 8.14 et 8.15) et sans biais articulatoire ni adaptation (figures 8.16 et 8.17). La comparaison entre ces différentes simulations met en évidence les traits saillants de nos résultats que nous décrivons dans la suite.

8.5.1 Rôle de l'adaptation

En comparant les simulations sans adaptation avec celles avec adaptation, nous pouvons constater que le mécanisme d'adaptation conduit à des bascules perceptives. Il est cependant à noter que le caractère probabiliste du niveau des percepts peut conduire, même en absence de l'adaptation, à des transformations verbales mais le nombre de bascules reste très faible comparativement aux données expérimentales. Notamment, lors de la présence des biais articulatoires, l'effet de transformation verbale a tendance à disparaître (voir figure 8.15). Ce résultat vient du fait que les biais articulatoires renforcent la probabilité du percept cohérent avec l'unité lexicale ayant l'activité maximum (séquence initiale, /pata/ ou /tapa/). Ainsi, sans adaptation, cette probabilité reste élevée tout au long d'une simulation ce qui entraîne une stabilité importante de la séquence initiale surtout dans le cas du stimulus /pata/.

8.5.2 Rôle des biais articulatoires

L'effet des biais articulatoires rend possible la préférence de certaines formes à d'autres (CVCV à VCVC et LC à CL). Deux tendances peuvent être observées sans biais articulatoire. En présence de l'adaptation les *delta* convergent vers zéro et la

[4] Nous rappelons que la valeur de *delta* correspond à la durée de la stabilité globale relative du percept /pata/ moins celle du percept /tapa/.

[5] Lorsque le stimulus auditif est /pata/, l'entrée visuelle du modèle dans les modalités AV, AV-pa et AV-ta est respectivement /pata/, /paaa/ et /aata/. Le flux visuel pendant le passage du stimulus /tapa/ est /tapa/ (en modalité AV), /aapa/ (en modalité AV-pa) et /taaa/ (en modalité AV-ta).

stabilité des percepts « autres » devient très importante par rapport aux percepts /pata/ et /tapa/ (figures 8.12 et 8.13). En l'absence de l'adaptation, le percept initial reste relativement stable, ce qui conduit à une valeur delta positive pour les stimuli de type /pata/ et une valeur négative pour les stimuli de type /tapa/ (figures 8.16 et 8.17).

Les figures 8.14 et 8.15 illustrent les résultats d'une série de simulations utilisant les biais articulatoires sans adaptation. Nous pouvons constater que les transformations mono-syllabiques et celles de type /VCVC/ deviennent peu probables ce qui semble cohérent avec nos données expérimentales. Cependant, en l'absence de l'adaptation, les transformations verbales se produisent à peine et les percepts initiaux (/pata/ ou /tapa/) restent très stables. En effet, c'est en utilisant les biais articulatoires avec les mécanismes d'adaptation que nous pouvons trouver les tendances observées dans nos expériences expérimentales (figures 8.10 et 8.11). Notamment, en ce qui concerne l'effet de l'onset visuel, nous pouvons noter la différence de *delta* entre la modalité AV-pa et AV-ta : en accord avec nos résultats expérimentaux et nos hypothèses de modélisation, l'onset visuel /pa#a/ favorise le percept /pata/ et l'onset visuel /ta#a/ facilite la stabilité du percept /tapa/. Nous analysons ce résultat plus en détail dans la sous-section 8.5.4.

Il est à préciser que la préférence pour les percepts dissyllabiques /pata/ et /tapa/ par rapport aux percepts mono-syllabiques /pa/ et /ta/ ne provient pas des mécanismes que nous avons implémentés mais elle est le résultat direct du modèle TRACE qui favorise les séquences longues par rapport aux séquences courtes (voir sous-section 8.2.2).

8.5.3 Effet de la séquence auditive

Les données comportementales sur l'effet de transformation verbale montrent souvent un effet significatif de la séquence auditive (voir section 4 et Sato *et al.*, 2007) [sans que les données ne montrent clairement si cet effet perdure – par effet de mémoire – ou est dû simplement à la préférence initiale – onset]. L'effet de la séquence auditive dans TRACE-VT existe seulement avant que la première transformation verbale ait lieu. Après la première transformation, la trace de la séquence initiale n'est plus visible. Ainsi, l'effet de la séquence auditive disparaît rapidement au début du passage du stimulus. Une des possibilités pour intégrer cet effet dans TRACE-VT serait d'y ajouter des processus de type mémoire. N'ayant pas d'hypothèse théorique sur cet effet et en l'absence d'études expérimentales sur les mécanismes le produisant, nous avons préféré ne pas l'inclure dans ce travail.

8.5.4 Analyse des valeurs de *delta*

Dans cette sous-section, nous analysons les différences entre les valeurs de *delta* dans les différentes modalités de présentation utilisées dans nos simulations. Nous présentons également des comparaisons entre ces simulations et nos résultats expérimentaux. Pour cela, nous prenons en compte les tendances observées dans les données et non pas les valeurs exactes des durées de stabilité et de *delta*.

Effet LC

La figure 8.11 montre que les valeurs de *delta* sont positives en modalité A et AV. Ce résultat vient du fait que les biais articulatoires renforcent plus l'activité de l'unité lexicale /pata/ que celle de /tapa/ à la sortie de TRACE. Ainsi, ces activités fournissant l'entrée externe du niveau des percepts entraînent une préférence pour le percept /pata/ par rapport à /tapa/.

Modalité A vs. AV

Selon nos hypothèses présentées dans la section 8.3, la récupération du geste d'ouverture de la mâchoire (biais articulatoires) doit être plus importante à partir du flux visuel que du flux auditif. Cette hypothèse avec les valeurs sélectionnées pour les coefficients a et b conduit à un renforcement de l'entrée des percepts /pata/ et /tapa/ en modalité audio-visuelle par rapport à la modalité auditive. Étant donné que l'écart entre les valeurs J_V de /p/ et /t/ est plus important que celui entre les valeurs J_A de /p/ et /t/, le percept /pata/ est plus renforcé que le percept /tapa/ en modalité AV. Ainsi, la valeur *delta* est plus grande dans la modalité AV que dans la modalité A. Cette tendance est cohérente avec nos données expérimentales.

Modalité AV vs. AV-pa

Nos simulations montrent que la valeur *delta* est plus élevée dans la modalité AV-pa que dans la modalité AV. Ce résultat est directement prévisible à partir des mécanismes proposés dans la sous-section 8.3.4. En effet, le percept /pata/ dans la modalité AV et AV-pa reçoit la même excitation externe correspondant à l'activité de l'unité lexicale /pata/ multipliée par J_{AV} de /p/ mais ce n'est pas le cas pour le percept /tapa/. Dans la modalité AV le percept /tapa/ a comme entrée externe l'activité du mot /tapa/ multiplié par J_{AV} de /t/ tandis que dans la modalité AV-pa l'activité du mot /tapa/ est multiplié par la somme pondérée de J_A de /t/ et J_V de /a/ (à cause du remplacement de /t/ par /a/ dans le flux visuel), ce qui est un biais moins important que J_{AV} de /t/. Le percept /tapa/ dans la modalité AV est un concurrent avec une probabilité plus élevée que dans la modalité AV-pa. Ainsi, la valeur *delta* est plus importante dans la modalité AV-pa que dans la modalité AV.

Malgré nos hypothèses, nos résultats expérimentaux n'ont pas montré de différence significative entre ces deux modalités. Nous avons commenté cette tendance dans la sous-section 4.3.3 en notant que le geste d'ouverture de /pa/ dans la modalité AV est beaucoup plus important que celui de /ta/. Ainsi, l'effet de l'onset visuel de /pa/ au sein de la séquence /pata/ ou /tapa/ en modalité AV serait très similaire à celui de l'onset visuel de /pa/ en modalité AV-pa. Cependant, comme décrit dans le paragraphe précédent, les mécanismes de récupération de l'onset tels qu'ils sont implémentés dans ce modèle ne conduisent pas à ce résultat.

Pour pouvoir analyser cette tendance de TRACE-VT par rapport aux résultats expérimentaux, il faudra étudier de plus près l'influence des autres mécanismes que l'onset visuel sur la stabilité des percepts. Par exemple, est-ce que la présence des transformations de type lexical et phonémique influence les valeurs de *delta* ? À quel

point l'onset visuel /ta/ est perçu en modalité A, V et AV ? Ces analyses pourraient conduire à une révision dans nos hypothèses centrales ou à des révisions au niveau de l'implémentation computationnelle de TRACE-VT (par exemple sur les valeurs des biais articulatoires auditifs et visuels et sur leur fusion en modalité AV). Pour cela, des expériences plus ciblées doivent être réalisées afin de mieux comprendre les mécanismes à la base des similarités entre la modalité AV et AV-pa.

Modalité A vs. AV-ta

Concernant la différence entre les valeurs de *delta* dans la modalité auditive et la modalité AV-ta, nous pouvons faire une analyse similaire à celles présentées précédemment. Dans la modalité A, les percepts /pata/ et /tapa/ sont renforcés respectivement par J_A de /p/ et J_A de /t/. Dans la modalité AV-ta, le percept /pata/ est renforcé par un coefficient correspondant à J_A de /p/ et J_V de /a/ (car le geste de /p/ est remplacé par /a/) et l'activation de /tapa/ est multiplié par J_{AV} de /t/. Ainsi, le percept /tapa/ est relativement plus favorisé par rapport au percept /pata/ en modalité AV-ta. Cette tendance est cohérente avec nos hypothèses mais elle n'a pas été significative selon nos résultats comportementaux.

Modalité AV-pa vs. AV-ta

Comme décrit dans les paragraphes précédentes, le biais articulatoire utilisé pour le percept /pata/ en modalité AV-pa correspond à J_{AV} de /p/ et celui utilisé pour le percept /tapa/ correspond à J_A de /t/ et J_V de /a/. Les valeurs utilisées dans ces simulations pour la matrice J et les coefficients a et b conduisent donc à un renforcement important de /pata/ en modalité AV-pa par rapport à /tapa/. En ce qui concerne la modalité AV-ta, le percept /pata/ est renforcé par J_A de /p/ et J_V de /a/ et le percept /tapa/ par J_{AV} de /t/. D'une manière inverse à la modalité AV-pa, la stabilité du percept /tapa/ sera donc plus importante que celle du percept /pata/ dans cette modalité.

Il est important de noter que l'effet de l'onset visuel est plus important en modalité AV-pa qu'en modalité AV-ta. Ce résultat suggère que l'hypothèse sur le liage en fonction du degré d'ouverture de la mâchoire pourrait expliquer cette tendance que nous avions également observée dans nos expériences comportementales. Des expériences plus précises sont nécessaires pour étudier la capacité perceptive à récupérer le degré d'ouverture de la mâchoire à partir des signaux de parole auditif, visuel et audio-visuel (voir sous-section 9.2.1).

8.5.5 Percepts instables

Bien que la probabilité des percepts mono-syllabiques et des percepts de type /VCVC/ soit très faible, ces percepts peuvent émerger de temps en temps dans nos simulations. Cependant, leur durée de stabilité est souvent très faible. Dans les figures présentées à la fin de cette section, nous avons exclu les transformations qui ne durent qu'un seul cycle de traitement. Nous présentons le nombre moyen de ces percepts instables (durée de stabilité d'un cycle) et les écart-types correspondants

dans le tableau 8.5. La présence de ces transformations ne semble pas être cohérente avec celles perçues par les sujets dans nos expériences comportementales. En effet, même si les sujets percevaient ces bascules extrêmement rapides pendant les expériences, ils ne pouvaient pas les signaler faut de temps entre les deux bascules consécutives. Cependant, dans la phase de débriefing, les sujets ne nous avons jamais signalé la perception de ces transformations rapides. Les mécanismes prévus dans TRACE-VT ne permettent pas d'éliminer ces percepts instables.

8.5.6 Différents types de transformation

Les transformations étudiées dans ce travail étant de type resegmentation, les transformations « autres » illustrés sur les figures à la fin de ce chapitre ne correspondent pas aux transformations classées en « autres » dans nos expériences comportementales. Certaines de ces transformations « autres » peuvent être facilement ajoutées à ce modèle. Par exemple, le mécanisme d'adaptation peut entraîner des transformations correspondant à des substitutions de phonèmes de la même manière que ce que MacKay *et al.* (1993) proposent dans le cadre du modèle NST (voir sous-section 7.3.1). En revanche, les transformations de type lexical ou de streaming ne peuvent émerger dans TRACE-VT. Afin de rendre compte des transformations de type streaming, il serait nécessaire d'ajouter des mécanismes de streaming auditif. Pour cela, l'entrée sous format textuel du modèle doit être remplacée par le signal auditif et les techniques du traitement du signal doivent s'intégrer au modèle. Ainsi, les traits phonétiques et les biais articulatoires seront directement calculés à partir des propriétés acoustiques du stimulus.

En ce qui concerne les transformations de type lexical, il nous semble que les mécanismes utilisés dans TRACE ne permettent pas à ce type de transformations d'émerger. En effet, ces percepts n'étant pas cohérents avec l'entrée du modèle, les unités lexicales dans TRACE représentant ces percepts ne peuvent jamais être activées. Ceci conduit à une absence d'excitation du niveau lexical vers le niveau de percept. Ainsi, la probabilité d'activation des transformations de type lexical est très faible même en présence de l'adaptation. La perception de ce type de transformations suggère qu'une fois entrés dans le processus de transformation verbale, les sujets peuvent percevoir des mots qui n'ont pas d'appuis sensoriels. Un modèle basé uniquement sur la reconnaissance de l'entrée ne peut donc pas rendre compte de ce type de transformation, et d'autres niveaux incorporant des mécanismes de proximité et de priming sémantiques sont nécessaires ici.

8.6 Proposition neuro-anatomique

Bien que l'attribution neuro-anatomique de différentes composantes de TRACE-VT soit au-delà de cette thèse, nos données iEEG présentées dans la section 5 et celles de la littérature (Sato *et al.*, 2004; Kondo et Kashino, 2007) permettent de suggérer quelques éléments de réponse. Il est à préciser que nous ne défendons pas la plausibilité de l'architecture de ce modèle telle qu'elle est proposée ici mais nous

essayons de placer les fonctionnalités proposées dans ce travail par rapport aux données neuro-anatomiques.

Les mécanismes présents dans TRACE-VT utiliseraient les deux circuits ventral et dorsal du traitement de la parole proposé par Hickok et Poeppel (voir sous-section 2.2.2) : l'aspect de la reconnaissance de la parole réalisé par TRACE serait cohérent avec le circuit ventral et les biais articulatoires de TRACE-VT et son influence sur les sorties du niveau lexical de TRACE pourraient exister sous-forme d'implication du circuit dorsal. Ainsi, un scénario possible serait le suivant : les unités représentant les percepts siègeraient dans le gyrus temporal supérieur. La matrice J correspondant aux biais articulatoires pourrait être représentée dans le gyrus frontal inférieur gauche et pourrait s'activer par une interaction feedforward entre le gyrus temporal supérieur, le gyrus supramarginal gauche et le gyrus frontal inférieur gauche. L'influence des biais articulatoires sur les représentations de type lexical pourrait exister sous forme de connexions de type feedback à partir du gyrus frontal inférieur gauche vers le gyrus temporal supérieur gauche. Il est important de noter que ce mécanisme de feedback n'a jamais été observé dans les expériences sur l'effet de transformation verbale mais il serait cohérent avec des études sur la multistabilité visuelle sur le rôle des aires frontales dans la multistabilité perceptive (Sterzer *et al.*, 2009, voir sous-section 3.1.2). La question des liens feedforward et feedback entre les aires temporales, pariétales et frontales dans la perception de la parole fait partie d'un projet de recherche proposé par Marc Sato au Gipsa-lab.

La proposition présentée ci-dessus semble être cohérente avec le modèle de Skipper et collègues (Skipper *et al.*, 2005; van Wassenhove *et al.*, 2005; Skipper *et al.*, 2007). Comme décrit dans la sous-section 2.3.5, selon ces auteurs, la projection feedback vers les aires temporales sert de « prédiction » pendant la perception audio-visuelle de la parole. Cependant, Schwartz *et al.* (2008a) remarquent que la prédiction basée sur les connaissances motrices ne semble pas être nécessaire d'une manière systématique dans la perception de la parole. Les auteurs suggèrent que ce feedback moteur pourrait jouer un rôle dans des conditions adverses telles qu'en présence du bruit afin d'intégrer les événements sensoriels les uns avec les autres d'une façon cohérente, autrement dit, pour une meilleure organisation de la scène de parole. Ainsi, le scénario que nous proposons dans cette section concernant les projections feedforward et feedback peut être compris dans le cadre du modèle de Skipper et collègues mais en ajoutant à ce modèle un contenu « liage ».

