Chapitre 1

Appariement entre signal acoustique et représentations lexicales

APPARIEMENT ENTRE SIGNAL ACOUSTIQUE ET REPRESENTATIONS LEXICALES

1. Du signal acoustique aux représentations linguistiques

La parole est un signal physique produit par la mise en mouvement d'un nombre considérable d'organes (poumons, cordes vocales, langue, lèvres, etc.). Les modifications rapides de la configuration des articulateurs dans le tractus vocal donnent lieu à des frottements qui, selon le modèle source-filtre ('source-filter model' ; Fant, 1960; Flanagan, 1972), produisent une onde. La répercussion de cette onde sur les diverses parois (buccale, nasale, …) du conduit vocal provoque l'amplification de certaines parties du spectre (ce qui équivaut à l'application d'un filtre). On appelle *formants* les composantes fréquentielles de la voix qui subissent une amplification. Certains phonèmes peuvent correspondre à une amplification de la quasi-totalité du spectre perceptible par l'humain ou à une évolution rapide de ces composantes, c'est le cas de la plupart des consonnes. Les voyelles quant à elles sont caractérisées par une structure spectrale relativement stable dans le temps. Le signal acoustique résultant de ces phénomènes mécaniques se caractérise par une organisation spectro-temporelle complexe qui, par l'intermédiaire du milieu de transmission (en général aérien), est transmise au système auditif périphérique par des

phénomènes de transduction mécanico-électriques et transformée en un *percept auditif* (Delgutte, 1987). Afin d'aboutir à un *percept linguistique*, le système perceptif doit apparier cette image auditive avec des représentations linguistiques abstraites phonémiques ou phonologiques, forme sous laquelle seraient représentés les mots dans le lexique mental.

1.1. Du signal acoustique à l'image auditive

Du fait de sa complexité, plusieurs étapes fonctionnelles de traitement sont nécessaires avant d'aboutir à une représentation linguistique d'un signal de parole. La première étape, réalisée par les organes de l'oreille interne, consiste à analyser ce signal acoustique (transformé en un signal électrique par les récepteurs sensoriels de l'oreille externe) afin d'en extraire une information qui pourra être utilisée par le système nerveux central. L'un des organes les plus importants pour le traitement de la parole et l'accès à des représentations linguistiques est la cochlée, qui est supportée par l'organe de Corti. Apte à décoder les différentes fréquences d'un son en temps réel (par l'intermédiaire des cellules ciliées internes) et dotée d'une sélectivité fréquentielle considérable (en raison de l'action des cellules ciliées externes), elle constitue un banc de filtres qui fournit au système auditif une analyse fréquentielle *et* temporelle des signaux acoustiques sensiblement comparable à ce que l'on peut observer sur un spectrogramme (Moore, 1997). On considère classiquement que l'information primordiale pour la perception de la parole est constituée par la sortie de ce filtre, qui fournit en temps réel une information sur l'évolution

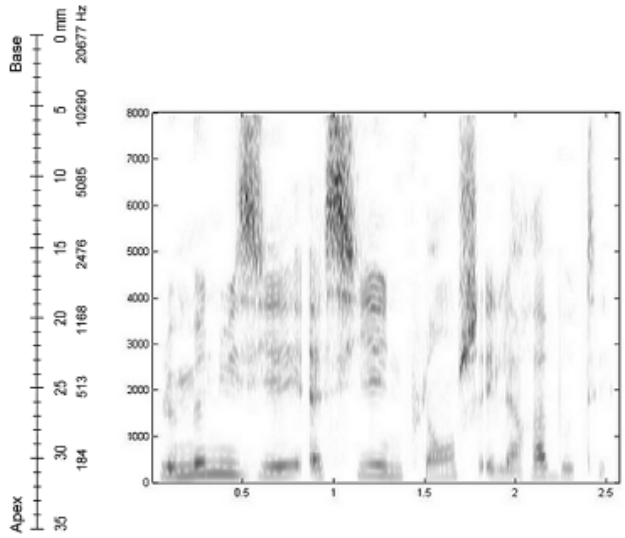

Figure 1 : Illustration du codage tonotopique effectué par la cochlée schématiquement représentée sur la gauche du graphique. Le traitement effectué peut en partie s'assimiler à une analyse en temps réel de l'enveloppe spectrale, le signal de sortie correspondant alors approximativement à ce que l'on observe sur un spectrogramme.