8.7 Conclusion

Le travail présenté dans ce chapitre est un travail préliminaire pour modéliser la multistabilité perceptive en parole. Cette modélisation nous a permis de rassembler nos hypothèses sur les mécanismes sous-jacents à l'effet de transformation verbale et, plus généralement, ceux sur le liage perceptif en parole. Le travail d'analyse préalable à l'implémentation de ces mécanismes nous a permis de rendre ces hypothèses explicites et, en conséquence, vérifiables par des expériences. Les résultats de nos simulations ont mis en évidence comment ces mécanismes interagissent et ont permis

de reproduire, dans une certaine mesure, nos résultats expérimentaux notamment sur le rôle de l'onset visuel.

Les perspectives de ce travail sont essentiellement de quatre ordres. Le premier enjeu est celui des données expérimentales qui permettront d'étudier plus en détail nos hypothèses théoriques et de vérifier celles utilisées au niveau de l'implémentation computationnelle (hypothèses périphériques). Des études expérimentales sur l'effet du geste de la mâchoire et sur la nature de la fenêtre de liage/décision devront notamment être réalisées. Dans ce sens, quelques idées d'expériences seront présentées dans le chapitre 9. D'autre part, des améliorations computationnelles devront être apportées au niveau des percepts dans TRACE-VT. En effet, malgré l'utilisation d'une architecture cohérente avec celle d'un réseau neuromimétique d'attracteurs au niveau des percepts, la dynamique de ce réseau n'est pas prise en compte. Pour améliorer ce point, il serait nécessaire de réviser le lien entre le modèle TRACE et le niveau des percepts et d'étudier la dynamique de ce réseau. De plus, des analyses computationnelles devront être effectuées au niveau des percepts instables. Une perspective importante de ce travail est celui des mécanismes neuronaux et cérébraux de l'effet de transformation verbale. Il serait par exemple intéressant de vérifier si les percepts peuvent exister sous forme des attracteurs d'un réseau de neurones comme supposé dans le niveau de percepts de TRACE-VT. Comme décrit dans la section 8.6, nous envisageons également de préciser l'attribution neuro-anatomique des différentes composantes de TRACE-VT en relation avec les données de neuroimagerie. Enfin, du point de vue théorique, il faudrait analyser le lien entre le phénomène de l'effet de transformation verbale et les mécanismes d'accès au lexique et de reconnaissance de la parole.

Résultats des simulations : TRACE-VT

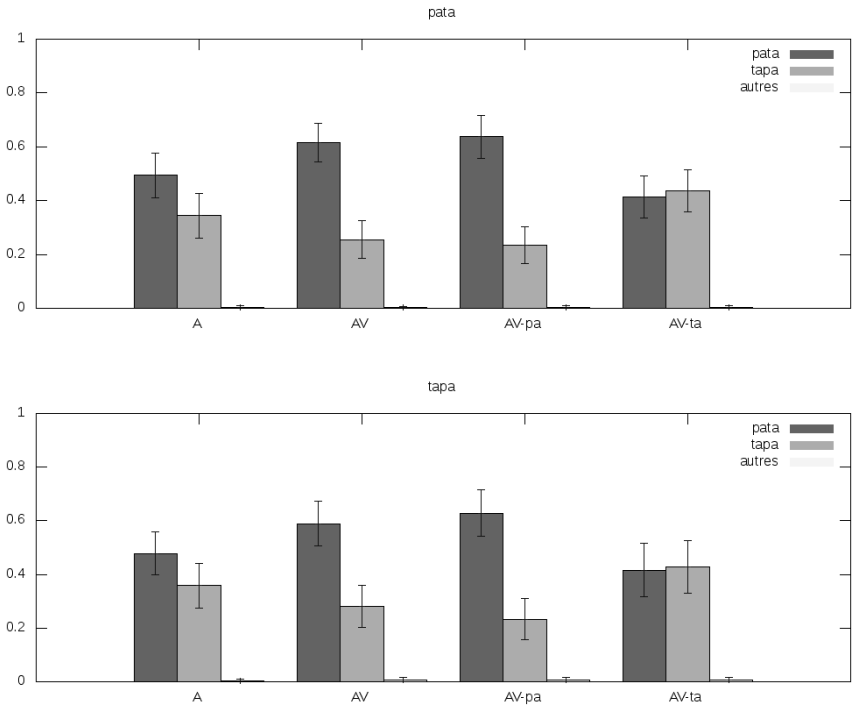


FIGURE 8.10 : Simulation de TRACE-VT : durée moyenne de stabilité des percepts. Entrée auditive : 150 répétitions de séquence /pata/ (en haut) et /tapa/ (en bas). Entrée visuelle : sans entrée (A), 150 répétitions de la séquence auditive (modalité AV), de la séquence /paaa/ (modalité AV-pa) ou de la séquence /taaa/ (modalité AV-ta). Nombre de simulations : 20. Les percepts « autres » peuvent être les séquences /atap/, /apat/, /pa/ et /ta/. Les barres d'erreur représentent les écart-types des moyennes.

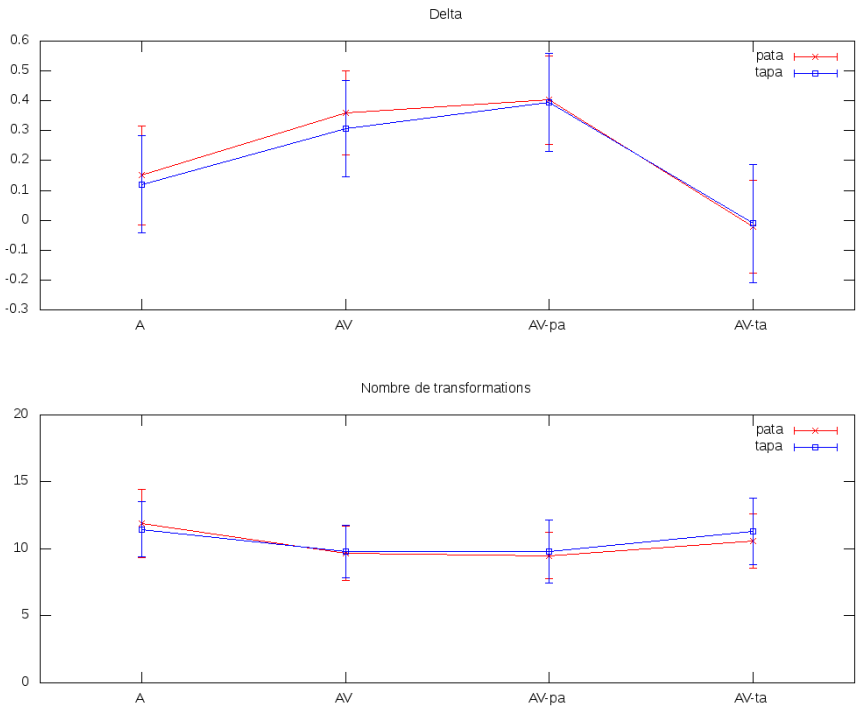


FIGURE 8.11 : Simulation de TRACE-VT : moyenne des valeurs de *delta* et du nombre de transformations correspondant aux simulations présentées sur la figure 8.10. Les barres d'erreur représentent les écart-types des moyennes.

TABLE 8.5 : Simulation de TRACE-VT : moyennes du nombre de transformations instables qui durent un seul cycle ± les écart-types des moyennes pour les différentes modalités de présentation (A, AV, AV-pa et AV-ta) et les deux séquences auditives /pata/ (en haut) et /tapa/ (en bas). Les moyennes nulles (avec évidemment écart-types égaux à zéro) ne sont pas présentées.

	/pata/	/tapa/	/atap/	/apat/	/pa/	/ta/
A	0.15±0.37	0.35±0.59			2.70±0.80	2.65±0.81
AV		0.35±0.59			2.20±1.00	1.80±0.77
AV-pa		0.60±0.69			2.00±1.02	1.90±0.85
AV-ta	0.25±0.55	0.15±0.37			2.70±0.98	2.25±1.07

	/pata/	/tapa/	/atap/	/apat/	/pa/	/ta/
A	0.20±0.41	0.35±0.67			2.85±0.93	2.45±0.76
AV	0.10±0.31	0.40±0.50			2.25±0.79	1.55±0.82
AV-pa	0.05±0.22	0.70±0.98			2.40±1.09	1.70±0.66
AV-ta	0.15±0.37	0.15±0.37			2.85±0.87	2.35±0.87

Résultats des simulations : TRACE-VT avec l'adaptation mais sans biais articulatoire

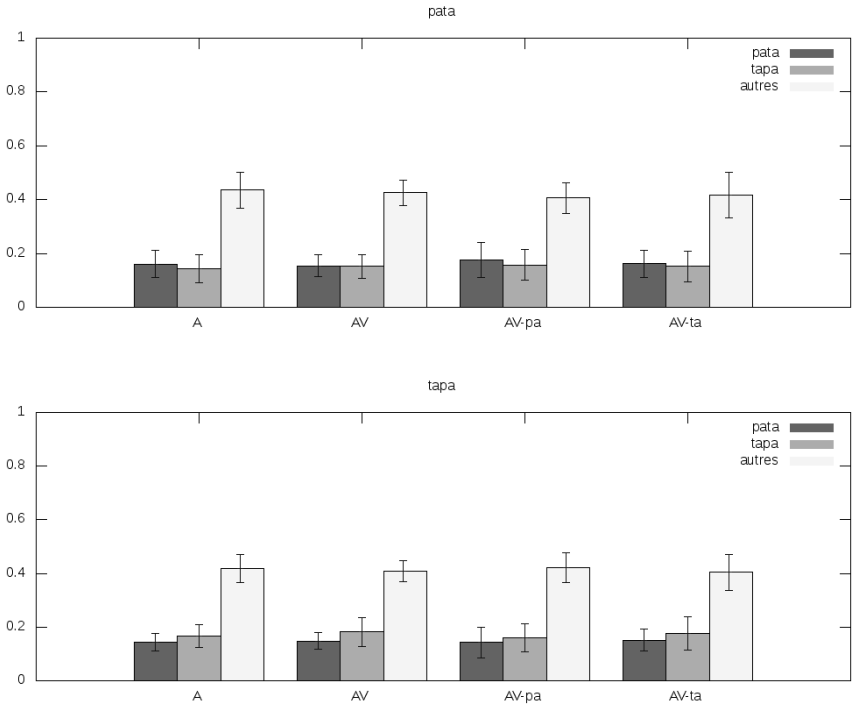


FIGURE 8.12 : Simulation de TRACE-VT avec l'adaptation mais sans biais articulatoire : durée moyenne de stabilité des percepts. Entrée auditive : 150 répétitions de séquence /pata/ (en haut) ou /tapa/ (en bas). Entrée visuelle : sans entrée (A), 150 répétitions de la séquence auditive (modalité AV), de la séquence /paaa/ (modalité AV-pa) ou de la séquence /taaa/ (modalité AV-ta). Nombre de simulations : 20. Les percepts « autres » peuvent être les séquences /atap/, /apat/, /pa/ et /ta/. Les barres d'erreur représentent les écart-types des moyennes.

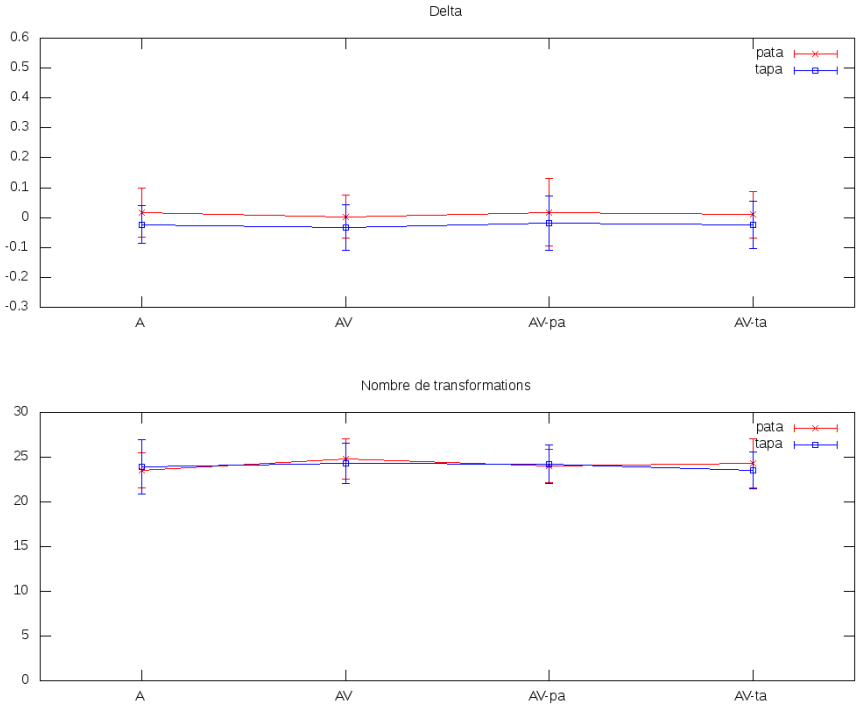


FIGURE 8.13 : Simulation de TRACE-VT avec l'adaptation mais sans biais articulatoire : moyenne des valeurs de *delta* et du nombre de transformations correspondant aux simulations présentées sur la figure 8.12. Les barres d'erreur représentent les écart-types des moyennes.

Résultats des simulations : TRACE-VT avec les biais articulatoires mais sans adaptation

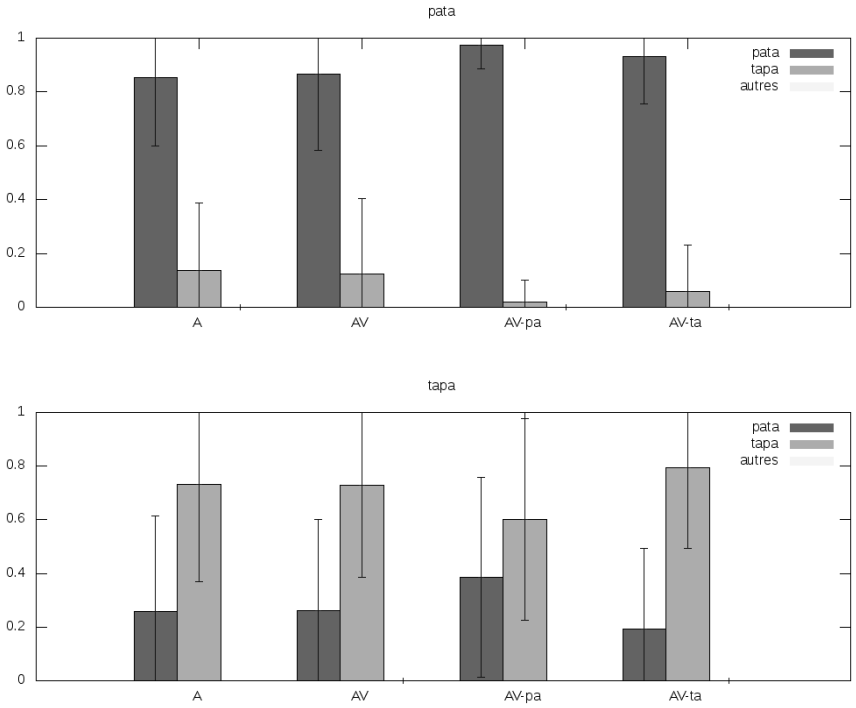


FIGURE 8.14 : Simulation de TRACE-VT avec les biais articulatoires mais sans adaptation : durée moyenne de stabilité des percepts. Entrée auditive : 150 répétitions de séquence /pata/ (en haut) ou /tapa/ (en bas). Entrée visuelle : sans entrée (A), 150 répétitions de la séquence auditive (modalité AV), de la séquence /paaa/ (modalité AV-pa) ou de la séquence /taaa/ (modalité AV-ta). Nombre de simulations : 20. Les percepts « autres » peuvent être les séquences /atap/, /apat/, /pa/ et /ta/. Les barres d'erreur représentent les écart-types des moyennes.

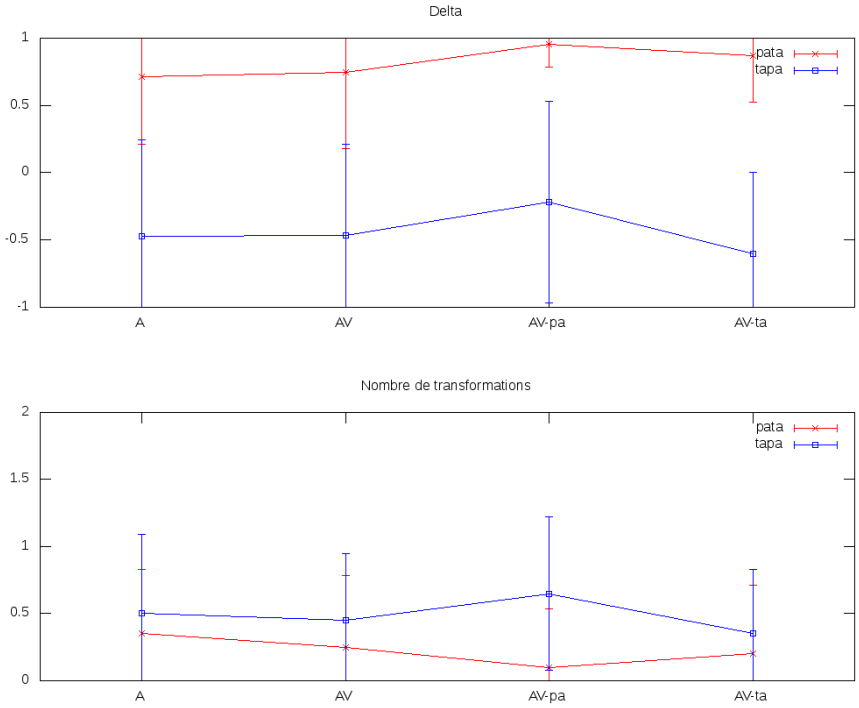


FIGURE 8.15 : Simulation de TRACE-VT avec les biais articulatoires mais sans adaptation : moyenne des valeurs de *delta* et du nombre de transformations correspondant aux simulations présentées sur la figure 8.14. Les barres d'erreur représentent les écart-types des moyennes.

Résultats des simulations : TRACE-VT sans biais articulatoire ni adaptation

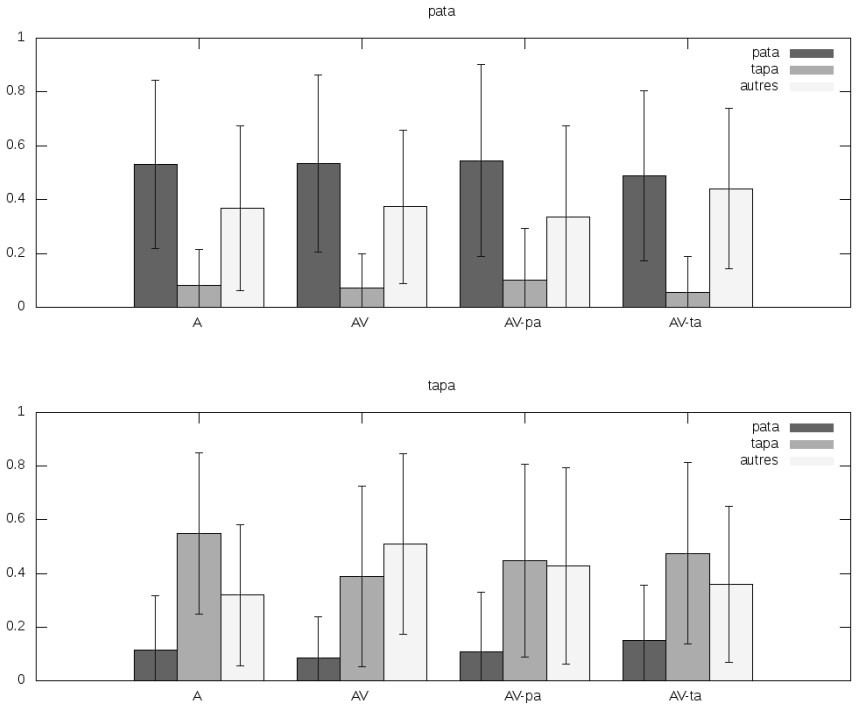


FIGURE 8.16 : Simulation de TRACE-VT sans biais articulatoire ni adaptation : durée moyenne de stabilité des percepts. Entrée auditive : 150 répétitions de séquence /pata/ (en haut) ou /tapa/ (en bas). Entrée visuelle : sans entrée (A), 150 répétitions de la séquence auditive (modalité AV), de la séquence /paaa/ (modalité AV-pa) ou de la séquence /taaa/ (modalité AV-ta). Nombre de simulations : 20. Les percepts « autres » peuvent être les séquences /atap/, /apat/, /pa/ et /ta/. Les barres d'erreur représentent les écart-types des moyennes.

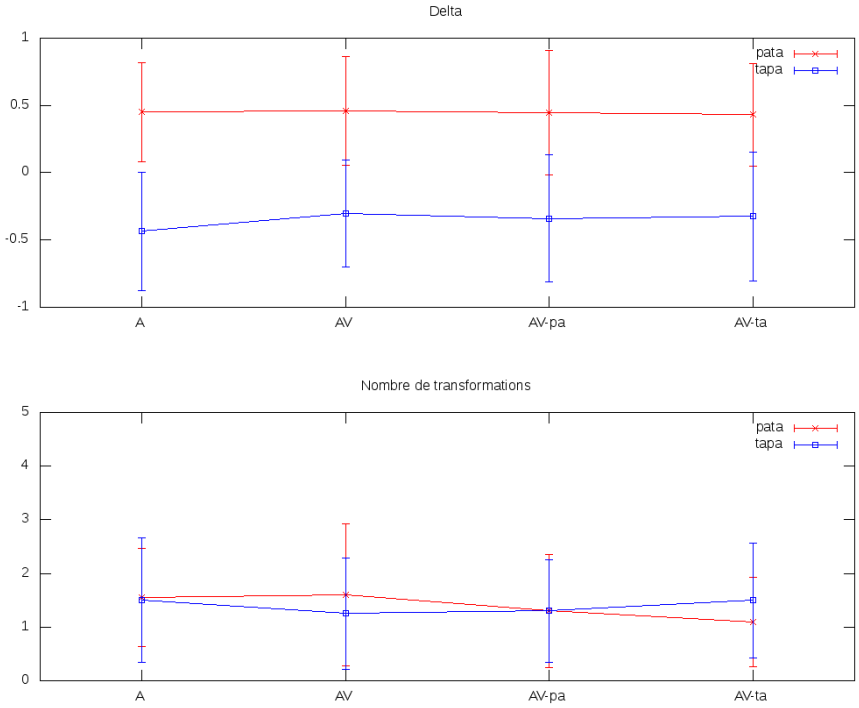


FIGURE 8.17 : Simulation de TRACE-VT sans biais articulatoire ni adaptation : moyenne des valeurs de *delta* et du nombre de transformations correspondant aux simulations présentées sur la figure 8.16. Les barres d'erreur représentent les écart-types des moyennes.

Quatrième partie

Conclusion générale

Chapitre 9

Vers une analyse perceptuo-motrice et multimodale de la scène de parole

Sommaire

9.1 Résumé des résultats

L'objectif de cette thèse était de mieux comprendre l'organisation perceptive de la scène de parole. Puisque la question de l'organisation perceptive d'une scène sensorielle n'est pas indépendante de la nature des objets la constituant, cette thèse portait également sur la nature des objets parole et sur les mécanismes de construction de ces objets dans le cerveau. Le paradigme expérimental utilisé dans cette thèse était la multistablité perceptive en parole appelé l'effet de transformation verbale. Ce paradigme permet d'étudier « en ligne » les mécanismes à la base de l'émergence et de la stabilisation des percepts phonétiques. Les études sur cette problématique ont été effectuées sous trois angles : l'expérimentation comportementale, l'expérimentation neurophysiologique et la modélisation cognitive computationnelle. Cette section présente un résumé des principaux résultats de ces études.

9.1.1 Organisation multimodale de la scène de parole

Nous ne percevons pas la parole uniquement à partir du signal sonore : les informations visuelles sur les gestes articulatoires du locuteur sont également utilisées pendant la perception de la parole. Les travaux comportementaux réalisés dans le

cadre de cette thèse avaient pour objectif d'étudier les mécanismes audio-visuels susceptibles de mettre en forme l'objet parole. Pour cela, nous avons effectué quatre expériences dont les résultats sont résumés ci-dessous.

Notre première expérience comportementale a mis en évidence, pour la première fois, que les informations visuelles influencent les transformations verbales. En effet, un flux visuel incongruent avec le flux auditif peut diminuer la durée de stabilité du percept cohérent avec le flux auditif, ceci en augmentant la durée de stabilité du percept cohérent avec les informations visuelles. À travers notre deuxième expérience, nous avons pu observé que la modalité visuelle joue un rôle « actif » dans l'émergence et la stabilisation des percepts phonétiques : les informations visuelles peuvent, dans une certaine mesure, contrôler l'émergence des transformations verbales. De plus, l'existence des bascules d'une séquence à l'autre dans le flux visuel renforce l'effet observé dans la première expérience. Nos deux dernières expériences comportementales avaient pour objectif d'investir les mécanismes à la base de l'influence de la modalité visuelle. Nous avons pu montrer que des informations visuelles sur le geste d'ouverture de la mâchoire correspondant ont la capacité de guider le liage perceptif audio-visuel (effet de l'onset visuel). Cette capacité semble être plus efficace lors d'une stimulation par un flux visuel contenant les gestes articulatoires du locuteur que pendant la présentation de formes géométriques simulant ces gestes.

Ces résultats suggèrent donc que l'organisation perceptive de la scène de parole est multimodale. Les informations qui guident cette organisation et qui sont impliquées dans la construction perceptive des objets parole ne sont pas des informations quelconques mais celles associées aux gestes articulatoires du locuteur. En lien avec la littérature sur l'effet de transformation verbale et face à ces résultat, il nous a paru vraisemblable que les représentations articulatoires interviennent dans l'organisation perceptive de la parole et dans l'émergence des nouveaux percepts phonétiques. Cette hypothèse a fait objet de notre étude sur le circuit cortical des transformations verbales que nous résumons dans la section suivante.

9.1.2 Implication du circuit dorsal

La méthode utilisée dans cette étude est celle de l'EEG intracrânienne. Nous avons confronté deux conditions conduisant toutes les deux aux changements perceptifs chez le sujet. La première condition consistait en la détection des changements « réels » dans le stimulus. La deuxième condition était une tâche de transformation verbale. Nous avons observé un couplage entre le gyrus supramarginal gauche et le gyrus frontal inférieur gauche en lien avec l'émergence des transformations verbales mais pas lors des bascules perceptives produites par des changements sensoriels. Du point de vue neuro-anatomique, ce résultat est en accord avec ceux obtenus par des études IRMf sur l'effet de transformation verbale (Sato *et al.*, 2004; Kondo et Kashino, 2007). Du point de vue de la dynamique des bascules, le résultat de notre étude ajoute un élément important à nos connaissances sur les mécanismes cérébraux de la perception multistable de la parole : ayant une résolution temporelle élevée, la méthode iEEG nous a permis de préciser l'activité de différentes zones cérébrales en lien avec la phase d'émergence des percepts phonétiques.

Le couplage pariéto-frontal observé dans cette étude est cohérent avec le circuit dorsal de la perception de la parole proposé par Hickok et Poeppel (2007, 2004). Ce résultat suggère un lien fort entre la perception et la production de la parole dans l'effet de transformation verbale. Ainsi, il confirmerait notre hypothèse issue de nos expériences comportementales sur les liens perceptuo-moteurs dans l'organisation perceptive de la scène de parole. Nous rappelons que les stimuli utilisés dans notre étude iEEG étaient en modalité auditive. Ceci élargit le champ d'application de notre proposition sur les liens perceptuo-moteurs que nous avons décrite en se basant sur nos résultats comportementaux : les liens perceptuo-moteurs ne sont pas seulement impliqués lors d'une présentation audio-visuelle de la parole mais ils joueraient également un rôle lors de la perception purement auditive de la parole.

9.1.3 Implémentation computationnelle

Nos travaux expérimentaux nous ont fourni certaines hypothèses sur l'organisation perceptive de la scène de parole. La dernière étape de cette thèse a été consacrée à un travail préliminaire de modélisation cognitive computationnelle basée sur ces hypothèses. L'intérêt de cette étape était double. Premièrement, la phase préalable d'analyse nous a permis de préciser d'une manière plus détaillée nos propositions et nos hypothèses en vue d'une implémentation computationnelle. Deuxièmement, cette implémentation a mis en évidence l'apport des mécanismes proposés et leurs interactions. Afin de ne pas rester à un niveau très abstrait par rapport au domaine de la perception de la parole, nous avons choisi d'intégrer nos propositions dans le cadre d'un modèle psycholinguistique. Pour cela, nous avons choisi le modèle TRACE (McClelland et Elman, 1986).

Le modèle TRACE, comme les autres modèles psycholinguistiques de la perception de la parole, ne peut pas rendre compte de l'effet de transformation verbale. Ainsi, la première phase de ce travail était d'implémenter le minimum de mécanismes nécessaires pour produire l'effet de transformation verbale dans TRACE. Pour cela, nous avons notamment ajouté des mécanismes d'adaptation des représentations perceptives. La deuxième phase consistait en l'intégration de nos propositions spécifiques sur les liens perceptuo-moteurs impliqués dans l'effet de transformation verbale. Des simulations de ce modèle baptisé TRACE-VT (pour *Verbal Transformation*) ont montré que les mécanismes d'adaptation, seuls, ne peuvent pas expliquer nos résultats comportementaux. C'est en ajoutant des mécanismes perceptuo-moteurs que TRACE-VT peut répliquer les tendances observées dans nos expériences comportementales notamment sur l'effet de l'onset visuel.

9.2 Perspectives de recherche

Les travaux effectués dans le cadre de cette thèse ouvrent la voie dans trois directions, expériences comportementales, études neurophysiologiques et modélisation computationnelle, que nous présentons dans les sous-sections suivantes.

9.2.1 Expériences comportementales

Deux interprétations que nous avons faites à partir des résultats de nos expériences comportementales méritent d'être étudiées plus en détail : l'effet du geste d'ouverture de la mâchoire et la nature spécifique à la parole de l'effet de l'onset visuel. Les paragraphes suivants présentent quelques idées d'expériences qui fourniraient des éléments de réponses à nos interrogations sur ces sujets. Les derniers paragraphes ouvrent des perspectives plus larges concernant l'effet de l'onset visuel.

Estimation perceptive du geste d'ouverture de la mâchoire

Rappelons que nous avons défendu l'idée selon laquelle l'organisation perceptive du flux de parole est basée sur la cohésion articulatoire, notamment lorsqu'il s'agit du flux audio-visuel de la parole pour lequel les gestes articulatoires sont visibles. Ces propositions impliquent l'hypothèse préalable que les sujets sont capables de déterminer à partir d'un flux auditif, visuel et audio-visuel de la parole la dynamique d'ouverture de la mâchoire. À notre connaissance, cette hypothèse n'a jamais été vérifiée dans la littérature. Voici une expérience perceptive simple qui permettrait de vérifier cette hypothèse. Les séquences mono-syllabiques (par exemple, /pa/ et /ta/) et dissyllabiques (par exemple, /pata/ et /tapa/) en différentes modalités et avec différents degrés d'ouverture seraient présentées au sujets. La tâche consisterait à signaler le degré d'ouverture des syllabes les unes par rapport aux autres. Ce type d'expérience existe dans la littérature en ce qui concerne les voyelles (par exemple Lindblom et Lubker, 1985; Lalain *et al.*, 2008). En fonction des résultats de cette expérience, des tâches sur l'effet de transformation verbale seraient envisagées afin de vérifier l'influence du degré perceptif d'ouverture de la mâchoire sur les transformations verbales.

Il est intéressant de noter ici que nous avons effectué dans le cadre d'un stage de master une expérience perceptive sur l'influence de l'accentuation des syllabes sur les transformations verbales (Basirat, 2005). Notre hypothèse était que les séquences /'pata/ et /ta'pa/ entraîneraient la préférence pour le percept /pata/ et, au contraire, les séquences /pa'ta/ et /'tapa/ conduiraient à la perception plus stable de /tapa/. Les résultats observés étaient en accord avec cette hypothèse. Il est donc légitime de vérifier, d'une manière plus systématique, l'influence du degré d'ouverture de la mâchoire sur la stabilité des percepts phonétiques à travers l'effet de transformation verbale.

Liens perception-production impliqués dans l'effet de l'onset visuel

Concernant les liens perception-production et leurs rôles dans l'organisation perceptive de la parole, il serait également intéressant de vérifier si un effet similaire à l'onset visuel pourrait exister lorsque les sujets réalisent une tâche perceptive de transformation verbale en produisant en même temps des séquences de parole. Si l'effet de l'onset visuel a des origines articulatoires, il ne serait pas surprenant que nous rencontrions ce type d'effet. Pour vérifier cette hypothèse, nous avons réalisé une étude pilote pendant cette thèse. La tâche consistait en l'écoute des séquences

/pata/ et /tapa/ en boucle pendant que les sujets répétaient les séquences /pata/, /tapa/, /pa#a/[1] et /ta#a/ d'une manière synchrone avec /pa/ et /ta/ dans les stimuli. Ces conditions sont respectivement comparables avec les modalités AV, AV-pa et AV-ta dans les expériences présentées dans les sections 4.3 et 4.4. Nous avions la même hypothèse que dans nos expériences audio-visuelles : lorsque les sujets répétaient la séquence /pa#a/, la valeur de *delta* devrait être plus élevée que lors de la répétition de la séquence /ta#a/. Cette étude pilote n'a pas abouti en une expérience car la tâche paraît difficile aux sujets et ils ne produisaient souvent pas des séquences synchrones avec les stimuli.

Cette tentative suggère que la vérification de cette idée expérimentale nécessiterait d'autres types d'analyses que celles dépendant de la production synchrone des sujets avec le signal auditif. Dans ce sens, une piste intéressante pourrait être l'analyse de la production désynchronisée des sujets. Ainsi, la tâche demandée aux sujets serait plus simple. On pourrait par exemple enregistrer des gestes articulatoires des sujets par la méthode électromyographique (EMG) afin de pouvoir analyser la corrélation entre la production et le stimulus auditif. Nous pourrions ensuite analyser les transformations verbales signalées par rapport à ce pattern. Deux possibilités devraient être étudiées : est-ce que la production de /pa#a/ et /ta#a/, synchrone ou non avec le stimulus, pourrait conduire à une préférence respectivement pour /pata/ et /tapa/ ? Est-ce que cette préférence est plus importante lors d'une production synchrone que pendant une production non synchrone ? Nous pourrions enfin basculer sur des expériences de perturbation motrice, de même type que celles décrites dans les tâches de transformation verbale (voir sous-section 3.3.2, Reisberg *et al.*, 1989) en déterminant si une activité motrice perturbatrice (telle que mâcher du chewing-gum) pourrait bloquer le mécanisme de détection d'onset visuel, et éliminer pour tout ou partie le biais en faveur de séquences LC en modalité auditive ou audio-visuelle.

Sa spécificité « parole » de l'effet d'onset visuel

Quant à la nature de l'onset visuel, nous défendons l'idée selon laquelle l'effet de l'onset visuel est un effet spécifique à la parole : ce seraient les informations fournies par les gestes articulatoires du locuteur qui guideraient l'organisation perceptive de la scène de parole. De ce point de vue, l'effet similaire mais moins fort de l'onset visuel induit par des barres présentées dans la section 4.4 a été expliqué en invoquant la capacité de ces stimuli à simuler les mouvements des lèvres. Il est intéressant de citer ici une étude réalisée par Devergie *et al.* (2008) sur le liage audio-visuel. Cette étude montre que certains stimuli visuels sont plus efficaces que d'autres pour l'extraction des éléments qui composent une séquence auditive (extraction de 3 voyelles pendant la répétition des voyelles /o u y a i e/). Les stimuli visuels consistaient en des formes géométriques dont le contraste ou le mouvement changeaient d'une manière synchrone avec la présentation auditive des 3 voyelles cibles parmi les six. Le résultat de cette expérience montre que seuls les stimuli incorporant des mouvements améliorent le score des sujets. Ceci, renforcerait notre hypothèse que l'effet

[1] # représente le prolongement de la voyelle /a/.

d'onset visuel n'est pas seulement basé sur la synchronisation des indices visuels avec le signal auditif mais que la nature de ces indices sont également importants. Pour pouvoir vérifier cette suggestion, il serait intéressant de mettre en place des expériences utilisant des formes géométriques plus ou moins similaires au patron d'ouverture des lèvres comme stimuli visuel, et notamment de remplacer des barres verticales par des barres horizontales.

Rôle des mécanismes de liage et de l'effet de l'onset visuel dans l'accès au lexique

Une question à laquelle nous nous sommes confrontés notamment dans la partie de modélisation de cette thèse est celle du lien entre les mécanismes de liage en parole et les processus à la base de l'accès au lexique. À notre connaissance, cette question n'a jamais été étudiée dans la littérature. Un cadre général de la perception de la parole doit comprendre, entre autres, des processus de liage perceptif relativement de bas niveau et des processus d'accès au lexique (impliquant une fusion multimodale) de plus haut niveau. C'est pour cela qu'il nous semble important d'étudier les liens entres ces mécanismes.

Les travaux qui se rapprochent probablement le plus de cette question sont les rares études sur le rôle des informations visuelles sur les gestes articulatoires du locuteur dans l'accès au lexique. Dans une expérience de type McGurk, Brancazio (2004) a observé que le statut lexical des stimuli (mot ou non-mot) influence l'effet McGurk : les réponses cohérentes avec le flux visuel étaient plus fréquentes lorsque le stimulus visuel était un mot et le stimulus auditif était un non-mot. Dans une tâche de détection du phonème en présence du bruit, Fort *et al.* (2010) ont observé que les informations visuelles améliorent le score des sujets d'une façon plus importante lorsque la cible était au sein d'un mot que lors de la présentation d'un non-mot. En lien avec ces résultats et avec certaines propositions psycholinguistiques (voir section 4.5), il serait intéressant de vérifier l'éventuel rôle de l'onset visuel dans l'accès au lexique. Ainsi, nous pourrions également vérifier la part de la fusion audio-visuelle et des mécanismes plus précoces de liage dans l'accès au lexique, ce qui n'a pas été démontré par les études citées ci-dessus.

9.2.2 Expériences neurophysiologiques

L'effet de transformation verbale est un phénomène dynamique qui est basé sur différentes aires cérébrales. Bien que notre étude iEEG présente un pas significatif dans la compréhension de l'émergence des transformations verbale, il reste muet sur la dynamique des transformations. Il nous semble qu'une perspective intéressante de ce travail serait l'étude de l'effet de transformation verbale dans le temps.

Décours temporel des transformations verbales

Pour pouvoir étudier le rôle fonctionnel des régions observées dans notre expérience, il serait nécessaire d'étudier le décours temporel des transformations. Autrement dit, il faudrait suivre l'activité de ces régions les unes par rapports aux autres.

Grâce à leur résolution temporelle et spatiale, il nous semble que les méthodes d'enregistrement iEEG et MEG sont des méthodes pertinentes pour des études sur le décours temporel (notre expérience d'iEEG décrite dans le chapitre 5 portait malheureusement sur trop peu de sujets pour permettre une telle analyse). Les méthodes d'analyse pourraient être similaires à celles utilisées au département Image et Signal du Gipsa-lab (par exemple, Achard et Bullmore, 2007; Amini *et al.*, 2009). Ainsi, nous pouvons déterminer le circuit cérébral et sa connectivité fonctionnelle en lien avec les transformations verbales. Le scénario présenté dans la section 8.6 sur l'éventuel retour du gyrus frontal inférieur vers le gyrus temporal supérieur serait également vérifiable dans ce contexte.

Transformations verbales audio-visuelles

L'intérêt des études neurophysiologiques sur l'effet de transformation verbale avec les stimuli audio-visuels est double. D'une part, ces études nous permettraient de vérifier si la modalité audio-visuelle, comme proposé dans cette thèse et en accord avec la littérature, entraînent des activités plus importantes qu'en modalité auditive dans les aires associées aux représentations articulatoires. D'autre part, des stimuli audio-visuels utilisés dans cette thèse fourniraient un matériel intéressant pour l'étude du décours temporel des transformations verbales. En effet, nous avons vu dans la section 4.2 que les bascules visuelles présentes dans la vidéo peuvent induire des tansformations verbales chez les sujets. Utilisant ces stimuli, nous pouvons estimer certaines transformations perçues par les sujets sans qu'ils les signalent par des tâches motrices comme l'appui sur un bouton. Ceci faciliterait l'analyse du décours temporel et de la connectivité cérébrale en lien avec les transformations verbales. Un projet a été aussi élaboré en collaboration avec Jean-Philippe Lachaux et Philippe Kahane, il reste hélas en attente de sujet pour passer l'expérience.

9.2.3 Perspectives computationnelles

Développement de TRACE-VT

Plusieurs perspectives sont envisageables afin d'améliorer TRACE-VT. En ce qui concerne les simulations, il faudrait réaliser des simulations des autres études que nous avons effectuées sur l'effet de transformation verbale. Par exemple, les entrées de type /psə/ et /səp/ devraient être introduites dans le modèle. Ainsi, nous pourrions étudier la capacité de TRACE-VT à rendre compte des deux principaux effets que nous avons observés dans l'expérience 1 et 2 présentées respectivement dans les sections 4.1 et 4.2, à savoir l'effet du flux visuel incongruent sur la stabilité des percepts et l'effet des bascules dans le flux visuel sur l'émergence et la stabilité des transformations. De la même façon, des simulations sur l'effet de la focalisation prosodique décrit dans la section 9.2.1 devraient également être envisagées.

Outre les travaux au niveau des simulations, de nouveaux mécanismes pourraient être intégrés à TRACE-VT. Il serait par exemple très utile d'ajouter un niveau de traitement du signal auditif et visuel à l'entrée du modèle. L'avantage de cette intégration serait double. D'une part, elle permettrait de rendre compte des trans-

formations de type streaming. D'autre part, ainsi, les biais articulatoires pourraient être estimés directement en fonction des signaux d'entrée, ce qui est notamment intéressant dans le cas de la focalisation prosodique.

Un autre élément qui pourrait être intégré à TRACE-VT concerne les différences individuelles observées lors des expériences sur l'effet de transformation verbale : par exemple, certains sujets perçoivent un nombre plus important de transformations que d'autres ou certains sont plus sensibles à la modalité visuelle que d'autres. TRACE-VT pourrait rendre compte de ces différences individuelles en contrôlant la constante de temps dans le mécanisme d'adaptation et les coefficients de pondération entre le flux auditif et le flux visuel des biais articulatoires. *A priori*, il ne serait donc pas compliqué d'introduire cet élément à TRACE-VT. Cependant, il faudrait analyser nos données expérimentales d'une façon plus fine afin de mieux caractériser ces différences individuelles.

Liens avec les substrats neuronaux

Les travaux d'analyse et de modélisation de cette thèse sont effectués dans un cadre de modélisation cognitive et ils sont basés sur un modèle psycholinguistique de la parole : il y manque naturellement des mécanismes neuronaux « réalistes ». La continuité de ces travaux doit être, à notre sens, sous forme d'une modélisation neuro-compatible de l'effet de transformation verbale. Pour cela, une plateforme intéressante est celle des réseaux d'attracteurs dont nous avons utilisé certaines notions dans TRACE-VT. Ces réseaux se sont montrés puissants dans la modélisation des processus cognitifs tels que la mémoire à long-terme, la mémoire à court-terme, l'attention et la prise de décision utilisant les neurones biologiquement plausibles du néocortex (Rolls, 2010). Du point de vue du comportement du réseau, ils fournissent une base parfaitement compatible avec la modélisation des phénomènes multistables. De plus, ces réseaux peuvent fournir un bon cadre pour expliquer les données comportementales et neurophysiologiques de l'effet d'amorçage et l'effet de l'*adaptation aftereffect* en vision (Akrami *et al.*, 2008; Akrami et Treves, 2009).

Un aspect intéressant de cette perspective serait les interactions entre les travaux de modélisation et les données neurophysiologiques. En effet, une modélisation de type réseau d'attracteurs de l'effet de transformation verbale nécessiterait l'étude des activités transitoires neurophysiologiques en lien avec des bascules perceptives. L'objectif serait de vérifier l'éventuelle convergence d'un pattern d'activités neuronales vers un autre, représentant le nouveau percept, pendant une transformation verbale. De plus, l'analyse de la dynamique de l'effet de transformation verbale serait au cœur des réseaux d'attracteurs, ce qui est complémentaire avec les travaux effectués dans cette thèse qui ne portaient pas sur la dynamique des transformations verbales.

9.3 Conclusion générale

Bien que le paradigme de la multistabilité perceptive en parole suscite beaucoup de curiosité scientifique, il n'a pas fait l'objet d'études d'une manière aussi

large que la multistabilité perceptive en vision. Dans cette thèse, en se basant sur une problématique plus générale, le liage perceptif, nous avons essayé de mettre en œuvre différents potentiels du paradigme de la multistabilité perceptive comme outil d'investigation sur l'organisation perceptive de la scène de parole. À travers cette approche, nos études suggèrent des arguments en faveur d'une analyse de la scène de parole guidée par des principes multisensoriels et perceptuo-moteurs. Du point de vue théorique, nos résultats s'inscrivent dans la cadre de PACT (Théorie de la Perception pour le Contrôle de l'Action, voir sous-section 2.1.3) dont l'architecture générale est illustrée sur la figure 9.1. Nous remarquons sur cette figure que les liens perceptuo-moteurs se manifestent par deux voies. La première liant les « schémas moteurs » et la « catégorisation auditive » est à la base de la co-structuration des représentions perceptives et motrices. C'est ce lien qui rend possible l'utilisation des connaissances motrices dans les mécanismes de la catégorisation auditive. La deuxième liant les « schémas moteurs » et la « caractérisation auditive » est celle qui peut intervenir, en ligne, dans l'analyse de la scène de la parole afin d'améliorer l'intégration des différentes composantes du flux de parole dans la scène. C'est ce lien qui est en jeu dans ces travaux, et qui ressort, à notre sens, renforcé par les résultats de cette thèse. C'est à ce titre que les travaux de cette thèse sont largement commentés dans l'article de synthèse de Schwartz *et al.* (sous presse) dont l'auteur de cette thèse est aussi un co-auteur.

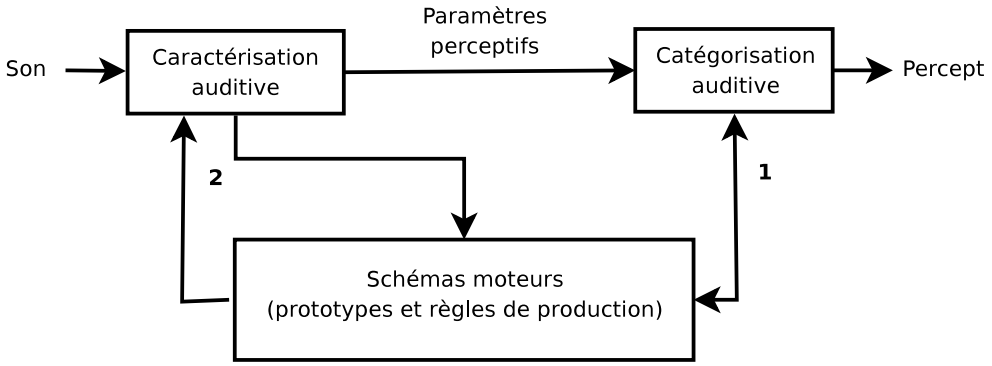


FIGURE 9.1 : Une architecture générale de PACT pour la perception de la parole. Les liens perceptuo-moteurs conduisent à la co-structuration des cartes sensorielles et motrices (1) et peuvent intervenir dans l'analyse de la scène de parole (2). Figure modifiée de Schwartz *et al.* (sous presse).

Une thèse se termine souvent par un nouveau départ. En ce qui concerne cette thèse, le départ se ferait naturellement vers l'étude plus approfondie des liens perception-action et leur rôle dans la perception de la parole chez l'adulte ainsi que lors de l'acquisition du langage. Il nous semble que c'est en traitant cette problématique de plusieurs abords (comportemental, neurophysiologique, modélisation computationnelle) et de différents niveaux (cognitif, neuronal, voire moléculaire) que nous parviendrons à ce but.

Cinquième partie

Bibliographie

Bibliographie

S. Achard et E. Bullmore : Efficiency and cost of economical brain functional networks. *PLoS Computational biology*, 3(2):e17, 2007. 207

A. Akrami, Y. Liu, A. Treves et B. Jagadeesh : Converging neuronal activity in inferior temporal cortex during the classification of morphed stimuli. *Cerebral Cortex*, 19(4):760–776, 2008. 208

A. Akrami et A. Treves : Neural basis of perceptual expectations : insights from transient dynamics of attractor neural networks. *BMC Neuroscience*, 10(Suppl 1):P174, 2009. 208

A. Alsius, J. Navarra, R. Campbell et S. Soto-Faraco : Audiovisual integration of speech falters under high attention demands. *Current Biology*, 15 (9):839–843, 2005. 47

L. Amini, C. Jutten, S. Achard, O. David, H. Soltanian-Zadeh, GA Hossein-Zadeh, P. Kahane, L. Minotti et L. Vercueil : Directed epileptic network from scalp and intracranial EEG of epileptic patients. *In IEEE International Workshop on Machine Learning for Signal Processing*, pages 1–6, 2009. 207

J.R. Anderson et C. Lebiere : *The atomic components of thought*. Lawrence Erlbaum Associates, 1998. 142

B. Arons : A review of the cocktail party effect. *Journal of the American Voice I/O Society*, 12(7):35–50, 1992. 22

A. Basirat : Le traitement de percepts phonetiques multistables : apport de la lecture labiale et etude neurophysiologique preliminaire. Mémoire de Master, Institut National Polytechnique de Grenoble, 2005. 180, 204

M.W. Beauvois et R. Meddis : Computer simulation of auditory stream segregation in alternating-tone sequences. *The Journal of the Acoustical Society of America*, 99:2270–2280, 1996. 18

D. Ben Shalom et D. Poeppel : Functional anatomic models of language : assembling the pieces. *The Neuroscientist*, 14(1):119–127, 2008. 37, 38

C. Benoit, T. Mohamadi et S. Kandel : Effects of phonetic context on audio-visual intelligibility of French. *Journal of Speech, Language and Hearing Research*, 37(5):1195, 1994. 42, 43

L.E. Bernstein, E.T. Auer et S. Takayanagi : Auditory speech detection in noise enhanced by lipreading. *Speech Communication*, 44(1-4):5–18, 2004. 44

L.E. Bernstein, Z.L. Lu et J. Jiang : Quantified acoustic–optical speech signal incongruity identifies cortical sites of audiovisual speech processing. *Brain Research*, 1242:172–184, 2008. 49

P. Bertelson, J. Vroomen, G. Wiegeraad et B. de Gelder : Exploring the relation between McGurk interference and ventriloquism. *In International conference on Spoken Language Processing*, volume 2, pages 559–562, 1994. 48

J. Besle, A. Fort, C. Delpuech et M-H. Giard : Bimodal speech : early suppressive visual effects in human auditory cortex. *European Journal of Neuroscience*, 20(8):2225–2234, 2004. 49

JR Binder, JA Frost, TA Hammeke, PSF Bellgowan, JA Springer, JN Kaufman et ET Possing : Human temporal lobe activation by speech and nonspeech sounds. *Cerebral Cortex*, 10(5):512–528, 2000. 37

J.R. Binder, E. Liebenthal, E.T. Possing, D.A. Medler et B.D. Ward : Neural correlates of sensory and decision processes in auditory object identification. *Nature Neuroscience*, 7(3):295–301, 2004. 41, 71, 129

R. Blake et N.K. Logothetis : Visual competition. *Nature Reviews Neuroscience*, 3(1):13–21, 2002. 56, 65

I. Bloch : Information combination operators for data fusion : A comparative review with classification. *In IEEE Transactions on Systems, Man and Cybernetics*, volume 2315, pages 148–159, 1994. 46

P. Boersma et D. Weenink : Praat, a system for doing phonetics by computer. *Glot International*, 5(9/10):341–345, 2001. 87, 89, 94, 99, 100

Y.S. Bonneh, M. Pavlovskaya, H. Ring et N. Soroker : Abnormal binocular rivalry in unilateral neglect : evidence for a non-spatial mechanism of extinction. *NeuroReport*, 15(3):473–477, 2004. 66

L. Brancazio : Lexical Influences in Audiovisual Speech Perception. *Journal of Experimental Psychology : Human Perception and Performance*, 30(3):445–463, 2004. 206

A.S. Bregman : *Auditory scene analysis : The perceptual organization of sound*. The MIT Press, 1990. 16, 17, 22, 25, 130

A.S. Bregman : Auditory scene analysis : hearing in complex environments. *In* E. McAdams, S. & Bigand, éditeur : *Thinking in Sound : the Cognitive Psychology of Human Audition*, pages 10–36. Oxford : Oxford University Press, 1993. 16

A.S. Bregman et J. Campbell : Primary auditory stream segregation and perception of order in rapid sequences of tones. *Journal of Experimental Psychology*, 89(2):244–249, 1971. 3, 4, 18, 68

S. Bringsjord : Declarative/logic-based computational cognitive modeling. *In* R. Sun, éditeur : *The Cambridge handbook of computational psychology*. Cambridge University Press New York, 2008. 142

J. Britz, T. Landis et C.M. Michel : Right parietal brain activity precedes perceptual alternation of bistable stimuli. *Cerebral Cortex*, 19(1):55–65, 2009. 66

B.R. Buchsbaum, R.K. Olsen, P.F. Koch, P. Kohn, J.S. Kippenhan et K.F. Berman : Reading, hearing, and the planum temporale. *NeuroImage*, 24(2):444–454, 2005. 51, 52

D. Burnham : Language specificity in the development of auditory-visual speech perception. *In* R. Campbell, B. Dodd et D. Burnham, éditeurs : *Hearing by eye II : Advances in the psychology of speechreading and auditory-visual speech*, pages 27–60. UK : Psychology Press, 1998. 47

D.E. Callan, J.A. Jones, A.M. Callan et R. Akahane-Yamada : Phonetic perceptual identification by native-and second-language speakers differentially activates brain regions involved with acoustic phonetic processing and those involved with articulatory–auditory/orosensory internal models. *NeuroImage*, 22 (3):1182–1194, 2004. 41, 129

D.E. Callan, J.A. Jones, K. Munhall, A.M. Callan, C. Kroos et E. Vatikiotis-Bateson : Neural processes underlying perceptual enhancement by visual speech gestures. *NeuroReport*, 14(17):2213–2218, 2003. 50, 93

G.A. Calvert, R. Campbell et M.J. Brammer : Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. *Current Biology*, 10(11):649–657, 2000. 48, 49, 52

R. Campbell : The processing of audio-visual speech : empirical and neural bases. *Philosophical Tranactions of The Royal British Society*, 363(1493):1001–1010, 2008. 45

R.P. Carlyon, R. Cusack, J.M. Foxton et I.H. Robertson : Effects of attention and unilateral neglect on auditory stream segregation. *Journal of Experimental Psychology : Human Perception and Performance*, 27(1):115–127, 2001. 21

G.A. Carpenter et S. Grossberg : A massively parallel architecture for a self-organizing neural pattern recognition machine. *Computer Vision, Graphics, and Image Processing*, 37(1):54–115, 1987. 141, 146, 153

GA Carpenter et S. Grossberg : The ART of adaptive pattern recognition by a self-organizing neural network. *Computer*, 21(3):77–88, 1988. 153

C.S. Carter, T.S. Braver, D.M. Barch, M.M. Botvinick, D. Noll et J.D. Cohen : Anterior cingulate cortex, error detection, and the online monitoring of performance. *Science*, 280(5364):747–749, 1998. 79

D. Chalmers : Facing up to the hard problem of consciousness. *Journal of Consciousness Studies*, 2(3):200–219, 1995. 8

X. Chen et S. He : Local factors determine the stabilization of monocular ambiguous and binocular rivalry stimuli. *Current Biology*, 14(11):1013–1017, 2004. 62, 63

E.C. Cherry : Some experiments on the recognition of speech, with one and with two ears. *Journal of the Acoustical Society of America*, 25(5):975–979, 1953. 15

J.M. Chowning : Computer synthesis of the singing voice by frequency modulation. *In* E. Jansson et J. Sundberg, éditeurs : *Sound generation in winds, strings, computers*, pages 4–13. Stockholm : Royal Swedish Academy of Music, 1980. 17

M.H. Christiansen, J. Allen et M.S. Seidenberg : Learning to segment speech using multiple cues : A connectionist model. *Language and Cognitive Processes*, 13(2–3):221–268, 1998. 149

M.H. Christiansen et N. Chater : Connectionist psycholinguistics in perspective. *In* M.H. Christiansen et N. Chater, éditeurs : *Connectionist psycholinguistics*, pages 19–75. Ablex Publishing Corporation, 2001. 151

C. Colin, M. Radeau, A. Soquet, D. Demolin, F. Colin et P. Deltenre : Mismatch negativity evoked by the McGurk-MacDonald effect : a phonetic representation within short-term memory. *Clinical Neurophysiology*, 113(4):495–506, 2002. 49

W.E. Cooper : Adaptation of phonetic feature analyzers for place of articulation. *The Journal of the Acoustical Society of America*, 56(2):617–627, 1974. 74

R. Cusack : The intraparietal sulcus and perceptual organization. *Journal of Cognitive Neuroscience*, 17(4):641–651, 2005. 70, 71

R. Cusack, D.J. Mitchell et J. Duncan : Discrete Object Representation, Attention Switching, and Task Difficulty in the Parietal Lobe. *Journal of Cognitive Neuroscience*, 22(1):32–47, 2010. 71

R. Cusack et B. Roberts : Effects of differences in timbre on sequential grouping. *Perception & Psychophysics*, 62(5):1112–1120, 2000. 19

A. Cutler : Prosody and the word boundary problem. *In* J. L. Morgan et K. Demuth, éditeurs : *Signal to syntax : Bootstrapping from speech to grammar in early acquisition*, pages 87–99. Lawrence Earlbaum, 1996. 111

A. Cutler et D.M. Carter : The predominance of strong initial syllables in the English vocabulary. *Compuer Speech and Language*, 2(2–3):133–142, 1987. 147

A. Cutler, J. Mehler, D. Norris et J. Segui : The syllable's differing role in the segmentation of French and English. *Journal of Memory and Language*, 25 (4):385–400, 1986. 147

A. Cutler et D. Norris : The role of strong syllables in segmentation for lexical. *Journal of Experimental Psychology : Human Perception and Performance*, 14 (1):113–121, 1988. 111, 146

D. Dahan et J.S. Magnuson : Spoken Word Recognition. *In* J. Traxler et M. A. Gernsbacher, éditeurs : *Handbook of Psycholinguistics*, pages 249–283. Academic Press, 2006. 148

A.R. Damasio : Time-locked multiregional retroactivation : a systems-level proposal for the neural substrates of recall and recognition. *Cognition*, 33(1-2):25–62, 1989. 1, 2

A. D'Ausilio, F. Pulvermüller, P. Salmas, I. Bufalari, C. Begliomini et L. Fadiga : The motor somatotopy of speech perception. *Current Biology*, 19 (5):381–385, 2009. 40, 41

B. Davis et P.F. MacNeilage : Universal Intrasyllabic Patterns in Early Acquisition. *In Proceedings of the 15th International Congress of Phonetic Sciences*, pages 379–382, 2003. 76

M.H. Davis : *Lexical segmentation in spoken word recognition.* Thèse de doctorat, Birkbeck College, University of London., 2000. 146

M.H. Davis, W.D. Marslen-Wilson et M.G. Gaskell : Leading up the lexical garden path : Segmentation and ambiguity in spoken word recognition. *Journal of Experimental : Psychology : Human Perception and Performance*, 28(1):218–241, 2002. 148

M.P. Deiber, V. Ibanez, N. Sadato et M. Hallett : Cerebral structures participating in motor preparation in humans : a positron emission tomography study. *Journal of Neurophysiology*, 75(1):233–247, 1996. 116

J.F. Demonet, G. Thierry et D. Cardebat : Renewal of the neurophysiology of language : functional neuroimaging. *Physiological Reviews*, 85(1):49–95, 2005. 36

J.E. Desmond, J.D.E. Gabrieli, A.D. Wagner, B.L. Ginier et G.H. Glover : Lobular patterns of cerebellar activation in verbal working-memory and finger-tapping tasks as revealed by functional MRI. *Journal of Neuroscience*, 17(24): 9675–9685, 1997. 79

A. Devergie, N. Grimault, F. Berthommier et E. Gaudrain : Pairing of vocalic streams and visual non speech cues. *In Speech and Face to Face Communication*, page 104, 2008. 205

H. Dewart : Disorders of language. *In* D. Groome, éditeur : *An introduction to cognitive psychology : Processes and Disorders*, pages 241–260. Routledge, 1999. 37

R.L. Diehl : Acoustic and auditory phonetics : the adaptive design of speech sound systems. *Philosophical Transactions of the Royal Society B : Biological Sciences*, 363(1493):965–978, 2008. 32

R.L. Diehl et K.R. Kluender : On the objects of speech perception. *Ecological Psychology*, 1(2):121–144, 1989a. 32

R.L. Diehl et K.R. Kluender : Reply to commentators. *Ecological Psychology*, 1 (2):195–225, 1989b. 32

R.L. Diehl, A.J. Lotto et L.L. Holt : Speech perception. *Annual Review of Psychology*, 55:149–179, 2004. 28, 31, 33

I. Dinstein : Human cortex : reflections of mirror neurons. *Current Biology*, 18 (20):956–959, 2008. 31

T. Ditzinger et H. Haken : A synergetic model of multistability in perception. *In* P. Kruse et M. Stadler, éditeurs : *Ambiguity in Mind and Nature : Multistable Cognitive Phenomena*, pages 255–274. Springer, 1995. 158, 160

T. Ditzinger, B. Tuller, H. Haken et J.A.S. Kelso : A synergetic model for the verbal transformation effect. *Biological cybernetics*, 77(1):31–40, 1997a. 158, 161

T. Ditzinger, B. Tuller et J.A.S. Kelso : Temporal patterning in an auditory illusion : The verbal transformation effect. *Biological cybernetics*, 77(1):23–30, 1997b. 161

B. Dodd : The role of vision in the perception of speech. *Perception*, 6(1):31–40, 1977. 44

M. Dohen, H. Lœvenbruck, M.A. Cathiard et J.L. Schwartz : Visual perception of contrastive focus in reiterant French speech. *Speech Communication*, 44 (1-4):155–172, 2004. 111

S. Dufour, R. Peereman, C. Pallier et M. Radeau : VoCoLex : A lexical database on phonological similarity between french words. *L'Année Psychologique*, 102(4):725–746, 2002. 86, 99

J.L. Elman : Finding structure in time. *Cognitive Science*, 14(2):179–211, 1990. 145, 151

A.K. Engel, P. König, A.K. Kreiter et W. Singer : Interhemispheric synchronization of oscillatory neuronal responses in cat visual cortex. *Science*, 252:1177–1179, 1991. 9

J. Eriksson, A. Larsson, K.R. Ahlstrom et L. Nyberg : Similar frontal and distinct posterior cortical regions mediate visual and auditory perceptual awareness. *Cerebral Cortex*, 17(4):760, 2007. 130

L. Fadiga, L. Craighero, G. Buccino et G. Rizzolatti : Speech listening specifically modulates the excitability of tongue muscles : a TMS study. *European Journal of Neuroscience*, 15(2):399–402, 2002. 31

P.F. Ferrari, V. Gallese, G. Rizzolatti et L. Fogassi : Mirror neurons responding to the observation of ingestive and communicative mouth actions in the monkey ventral premotor cortex. *European Journal of Neuroscience*, 17(8):1703–1714, 2003. 30

Y.I. Fishman, D.H. Reser, J.C. Arezzo et M. Steinschneider : Neural correlates of auditory stream segregation in primary auditory cortex of the awake monkey. *Hearing Research*, 151(1-2):167–187, 2001. 19, 20

M. Fort, E. Spinelli, C. Savariaux et S. Kandel : The word superiority effect in audiovisual speech perception. *Speech Communication*, 52(6):525–532, 2010. 111, 206

C.A. Fowler : An Event Approach to the Study of Speech Perception from a Direct-realistic Perspective. *Journal of Phonetics*, 14(1):3–28, 1986a. 29

C.A. Fowler : Reply to commentators. *Journal of Phonetics*, 14:149–170, 1986b. 29

C.A. Fowler : Speech perception : direct realist theory. *In* R.E. Asher, éditeur : *Encyclopedia of language and linguistics*, volume 8, pages 4199–4203. Oxofrd : Pergamon, 1994. 29

C.A. Fowler : Listeners do hear sounds, not tongues. *Journal of the Acoustical Society of America*, 99(3):1730–1741, 1996. 29

C.A. Fowler et M.R. Smith : Speech perception as "vector analysis" : an approach to the problems of invariance and segmentation. *In* Perkell J.S. et Klatts D.H., éditeurs : *Invariance and Variability in Speech Processes*, pages 123–139. Erlbaum, Hillsdale, NJ, 1986. 30

A.D. Friederici : Towards a neural basis of auditory sentence processing. *Trends in Cognitive Sciences*, 6(2):78–84, 2002. 37

A.D. Friederici et J.M.I. Wessels : Phonotactic knowledge of word boundaries and its use in infant speech perception. *Perception and Psychophysics*, 54(3):287–287, 1993. 147

P. Fries, D. Nikolić et W. Singer : The gamma cycle. *Trends in Neurosciences*, 30(7):309–316, 2007. 84, 114

B. Galantucci, C.A. Fowler et M.T. Turvey : The motor theory of speech perception reviewed. *Psychonomic bulletin & review*, 13(3):361–377, 2006. 31

T. Gambell et C. Yang : Word segmentation : Quick but not dirty. Yale University. Manuscrit non publié., 2005. 146, 147

W.F. Ganong : Phonetic categorization in auditory word perception. *Journal of Experimental Psychology : Human Perception and Performance*, 6(1):110–125, 1980. 150

M.G. Gaskell : Statistical and connectionist models of speech perception and word recognition. *In* M.G. Gaskell, éditeur : *The Oxford Handbook of Psycholinguistics*, pages 55–69. Oxford University Press, 2007. 150

M.G. Gaskell et W.D. Marslen-Wilson : Integrating form and meaning : A distributed model of speech perception. *In* G.T.M. Altman, éditeur : *A special issue of language and cognitive process : psycholinguistics and computational perspectives on the lexicon*, volume 12, pages 613–656. Psychology Press, 1997. 145, 148

N. Geschwind : Language and the brain. *Scientific American*, 226(4):76–83, 1972. 37

A.A. Ghazanfar, C. Chandrasekaran et N.K. Logothetis : Interactions between the superior temporal sulcus and auditory cortex mediate dynamic face/voice integration in rhesus monkeys. *Journal of Neuroscience*, 28(17):4457–4469, 2008. 49, 52

A.A. Ghazanfar, J.X. Maier, K.L. Hoffman et N.K. Logothetis : Multisensory integration of dynamic faces and voices in rhesus monkey auditory cortex. *Journal of Neuroscience*, 25(20):5004–5012, 2005. 44

K.W. Grant et P.F. Seitz : The use of visible speech cues for improving auditory detection of spoken sentences. *The Journal of the Acoustical Society of America*, 108:1197–1208, 2000. 43

C.M. Gray, P. König, A.K. Engel et W. Singer : Oscillatory responses in cat visual cortex exhibit inter-columnar synchronization which reflects global stimulus properties. *Nature*, 338(6213):334–337, 1989. 9

K. Green, P. Kuhl, A. Meltzoff et E. Stevens : Integrating speech information across talkers, gender, and sensory modality : female faces and male voices in the McGurk effect. *Perception & Psychophysics*, 50(6):524–536, 1991. 48

T. L. Griffiths, C. Kemp et J.B. Tenenbaum : Bayesian models of cognition. *In* R. Sun, éditeur : *The Cambridge handbook of computational psychology*. Cambridge University Press New York, 2008. 140, 141

N. Grimault, S.P. Bacon et C. Micheyl : Auditory stream segregation on the basis of amplitude-modulation rate. *The Journal of the Acoustical Society of America*, 111(3):1340–1348, 2002. 19

N. Grimault, C. Micheyl, R.P. Carlyon, P. Arthaud et L. Collet : Influence of peripheral resolvability on the perceptual segregation of harmonic complex tones differing in fundamental frequency. *The Journal of the Acoustical Society of America*, 108(1):263–271, 2000. 19

S. Grossberg, I. Boardman, M. Cohen et S. Percepts : Neural dynamics of variable-rate speech categorization. *Journal of Experimental Psychology : Human Perception and Performance*, 23(2):481–503, 1997. 154

S. Grossberg et C.W. Myers : The resonant dynamics of speech perception : Interword integration and duration-dependent backward effects. *Psychological Review*, 107(4):735–767, 2000. 154, 155

A. Gutschalk, C. Micheyl, J.R. Melcher, A. Rupp, M. Scherg et A.J. Oxenham : Neuromagnetic correlates of streaming in human auditory cortex. *Journal of Neuroscience*, 25(22):5382–5388, 2005. 70, 71

H. Haken : *Synergetic computers and cognition : a top-down approach to neural nets*. Springer, 1991. 158

W.M. Hartmann et D. Johnson : Stream segregation and peripheral channeling. *Music Perception*, 9(2):155–184, 1991. 18

HR Heekeren, S. Marrett, PA Bandettini et LG Ungerleider : A general mechanism for perceptual decision-making in the human brain. *Nature*, 431(7010): 859–862, 2004. 80

H. Helmholtz : *On the sensations of tone*. New York, Dover, 1877. English translation A.J. Ellis, 1954. 16

G. Hickok : Eight problems for the mirror neuron theory of action understanding in monkeys and humans. *Journal of Cognitive Neuroscience*, 21(7):1229–1243, 2009a. 31

G. Hickok : The functional neuroanatomy of language. *Physics of Life Reviews*, 6 (3):121–143, 2009b. 39

G. Hickok : What does the motor system contribute to speech perception ? Transparents de l'exposé orale à Neurobiology of Language Conference, Chicago, États-Unis, 2009c. 40

G. Hickok, B. Buchsbaum, C. Humphries et T. Muftuler : Auditory-motor interaction revealed by fMRI : speech, music, and working memory in area Spt. *Journal of Cognitive Neuroscience*, 15(5):673–682, 2003. 51

G. Hickok et D. Poeppel : Dorsal and ventral streams : a framework for understanding aspects of the functional anatomy of language. *Cognition*, 92(1-2):67–99, 2004. 36, 37, 39, 203

G. Hickok et D. Poeppel : The cortical organization of speech processing. *Nature Reviews Neuroscience*, 8(5):393–402, 2007. 36, 37, 39, 40, 114, 116, 203

G.E. Hinton et T.J. Sejnowski : Learning and relearning in Boltzmann machines. *In* D.E. Rumelhart et J.L. Mcclelland, éditeurs : *Parallel distributed processing : explorations in the microstructure of cognition*, volume 1, pages 282–317. MIT Press, Cambridge, MA, 1986. 176, 177

J. Hocking et C.J. Price : The role of the posterior superior temporal sulcus in audiovisual processing. *Cerebral Cortex*, 18(10):2439–2449, 2008. 49

G.D. Honey, E.T. Bullmore et T. Sharma : Prolonged Reaction Time to a Verbal Working Memory Task Predicts Increased Power of Posterior Parietal Cortical Activation. *NeuroImage*, 12(5):495–503, 2000. 116

JJ Hopfield : Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences of the United States of America*, 79(8):2554–2558, 1982. 176

J.F. Houde et M.I. Jordan : Sensorimotor adaptation of speech I : Compensation and adaptation. *Journal of Speech, Language, and Hearing Research*, 45(2):295–310, 2002. 39

J.E. Hummel : Complementary solutions to the binding problem in vision : Implications for shape perception and object recognition. *Visual Cognition*, 8(3-5):489–517, 2001. 15

J.M. Hupé, L.M. Joffo et D. Pressnitzer : Bistability for audiovisual stimuli : Perceptual decision is modality specific. *Journal of Vision*, 8(7):1–15, 2008. 69, 70

R. Jakobson, G. Fant et M. Halle : *Preliminaries to speech analysis : The distinctive features and their correlates*. MIT press Cambridge, 1969. 166

O. Jensen, J. Kaiser et J.P. Lachaux : Human gamma-frequency oscillations associated with attention and memory. *Trends in Neurosciences*, 30(7):317–324, 2007. 11, 13

D. Jilk, C. Lebiere, R. O'Reilly et J. Anderson : SAL : an explicitly pluralistic cognitive architecture. *Journal of Experimental & Theoretical Artificial Intelligence*, 20(3):197–218, 2008. 142

G. Johansson : Visual perception of biological motion and a model for its analysis. *Perception & psychophysics*, 14(2):201–211, 1973. 30

M. Joliot, U. Ribary et R. Llinas : Human oscillatory brain activity near 40 Hz coexists with cognitive temporal binding. *Proceedings of the National Academy of Sciences*, 91(24):11748–11751, 1994. 11, 12

T.R. Jordan et P. Sergeant : Effects of distance on visual and audiovisual speech recognition. *Language and Speech*, 43(1):107–124, 2000. 44

P. Kahane, L. Minotti, D. Hoffmann, J.P. Lachaux et P. Ryvlin : Invasive EEG in the definition of the seizure onset zone : depth electrodes. *In Handbook of clinical neurophysiology*, volume 3, pages 109–133. Elsevier, 2004. 115

Z. Kaminska, M. Pool et P. Mayer : Verbal Transformation : Habituation or Spreading Activation ? *Brain and Language*, 71(2):285–298, 2000. 73

B. Kast : Decisions, decisions... *Nature*, 411(6834):126–128, 2001. 79, 80

N. Kawabata et T. Mori : Disambiguating ambiguous figures by a model of selective attention. *Biological Cybernetics*, 67(5):417–425, 1992. 56

R.D. Kent : Sonority theory and syllable pattern as keys to sensory-motor-cognitive interactions in infant vocal development. *In* B. Boysson-Bardies, S. de Schonen, P. Jusczyk, P. McNeilage et J. Morton, éditeurs : *Developmental neurocognition : Speech and face processing in the first year of life*, pages 329–340. Kulwer Academic Publishers, 1993. 181

J.N. Kim et M.N. Shadlen : Neural correlates of a decision in the dorsolateral prefrontal cortex of the macaque. *Nature Neuroscience*, 2(2):176–185, 1999. 79

P.C. Klink, R. van Ee, M.M. Nijs, G.J. Brouwer, A.J. Noest et R.J.A. van Wezel : Early interactions between neuronal adaptation and voluntary control determine perceptual choices in bistable vision. *Journal of Vision*, 8(5):1–18, 2008. 64

E. Kohler, C. Keysers, M.A. Umilta, L. Fogassi, V. Gallese et G. Rizzolatti : Hearing sounds, understanding actions : action representation in mirror neurons. *Science*, 297(5582):846–848, 2002. 30

B. Kokinov : The DUAL cognitive architecture : A hybrid multi-agent approach. *In Proceedings of the Eleventh European Conference of Artificial Intelligence*, pages 203–207, 1994. 143

H.M. Kondo et M. Kashino : Neural mechanisms of auditory awareness underlying verbal transformations. *NeuroImage*, 36(1):123–130, 2007. 78, 84, 111, 113, 114, 117, 129, 187, 202

J. Kornmeier, W. Ehm, H. Bigalke et M. Bach : Discontinuous presentation of ambiguous figures : How interstimulus-interval durations affect reversal dynamics and ERPs. *Psychophysiology*, 44(4):552–560, 2007. 62, 63, 110

A.K. Kreiter et W. Singer : Stimulus-dependent synchronization of neuronal responses in the visual cortex of the awake macaque monkey. *Journal of Neuroscience*, 16(7):2381–2396, 1996. 9

M. Lalain, N. Vallée et J.L. Schwartz : Du percept auditif au geste articulatoire : les capacités perceptuo-motrices chez les enfants normolecteurs et dyslexiques. *In Actes des XXVIIèmes Journées d'Études sur la parole*, 2008. 204

M.T. Lallouache : Un poste « visage-parole ». acquisition et traitement de contours labiaux. *In Actes des XVIIIèmes Journées d'Études sur la Parole*, pages 282–286, 1990. 87, 99

L. Lancia : *Dynamique non linéaire de la perception de la parole*. Thèse de doctorat, Université de Provence - Aix-Marseille 1, 2009. 141

L. Lancia, N. Nguyen et B. Tuller : Nonlinear dynamics of speech categorization : critical slowing down and critical fluctuations. *The Journal of the Acoustical Society of America*, 123(5):3077, 2008. 141

P. Langley, J.E. Laird et S. Rogers : Cognitive architectures : Research issues and challenges. *Cognitive Systems Research*, 10(2):141–160, 2008. 143

N.J. Lass et S.S. Golden : The use of isolated vowels as auditory stimuli in eliciting the verbal transformation effect. *Canadian Journal of Psychology*, 25(4):349–359, 1971. 73

J.H. Lee, B.E. Russ, L.E. Orr, Y.E. Cohen et Y. Cohen : Prefrontal activity predicts monkeys' decisions during an auditory category task. *Frontiers in Integrative Neuroscience*, 3(16):1–12, 2009. 71

S.H. Lee, R. Blake et D.J. Heeger : Traveling waves of activity in primary visual cortex during binocular rivalry. *Nature Neuroscience*, 8(1):22–23, 2005. 60, 61

I. Lehiste : The timing of utterances and linguistic boundaries. *The Journal of the Acoustical Society of America*, 51(6):2018–2024, 1972. 147

S.R. Lehky : An astable multivibrator model of binocular rivalry. *Perception*, 17 (2):215–228, 1988. 56

D.A. Leopold et N.K. Logothetis : Multistable phenomena : changing views in perception. *Trends in Cognitive Sciences*, 3(7):254–264, 1999. 56, 57, 59, 62, 65, 67, 69, 70, 80, 129

D.A. Leopold, M. Wilke, A. Maier et N.K. Logothetis : Stable perception of visually ambiguous patterns. *Nature Neuroscience*, 5(6):605–609, 2002. 62, 110

R.L. Lewis : Computational psycholinguistics. *In Encyclopedia of Cognitive Science*. London : Macmillon (Nature Publishing Group), 2000. 143

A.M. Liberman : *Speech : A special code*. The MIT Press, 1996. 28, 29

A.M. Liberman, F.S. Cooper, D.P. Shankweiler et M. Studdert-Kennedy : Perception of speech code. *Psychological Review*, 74(6):431–461, 1967. 28

A.M. Liberman et I.G. Mattingly : The motor theory of speech perception revised. *Cognition*, 21(1):1–36, 1985. 28

A.M. Liberman et D.H. Whalen : On the relation of speech to language. *Trends in Cognitive Sciences*, 4(5):187–196, 2000. 28

J. Liljencrants et B. Lindblom : Numerical simulation of vowel quality systems : the role of perceptual contrast. *Language*, 48(4):839–862, 1972. 31

B. Lindblom : Phonetic universals in vowel systems. *In* Ohala J. et Jaeger J., éditeurs : *Experimental phonology*, pages 13–44. Oralndo : Academic Press, 1986. 32

B. Lindblom : On the notion of "possible speech sound". *Journal of phonetics*, 18 (2):135–152, 1990. 32

B. Lindblom et J. Lubker : The speech Humunculus and a problem of phonetic linguistics. *In* V. Fromkin, éditeur : *Phonetic Linguistics, Essays in Honor of Peter Ladefoged*, pages 169–192. Orlando : Academic Press, 1985. 204

G.M. Long et T.C. Toppino : Enduring interest in perceptual ambiguity : Alternating views of reversible figures. *Psychological Bulletin*, 130(5):748–768, 2004. 56, 67, 68

A.J. Lotto, G.S. Hickok et L.L. Holt : Reflections on mirror neurons and speech perception. *Trends in Cognitive Sciences*, 13(3):110–114, 2009. 31

P.A. Luce, D.B. Pisoni et S.D. Goldinger : Similarity neighborhoods of spoken words. *In* G.T.M. Altman, éditeur : *Cognitive models of speech processing : Psycholinguistic and computational perspectives*, pages 122–147. Cambridge, MIT Press, 1990. 144

R.D. Luce : *Individual choice behavior*. Wiley New York, 1959. 172

E.D. Lumer, K.J. Friston et G. Rees : Neural correlates of perceptual rivalry in the human brain. *Science*, 280(5371):1930–1934, 1998. 65, 66

W. Maass, T. Natschläger et H. Markram : Real-time computing without stable states : A new framework for neural computation based on perturbations. *Neural computation*, 14(11):2531–2560, 2002. 179

D.G. MacKay : The problems of flexibility, fluency, and speed-accuracy trade-off in skilled behavior. *Psychological Review*, 89(5):483–506, 1982. 73, 156

D.G. MacKay : The organization of perception and action. A theory for language and other cognitive skills. *The Italian Journal of Neurological Sciences*, 9(3):303–303, 1988. 73, 156

D.G. MacKay, G. Wulf, C. Yin et L. Abrams : Relations between word perception and production : new theory and data of the verbal transformation effect. *Journal of Memory and Language*, 32(5):624–646, 1993. 73, 156, 158, 161, 187

A. MacLeod et Q. Summerfield : Quantifying the contribution of vision to speech perception in noise. *British Journal of Audiology*, 21(2):131–141, 1987. 43

P.F. MacNeilage et B.L. Davis : On the origin of internal structure of word forms. *Science*, 288(5465):527–531, 2000. 76

S. Maeda : An articulatory model of the tongue based on a statistical analysis. *The Journal of the Acoustical Society of America*, 65:S22, 1979. 35

A. Maier, M. Wilke, N.K. Logothetis et D.A. Leopold : Perception of temporally interleaved ambiguous patterns. *Current Biology*, 13(13):1076–1085, 2003. 62

D. Mareschal et M.S.C. Thomas : Computational modeling in developmental psychology. *IEEE Transactions on Evolutionary Computation*, 11(2):137–150, 2007. 138

D. Marr : *Vision : A computational investigation into the human representation and processing of visual information*. San Francisco : W.H. Freeman, 1982. 141

W. Marslen-Wilson : Linguistic structure and speech shadowing at very short latencies. *Nature*, 244(5417):522–523, 1973. 95

W. Marslen-Wilson : Issues of process and representation in lexical access. *In* G.T.M. Altman et R. Shillcock, éditeurs : *Cognitive models of speech processing : The second Sperlonga meeting*, pages 187–210, 1993. 148

W.D. Marslen-Wilson et A. Welsh : Processing Interactions and Lexical Access during Word Recognition in Continuous Speech. *Cognitive Psychology*, 10(1):29–63, 1978. 111, 147

D.W. Massaro : *Perceiving talking faces : From speech perception to a behavioral principle*. The MIT Press, 1998. 46

J. L. McClelland et J. L. Elman : The TRACE model of speech perception. *Cognitive Psychology*, 18(1):1–86, 1986. 149, 163, 167, 169, 170, 183, 203

J.L. McClelland : The place of modeling in cognitive science. *Topics in Cognitive Science*, 1(1):11–38, 2009. 137, 138, 139, 140

M. McGrath et Q. Summerfield : Intermodal timing relations and audio-visual speech recognition by normal-hearing adults. *The Journal of the Acoustical Society of America*, 77(2):678–685, 1985. 48

H. McGurk et J. MacDonald : Hearing lips and seeing voices. *Nature*, 264 (5588):746–748, 1976. 44

J.M. McQueen, A. Cutler, T. Briscoe et D. Norris : Models of continuous speech recognition and the contents of the vocabulary. *Language and Cognitive Processes*, 10(3-4):309–331, 1995. 148

J. Mehler, J.Y. Dommergues, U. Frauenfelder et J. Segui : The syllable's role in speech segmentation. *Journal of Verbal Learning & Verbal Behavior.*, 20 (3):298–305, 1981. 144

I.G. Meister, S.M. Wilson, C. Deblieck, A.D. Wu et M. Iacoboni : The essential role of premotor cortex in speech perception. *Current Biology*, 17(19): 1692–1696, 2007. 40

C. Micheyl, B. Tian, R.P. Carlyon et J.P. Rauschecker : Perceptual organization of tone sequences in the auditory cortex of awake macaques. *Neuron*, 48 (1):139–148, 2005. 19, 20, 21, 71

L.M. Miller et M. D'Esposito : Perceptual fusion and stimulus coincidence in the cross-modal integration of speech. *Journal of Neuroscience*, 25(25):5884–5893, 2005. 51

D. Mirman, J.L. McClelland et L.L. Holt : An interactive Hebbian account of lexically guided tuning of speech perception. *Psychonomic Bulletin & Review*, 13 (6):958–965, 2006. 150, 165

J. Mishra, A. Martinez, T.J. Senkowski et S.A. Hillyard : Early cross-modal interactions in auditory and visual cortex underlie a sound-induced visual illusion. *Journal of Neuroscience*, 27(15):42120–4131, 2007. 12, 13

R. Moreno-Bote, J. Rinzel et N. Rubin : Noise-induced alternations in an attractor network model of perceptual bistability. *Journal of Neurophysiology*, 98 (3):1125–1139, 2007. 56

L. Ménard, J.L. Schwartz et J. Aubin : Invariance and variability in the production of the height feature in French vowels. *Speech Communication*, 50(1):14–28, 2008. 34, 35, 36

R. Näätänen : The perception of speech sounds by the human brain as reflected by the mismatch negativity (MMN) and its magnetic equivalent (MMNm). *Psychophysiology*, 38(1):1–21, 2001. 116

H.R. Naghavi et L. Nyberg : Common fronto-parietal activity in attention, memory, and consciousness : shared demands on integration ? *Consciousness and Cognition*, 14(2):390–425, 2005. 66, 130

T. Natsoulas : A study of the verbal-transformation effect. *The American Journal of Psychology*, 78(2):257–263, 1965. 74

L.A. Necker : Observations on some remarkable phenomena seen in switzerland and optical phenomenon which occurs on viewing a figure of a crystal or geometrical solid. *The London and Edinburgh Philosophical Magazine and Journal of Science*, 1:329–337, 1832. 3

U. Neisser : *Cognitive psychology*. Appleton-Century-Crofts New York, 1967. 16

A. Newell et H.A. Simon : Computer science as empirical inquiry : Symbols and search. *Communications of the ACM*, 19(3):113–126, 1976. 142

A.J. Noest, R. Van Ee, M.M. Nijs et R.J. Van Wezel : Percept-choice sequences driven by interrupted ambiguous stimuli : A low-level neural model. *Journal of vision*, 7(8):1–14, 2007. 64

O. Nohorna : L'émergence des formes audiovisuelles dans le traitement multisensoriel de la parole : expériences et modélisation. Mémoire de Master, Grenoble INP, 2009. 47

D. Norris : Shortlist : A connectionist model of continuous speech recognition. *Cognition*, 52(3):189–234, 1994. 138, 151

D. Norris : How do computational models help us build better theories. *In* A. Cutler, éditeur : *Twenty-first century psycholinguistics : Four cornerstones*, pages 331–346. Lawrence Erlbaum Associates, 2005. 138, 139, 151, 173

D. Norris et J.M. McQueen : Shortlist B : A Bayesian model of continuous speech recognition. *Psychological Review*, 115(2):357–395, 2008. 152

D. Norris, JM McQueen et A. Cutler : Competition and segmentation in spoken-word recognition. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, 21(5):1209–1228, 1995. 149

D. Norris, J.M. McQueen et A. Cutler : Merging information in speech recognition : Feedback is never necessary. *Behavioral and Brain Sciences*, 23(3):299–325, 2000. 176

J. Obleser, H. Boecker, A. Drzezga, B. Haslinger, A. Hennenlotter, M. Roettinger, C. Eulitz et J.P. Rauschecker : Vowel sound extraction in anterior superior temporal cortex. *Human Brain Mapping*, 27(7):562–571, 2006. 41

G.C. Oden et D.W. Massaro : Integration of featural information in speech perception. *Psychological Review*, 85(3):172–191, 1978. 46

V. Ojanen, R. Möttönen, J. Pekkola, I.P. Jääskeläinen, R. Joensuu, T. Autti et M. Sams : Processing of audiovisual speech in Broca's area. *NeuroImage*, 25(2):333–338, 2005. 50

K. Okada et G. Hickok : Two cortical mechanisms support the integration of visual and auditory speech : A hypothesis and preliminary data. *Neuroscience Letters*, 452(3):219–223, 2009. 51, 52

J. Orbach, D. Ehrlich et H.A. Heath : Reversibility of the necker cube. i. an examination of the concept of "satiation of orientation". *Perceptual and Motor Skills*, 17(2):439–458, 1963. 62

R.C. O'Reilly et Y. Munakata : *Computational explorations in cognitive neuroscience : Understanding the mind by simulating the brain.* The MIT Press, 2000. 142

T. Otake, G. Hatano, E.A. Cutler et J. Mehler : Mora or syllable ? Speech segmentation in Japanese. *Journal of Memory and Language*, 32(2):258–278, 1993. 147

J. Pa et G. Hickok : A parietal–temporal sensory–motor integration area for the human vocal tract : Evidence from an fMRI study of skilled musicians. *Neuropsychologia*, 46(1):362–368, 2008. 52

J.S. Pardo et R.E. Remez : *The perception of speech*, pages 201–248. New York : Academic Press., 2006. 24

L. Parkkonen, J. Andersson, M. Hämäläinen et R. Hari : Early visual brain areas reflect the percept of an ambiguous scene. *Proceedings of the National Academy of Sciences*, 105(51):20500–20504, 2008. 60, 61

J. Pearson et J. Brascamp : Sensory memory for ambiguous vision. *Trends in Cognitive Sciences*, 12(9):334–341, 2008. 62

M.A. Pitt et I.J. Myung : When a good fit can be bad. *Trends in Cognitive Sciences*, 6(10):421–425, 2002. 139

M.A. Pitt et L. Shoaf : The source of a lexical bias in the Verbal Transformation Effect. *Spoken Word Access Processes*, 16(5–6):715–721, 2001. 73

M.A. Pitt et L. Shoaf : Linking verbal transformations to their causes. *Journal of Experimental : Psychology : Human Perception and Performance*, 28(1):150–162, 2002. 72, 74, 130

M.L. Platt : Neural correlates of decisions. *Current Opinion in Neurobiology*, 12 (2):141–148, 2002. 79, 80

D.C. Plaut : Methodologies for the computer modeling of human cognitive processes. *In* F. Boller, J. Grafman et G. Rizzotti, éditeurs : *Handbook of Neuropsychology*, volume 1, pages 259–268. Elsevier, 2000. 138, 139

A. Polonsky, R. Blake, J. Braun et D.J. Heeger : Neuronal activity in human primary visual cortex correlates with perception during binocular rivalry. *Nature Neuroscience*, 3(11):1153–1159, 2000. 59

R.F. Port : The dynamical systems hypothesis in cognitive science. *In MacMillan Encyclopedia of Cognitive Science*. London. Amy Lockyer, 2000. 141

R.J. Porter et F.X. Castellanos : Speech-production measures of speech perception : Rapid shadowing of VCV syllables. *The Journal of the Acoustical Society of America*, 67(4):1349–1356, 1980. 95

D. Pressnitzer et J.M. Hupé : Is auditory streaming a bistable percept ? *In Forum Acusticum 2005 Budapest*, pages 1557–1561, 2005. 57, 58

D. Pressnitzer et J.M. Hupé : Temporal dynamics of auditory and visual bistability reveal common principles of perceptual organization. *Current Biology*, 16 (13):1351–1357, 2006. 68, 69

D. Pressnitzer, M. Sayles, C. Micheyl et I.M. Winter : Perceptual organization of sound begins in the auditory periphery. *Current Biology*, 18(15):1124–1128, 2008. 21

C.J. Price : Functional imaging studies of aphasia. *In* J.C. Mazziotta, A.W. Toga et R.S. Frackowiak, éditeurs : *Brain Mapping : the Disorders.*, pages 181–200. Academic Press, 2000. 37

A. Protopapas : Connectionist modeling of speech perception. *Psychological Bulletin*, 125:410–436, 1999. 150, 152

F. Pulvermüller, M. Huss, F. Kherif, F. Moscoso del Prado Martin, O. Hauk et Y. Shtyrov : Motor cortex maps articulatory features of speech sounds. *Proceedings of the National Academy of Sciences*, 103(20):7865–7870, 2006. 31, 41

H. Purwins, B. Blankertz et K. Obermayer : Computing auditory perception. *Organised Sound*, 5(3):159–171, 2001. 17

Z.W. Pylyshyn : The role of cognitive architecture in theories of cognition. *In* K. VanLehn, éditeur : *Architectures for intelligence : the twenty-second Carnegie Mellon Symposium on Cognition*, pages 189–223. Lawrence Erlbaum Associates, 1991. 142

C.M. Reed : The implications of the Tadoma method of speechreading for spoken language processing. *In Fourth International Conference on Spoken Language Processing*, volume 3, pages 1489–1492, 1996. 42

D. Reisberg, J. McLean et A. Goldfield : Easy to hear but hard to understand : A lip-reading advantage with intact auditory stimuli. *In* B. Dodd et R. Campbell, éditeurs : *Hearing by eye : The psychology of lip-reading*, pages 97–113. Lawrence Erlbaum Associates, 1987. 43

D. Reisberg, J.D. Smith, D.A. Baxter et M. Sonenshine : “Enacted” auditory images are ambiguous ;" pure" auditory images are not. *The Quarterly journal of experimental psychology. A, Human experimental psychology*, 41(3):619–641, 1989. 75, 205

R.E. Remez, P.E. Rubin, S.M. Berns, J.S. Pardo et J.M. Lang : On the PerceptualOrganizationofSpeech. *Psychological Review*, 101(1):129–156, 1994. 22, 23, 24, 25, 48, 130

R.E. Remez, P.E. Rubin, D.B. Pisoni et T.D. Carrell : Speech perception without traditional speech cues. *Science*, 212(4497):947–949, 1981. 23

A. Revonsuo : Binding and the phenomenal unity of consciousness. *Consciousness and Cognition*, 8(2):173–185, 1999. 7

G. Rizzolatti et M.A. Arbib : Language within our grasp. *Trends in Neurosciences*, 21(5):188–194, 1998. 31

G. Rizzolatti et L. Craighero : The mirror-neuron system. *Annual Review of Neuroscience*, 27:169–192, 2004. 31

G. Rizzolatti, L. Fogassi et V. Gallese : Neurophysiological mechanisms underlying the understanding and imitation of action. *Nature Reviews Neuroscience*, 2(9):661–670, 2001. 30

J. Robert-Ribes, J.L. Schwartz et P. Escudier : A comparison of models for fusion of the auditory and visual sensors in speech perception. *Artificial Intelligence Review*, 9(4):323–346, 1995. 45

S. Roberts et H. Pashler : How persuasive is a good fit ? a comment on theory testing. *Psychological Review*, 107(2):358–367, 2000. 139

A. Rochet-Capellan et J.L. Schwartz : An articulatory basis for the labial-to-coronal effect : /pata/ seems a more stable articulatory pattern than /tapa/. *The Journal of the Acoustical Society of America*, 121(6):3740–3754, 2007. 76

I. Rock et K. Mitchener : Further evidence of failure of reversal of ambiguous figures by uninformed subjects. *Perception*, 21(1):39–45, 1992. 62

R. Rojas : *Neural Networks : A Systematic Introduction*. Springer, 1996. 177, 178

E.T. Rolls : Attractor networks. *Wiley Interdisciplinary Reviews : Cognitive Science*, 1(1):119–134, 2010. 208

L. Romero, V. Walsh et C. Papagno : The Neural Correlates of Phonological Short-term Memory : A Repetitive Transcranial Magnetic Stimulation Study. *Journal of Cognitive Neuroscience*, 18(7):1147–1155, 2006. 116

L.D. Rosenblum, J.A. Johnson et H.M. Saldaña : Point-light facial displays enhance comprehension of speech in noise. *Journal of Speech, Language and Hearing Research*, 39(6):1159–1170, 1996. 45

L.D. Rosenblum et H.M. Saldaña : An audiovisual test of kinematic primitives for visual speech perception. *Journal Of Experimental Psychology : Human Perception And Performance*, 22:318–331, 1996. 45

E. Rubin : Figure and ground. *In* D.C. Beardslee et M. Wertheimer, éditeurs : *Readings in perception*, pages 194–203. Van Nostrand : Princeton, NJ, 1958. 3

N. Rubin et J.M. Hupé : Dynamics of perceptual bi-stability : plaids and binocular rivalry compared. *In* A. Alais et R. Blake, éditeurs : *Binocular Rivalry*, pages 137–154. Cambridge : MIT Press, 2005. 3, 69

J.R. Saffran, R.N. Aslin et E.L. Newport : Statistical learning by 8-month-old infants. *Science*, 274(5294):1926–1928, 1996a. 146

J.R. Saffran, E.L. Newport et R.N. Aslin : Word segmentation : The role of distributional cues. *Journal of Memory and Language*, 35(4):606–621, 1996b. 146

M. Sams, R. Möttönen et T. Sihvonen : Seeing and hearing others and oneself talk. *Cognitive Brain Research*, 23(2-3):429–435, 2005. 50

M. Sato : *Représentations Verbales Multistables en Mémoire de Travail : Vers une Perception Active des Unités de Parole*. Thèse de doctorat, Institut National Polytechnique de Grenoble, 2004. 1, 77, 86, 87, 111

M. Sato, M. Baciu, H. Lœvenbruck, J.L. Schwartz, M.A. Cathiard, C. Segebarth et C. Abry : Multistable representation of speech forms : A functional MRI study of verbal transformations. *NeuroImage*, 23(3):1143–1151, 2004. 77, 79, 84, 113, 114, 117, 129, 131, 187, 202

M. Sato, J.L. Schwartz, C. Abry, M.A. Cathiard et H. Lœvenbruck : Multistable syllables as enacted percepts : A source of an asymmetric bias in the verbal transformation effect. *Perception and Psychophysics*, 68(3):458, 2006. 75, 76, 93, 111, 129, 131

M. Sato, P. Tremblay et V.L. Gracco : A mediating role of the premotor cortex in phoneme segmentation. *Brain and Language*, 111(1):1–7, 2009. 40

M. Sato, N. Vallee, J.L. Schwartz et I. Rousset : A perceptual correlate of the labial-coronal effect. *Journal of Speech, Language, and Hearing Research*, 50 (6):1466–1480, 2007. 76, 111, 129, 131, 184

J.L. Schwartz : La parole multisensorielle : Plaidoyer, problèmes, perspective. *In Actes des XXVème Journées d'Étude sur la Parole*, pages 11–18, 2004. 45, 46

J.L. Schwartz : A reanalysis of McGurk data suggests that audiovisual fusion in speech perception is subject-dependent. *The Journal of the Acoustical Society of America*, 127(3):1584–1594, 2010. 47

J.L. Schwartz, C. Abry, L.J. Boë et M. Cathiard : Phonology in a theory of perception-for-action-control. *In* J. Durand et B. Laks, éditeurs : *Phonetics, phonology and cognition*, pages 244–280. Oxford University Press, 2002. 34, 84

J.L. Schwartz, A. Basirat, L. Ménard et M. Sato : The perception for action control theory (PACT) : a perceptuo-motor theory of speech perception. *Journal of Neurolinguistics*, sous presse. 34, 41, 129, 209

J.L. Schwartz, F. Berthommier et C. Savariaux : Seeing to hear better : evidence for early audio-visual interactions in speech identification. *Cognition*, 93 (2):69–78, 2004. 44, 47, 129, 131

J.L. Schwartz, L.J. Boë et C. Abry : Linking the dispersion-focalization theory (DFT) and the maximum utilization of the available distinctive features (MUAF) principle in a perception-for-action-control theory (PACT). *In* M.J. Solé, P.S. Beddor et Ohala M., éditeurs : *Experimental approaches to phonology*, pages 104–124. Oxford University Press, 2007. 34, 35, 84

J.L. Schwartz et L. Ménard : Perceptuo-motor biases in the perceptual organization of the height feature in french vowels. soumis. 35

J.L. Schwartz, J. Robert-Ribes et P. Escudier : Ten years after Summerfield : A taxonomy of models for audio-visual fusion in speech perception. *In* R. Campbell, B. Dodd et D. Burnham, éditeurs : *Hearing by eye II : Advances in the psychology of speechreading and auditory-visual speech*, pages 85–108. UK : Psychology Press, 1998. 45, 47

J.L. Schwartz, M. Sato et L. Fadiga : The common language of speech perception and action : a neurocognitive perspective. *Revue Française de Linguistique Appliquée-Communiquer par la parole : des processus complexes*, 13:9–22, 2008a. 188

J.L. Schwartz, N. Vallée et S. Kandel : Hearing the tongue and lips of vowel gestures : a new differential paradigm. *In Acoustical Society of America Conference/International Conference on Acoustics*, 2008b. 41

S.K. Scott et I.S. Johnsrude : The neuroanatomical and functional organization of speech perception. *Trendsin Neurosciences*, 26(2):100–107, 2003. 36

S.K. Scott, C. McGettigan et F. Eisner : A little more conversation, a little less action-candidate roles for the motor cortex in speech perception. *Nature Reviews Neuroscience*, 10(4):295–302, 2009. 41, 42

K. Sekiyama et Y. Tohkura : Inter-language differences in the influence of visual cues in speech perception. *Journal of Phonetics*, 21(4):427–444, 1993. 47

D. Senkowski, T.R. Schneider, J.J. Foxe et A.K. Engel : Crossmodal binding through neural coherence : implications for multisensory processing. *Trends in Neurosciences*, 31(8):401–409, 2008. 12, 13, 14, 52, 53

M.N. Shadlen et W.T. Newsome : Neural basis of a perceptual decision in the parietal cortex (area LIP) of the rhesus monkey. *Journal of Neurophysiology*, 86 (4):1916–1936, 2001. 79

L.C. Shoaf et M.A. Pitt : Does node stability underlie the verbal transformation effect ? A test of node structure theory. *Perception and Psychophysics*, 64(5):795–803, 2002. 74, 86, 158

W. Singer et C.M. Gray : Visual feature integration and the temporal correlation hypothesis. *Annual Review of Neuroscience*, 18(1):555–586, 1995. 9

J.I. Skipper, H.C. Nusbaum et S.L. Small : Listening to talking faces : motor cortical activation during speech perception. *NeuroImage*, 25(1):76–89, 2005. 41, 50, 111, 129, 188

J.I. Skipper, V. van Wassenhove, H.C. Nusbaum et S.L. Small : Hearing lips and seeing voices : How cortical areas supporting speech production mediate audiovisual speech perception. *Cerebral Cortex*, 17(10):2387–2399, 2007. 41, 50, 51, 111, 129, 188

P.L. Smith et R. Ratcliff : Psychology and neurobiology of simple decisions. *Trends in Neurosciences*, 27(3):161–168, 2004. 80

J.R. Smythies : Requiem for the identity theory. *Inquiry*, 37(3):311–329, 1994a. 2

J.R. Smythies : *The walls of Plato's cave : the science and philosophy of (brain, consciousness, and perception)*. UK : Avebury, 1994b. 2

J.S. Snyder, O.L. Carter, S.K. Lee, E.E. Hannon et C. Alain : Effects of context on auditory stream segregation. *Journal of Experimental Psychology*, 34 (4):1007–1016, 2008. 21

K.A. Snyder, R.S. Calef, M.C. Choban et E. Scott Geller : Effects of word repetition and presentation rate on the frequency of verbal transformations : Support for habituation. *Bulletin of the Psychonomic Society*, 31(2):91–93, 1993. 73

S. Soto-Faraco, J. Navarra et A. Alsius : Assessing automaticity in audiovisual speech integration : evidence from the speeded classification task. *Cognition*, 92 (3):13–23, 2004. 47

R.B. Stein : The frequency of nerve action potentials generated by applied currents. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, 167 (6):64–86, 1967. 179

J.J. Stekelenburg et J. Vroomen : Neural correlates of multisensory integration of ecologically valid audiovisual events. *Journal of Cognitive Neuroscience*, 19 (12):1964–1973, 2007. 49

P. Sterzer et A. Kleinschmidt : A neural basis for inference in perceptual ambiguity. *Proceedings of the National Academy of Sciences*, 104(1):323–328, 2007. 65, 66

P. Sterzer, A. Kleinschmidt et G. Rees : The neural bases of multistable perception. *Trends in Cognitive Science*, 13(7):310–318, 2009. 56, 59, 65, 66, 67, 130, 188

P. Sterzer et G. Rees : A neural basis for percept stabilization in binocular rivalry. *Journal of Cognitive Neuroscience*, 20(3):389–399, 2008. 66

K.N. Stevens : The quantal nature of speech : Evidence from articulatory-acoustic data. *In* E.E. David et J.R. & P.B. Dens, éditeurs : *Human communication : A unified view*, pages 51–66. McGraw-Hill Companies, 1972. 32

K.N. Stevens : On the quantal nature of speech. *Journal of Phonetics*, 17(1):3–45, 1989. 32, 33

K.N. Stevens et S.E. Blumstein : Invariant cues for place of articulation in stop consonants. *The Journal of the Acoustical Society of America*, 64(5):1358–1368, 1978. 32

T. Strauss, D. Mirman et J. Magnuson : Computational modeling of spoken langauge processing. Transparents de tutorial. The 30th Annual Conference of the Cognitive Science Society, 2008. 171

T.J. Strauss, H.D. Harris et J.S. Magnuson : jTRACE : A reimplementation and extension of theTRACE model of speech perception and spoken word recognition. *Behavior Research Methods*, 39(1):19–30, 2007. 150, 165

W.H. Sumby et I. Pollack : Visual contribution to speech intelligibility in noise. *The Journal of the Acoustical Society of America*, 26(2):212–215, 1954. 42, 43

Q. Summerfield : Use of visual information for phonetic perception. *Phonetica*, 36(4-5):314, 1979. 43, 44

Q. Summerfield : Audio-visual speech perception, lipreading and artificial stimulation. *In* M. E. Lutman et M. P. Haggard, éditeurs : *Hearing Science and Hearing Disorders*, pages 131–182. London : Academic Press, 1983. 98

Q. Summerfield, A. MacLeod, M. McGrath et M. Brooke : Lips, teeth, and the benefits of lipreading. *In* A.W. Young et H.D. Ellis, éditeurs : *Handbook of research on face processing*, pages 223–233. North-Holland, 1989. 43

R. Sun : *Duality of the mind : A bottom-up approach toward cognition.* Lawrence Erlbaum Associates, 2002. 143

R. Sun : *The Cambridge handbook of computational psychology.* Cambridge University Press New York, NY, USA, 2008. 139

C. Tallon-Baudry et O. Bertrand : Oscillatory gamma activity in humans and its role in object representation. *Trends in Cognitive Sciences*, 3(4):151–162, 1999. 10

C. Tallon-Baudry, O. Bertrand, C. Delpuech et J. Pernier : Stimulus specificity of phase-locked and non-phase-locked 40 Hz visual responses in human. *Journal of Neuroscience*, 16(13):4240–4249, 1996. 9, 10

C. Tallon-Baudry, O. Bertrand, C. Delpuech et J. Pernier : Oscillatory gamma-band (30-70 Hz) activity induced by a visual search task in humans. *Journal of Neuroscience*, 17(2):722, 1997. 10, 11

P. Teissier, J. Robert-Ribes, J.L. Schwartz et A. Guerin-Dugue : Comparing models for audiovisual fusion in a noisy-vowelrecognition task. *IEEE Transactions on Speech and Audio Processing*, 7(6):629–642, 1999. 46

K. Tiippana, T.S. Andersen et M. Sams : Visual attention modulates audiovisual speech perception. *European Journal of Cognitive Psychology*, 16(3):457–472, 2004. 47

F. Tong et S.A. Engel : Interocular rivalry revealed in the human cortical blindspot representation. *Nature*, 411(6834):195–199, 2001. 59

F. Tong, M. Meng et R. Blake : Neural bases of binocular rivalry. *Trends in Cognitive Sciences*, 10(11):502–511, 2006. 3

T.C. Toppino et G.M. Long : Selective adaptation with reversible figures : Don't change that channel. *Perception & Psychophysics*, 42(1):37–48, 1987. 57

T.C. Toppino et G.M. Long : Top-down and bottom-up processes in the perception of reversible figures : Toward a hybrid model. *In* N. Ohta, C.M. Macleod et B. Uttl, éditeurs : *Dynamic Cognitive Processes*, pages 37–58. Springer Tokyo, 2005. 57, 62

J.T. Townsend et J. Busemeyer : Dynamic representation of decision-making. *In* R.F. Port et T. van Gelder, éditeurs : *Mind as motion : Explorations in the dynamics of cognition*, pages 101–120. The MIT Press, 1995. 141

A. Treisman : Features and objects : The fourteenth Bartlett memorial lecture. *Quarterly Journal of Experimental Psychology : Human Experimental Psychology*, 40(2):201–237, 1988. 14

A. Treisman : Feature binding, attention and object perception. *Philosophical Transactions of the Royal Society B : Biological Sciences*, 353(1373):1295–1306, 1998. 15

A. Treisman et H. Schmidt : Illusory conjunctions in the perception of objects. *Cognitive Psychology*, 14(1):107–141, 1982. 14

A.M. Treisman et G. Gelade : A feature-integration theory of attention. *Cognitive Psychology*, 12(1):97–136, 1980. 14

R. Van Ee, L.C.J. Van Dam et G.J. Brouwer : Voluntary control and the dynamics of perceptual bi-stability. *Vision Research*, 45(1):41–55, 2005. 62

L.P.A.S. Van Noorden : *Temporal Coherence in the Perception of Tone Sequences*. Thèse de doctorat, Institute for Perceptual Research, 1975. 18, 68

V. van Wassenhove, K.W. Grant et D. Poeppel : Visual speech speeds up the neural processing of auditory speech. *Proceedings of the National Academy of Sciences*, 102(4):1181–1186, 2005. 41, 50, 188

V. van Wassenhove, K.W. Grant et D. Poeppel : Temporal window of integration in bimodal speech. *Neuropsychologia*, 45(3):598–607, 2007. 48

F. Varela, J.P. Lachaux, E. Rodriguez et J. Martinerie : The brainweb : phase synchronization and large-scale integration. *Nature Reviews Neuroscience*, 2(4):229–239, 2001. 84

J. Vliegen et A.J. Oxenham : Sequential stream segregation in the absence of spectral cues. *The Journal of the Acoustical Society of America*, 105(1):339–346, 1999. 19

M.R. Warren et D.M. Meyers : Effects of listening to repeated syllables : Category boundary shifts versus verbal transformation. *Journal of Phonetics*, 15(2):169–181, 1987. 73

P. Warren et W. Marslen-Wilson : Continuous uptake of acoustic cues in spoken word recognition. *Perception & Psychophysics*, 41(3):262–275, 1987. 144

R.M. Warren : Illusory changes in repeated words : Differences between young adults and the aged. *The American Journal of Psychology*, 74(4):506–516, 1961a. 72

R.M. Warren : Illusory changes of distinct speech upon repetition–the verbal transformation effect. *British Journal of Psychology*, 52:249–258, 1961b. 3, 72

R.M. Warren : Auditory illusions and their relation to mechanisms normally enhancing accuracy of perception. *Journal of the Audio Engineering Society*, 31 (9):623–629, 1983. 73

R.M. Warren et R.L. Gregory : An auditory analogue of the visual reversible figure. *The American Journal of Psychology*, 71(3):612–613, 1958. 3, 71

R.M. Warren et R.P. Warren : Auditory illusions and confusions. *Scientific American*, 223(12):30–36, 1970. 4, 72, 73

K.E. Watkins, A.P. Strafella et T. Paus : Seeing and hearing speech excites the motor system involved in speech production. *Neuropsychologia*, 41(8):989–994, 2003. 31

H.R. Wilson, R. Blake et S.H. Lee : Dynamics of travelling waves in visual perception. *Nature*, 412(6850):907–910, 2001. 60

S. Windmann, M. Wehrmann, P. Calabrese et O. Güntürkün : Role of the prefrontal cortex in attentional control over bistable vision. *Journal of Cognitive Neuroscience*, 18(3):456–471, 2006. 66

D. Wolpert et Z. Ghahramani : Bayes rule in perception, action and cognition. *In* Gregory R.L., éditeur : *The Oxford Companion to the Mind*. Oxford University Press, 2005. 140

T. Womelsdorf et P. Fries : The role of neuronal synchronization in selective attention. *Current Opinion in Neurobiology*, 17(2):154–160, 2007. 11

T.M. Wright, K.A. Pelphrey, T. Allison, M.J. McKeown et G. McCarthy : Polysensory interactions along lateral temporal regions evoked by audiovisual speech. *Cerebral Cortex*, 13(10):1034–1043, 2003. 48

S. Zeki : *A vision of the brain*. Blackwell Oxford, 1992. 1

Y.H. Zhou, J.B. Gao, K.D. White, I. Merk et K. Yao : Perceptual dominance time distributions in multistable visual perception. *Biological Cybernetics*, 90 (4):256–263, 2004. 57, 58

Résumé : La problématique traitée dans le cadre de cette thèse est celle du liage perceptif en parole, ce qui amène à l'étude des principes de l'analyse de scène de parole (en analogie avec l'analyse de scène auditive). La littérature sur la perception de la parole met en évidence que ces principes sont en partie différents de ceux de l'analyse de scène auditive. Notre objectif dans cette thèse est de mieux caractériser ces principes « spécifiques à la parole ». Le paradigme que nous utilisons est celui de l'Effet de Transformation Verbale. À travers une série d'expériences comportementales et une étude en EEG intracrânienne, nous suggérons que cette organisation est basée sur des principes multisensoriels et perceptuo-moteurs. Nous mettons en œuvre quelques uns de ces mécanismes au sein du modèle psycholinguistique TRACE. Du point de vue théorique, les résultats obtenus dans le cadre de cette thèse s'inscrivent dans PACT (Théorie de la Perception pour le Contrôle de l'Action).

Mot-clés : Organisation perceptive de la parole, Effet de Transformation Verbale, liage perceptif, perception multisensorielle de la parole, lien perceptuo-moteur.

Abstract: The aim of this thesis is to study the principles of speech scene analysis (in analogy to auditory scene analysis). The literature on speech perception suggests that these principles are partly different from those underlying auditory scene analysis. We use the Verbal Transformation Effect to investigate these "speech specific" mechanisms. The behavioral and neuroimaging results obtained in this work suggest that the perceptuo (multisensory)-motor processes are involved in the perceptual organization of speech. We implement some of theses mechanisms in the TRACE model of speech perception. Our results can be understood within the framework of PACT (Perception for Action Control Theory), suggesting a link between speech perception and production systems in the perceptual organization of speech.

Keywords: Perceptual organization of speech, Verbal Transformation Effect, perceptual binding, multisensory speech perception, perceptuo-motor link.