des différentes composantes fréquentielles du signal en fonction du temps (cf. Figure 1). Chaque phonème de la langue peut être décrit par des caractéristiques spectrales spécifiques (Stevens, 1998). L'une des tâches du système d'identification de la parole consiste donc à apparier ces patterns spectraux à des représentations phonétiques. Selon Klatt (Lexical Access From Spectra, 1979 ; 1989), le système de traitement de la parole comparerait chaque spectre à court-terme avec des patterns spectraux prototypiques et rechercherait celui qui correspond le mieux au calcul effectué sur le signal. Chaque phonème de la langue serait lié à un nombre considérable de patterns prototypiques. L'appariement entre le produit des calculs et les représentations spectrales prototypiques stockées en mémoire permettrait alors d'avoir accès à des représentations phonétiques. Stevens (1960 ; 1996) accorde une importance essentielle au lien entre perception et production pour rendre compte du traitement de la parole mais se démarque des propositions avancées dans le cadre des théories motrice (Liberman & Mattingly, 1985) et directe-réaliste (Fowler, 1986). Pour dériver une représentation phonétique du signal acoustique, il propose de relier le système perceptif à un synthétiseur vocal. Ce synthétiseur générerait des spectres à court terme à partir d'un ensemble de règles de production du signal de parole. Ces règles de production permettraient de relier une représentation phonétique à un ensemble de paramètres articulatoires. Le système comparerait le spectre calculé avec chacun des spectres produits par le synthétiseur. Il évaluerait alors la distance entre le spectre calculé et le spectre produit afin de modifier les paramètres du synthétiseur. Cette étape serait répliquée jusqu'à ce que le synthétiseur produise un spectre suffisamment proche du spectre perçu. Il serait alors très facile d'identifier le phonème correspondant au spectre traité puisque le spectre généré par le synthétiseur correspondrait à un programme articulatoire (donc phonétique) connu.

Mais cette représentation spectrale du signal acoustique ne constitue pas la seule information utile à ces processus d'appariement entre signal acoustique et représentations linguistiques (Van Tasell, Soli, Kirby, & Widin, 1987 ; Shannon, Zeng, Kamath, Wygonski, & Ekelid, 1995). Le rôle de l'enveloppe d'intensité (des modulations temporelles à long terme) dans l'identification des phonèmes a ainsi pu être mis en évidence grâce à des méthodes de traitement du signal qui consistent à présenter à des auditeurs une bande de bruit lissée avec la forme de l'enveloppe d'intensité d'un signal de parole (cf. Figure 2). Dans cette situation il reste possible, malgré l'absence de toute information spectrale fine, d'accéder à une représentation linguistique du signal. Par exemple, après un entraînement intense, on peut observer des taux d'identification correcte de phrases qui avoisinent les 50 % (Shannon et al., 1995). Dans une tâche de choix forcé à 16 alternatives dans laquelle les auditeurs doivent identifier la consonne

médiane de logatomes[1] VCV[2], les auditeurs sont en mesure d'atteindre des taux d'identification correcte de l'ordre de 20 % (Apoux, Berthommier, Bacri, & Lorenzi, 1998) alors que le taux de réponses au hasard correspond à 6,25 %[3]. Que l'on considère la situation de compréhension de phrases ou d'identification de logatomes, ces stimuli ne fournissent au système auditif aucune information spectrale puisqu'ils ne sont en fait que des séquences de bruit blanc dont l'intensité évolue au cours du temps. Mais les modulations d'intensité sonore de ce bruit permettent d'accéder au moins partiellement à une représentation phonétique du signal. Cette aptitude des auditeurs humains à utiliser les indices fournis par les modulations temporelles à long terme présentes dans l'enveloppe constitue la preuve que la forme de l'enveloppe d'intensité d'un signal acoustique apporte des informations importantes pour l'identification de la parole -en tout cas pour la détection de certains indices linguistiques-, donc pour l'appariement entre signal acoustique et représentations linguistiques.

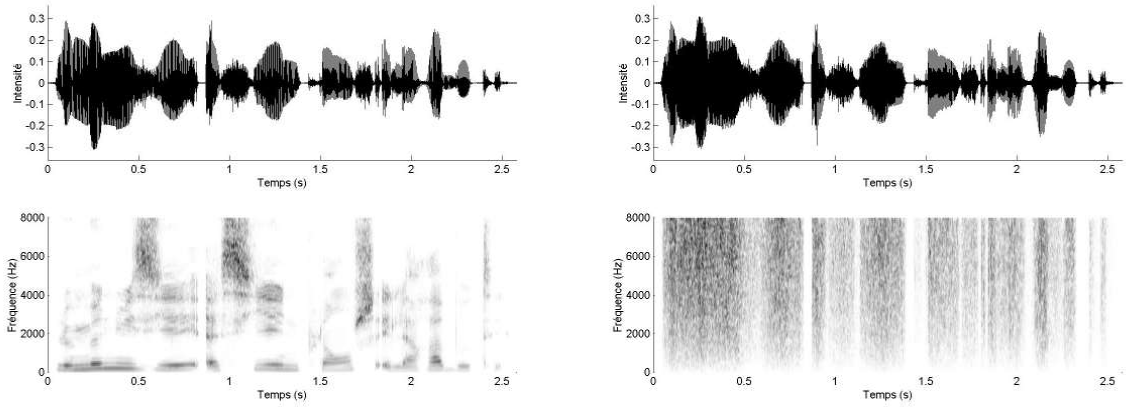

Figure 2 : La phrase 'Le menuisier a scié une planche et l'a rabotée' ; à gauche dans sa forme originale et à droite après suppression des composantes spectrales fines.

Il a également été mis en évidence qu'il n'est pas nécessaire, pour aboutir à un percept structuré, de disposer d'une information précise quant au moment d'occurrence des différents événements fréquentiels du signal de parole dans le temps. Si l'on applique une désynchronisation artificielle au signal de parole (cf. Figure 3), on observe une aptitude des auditeurs à conserver une compréhension correcte des messages linguistiques malgré des taux de réverbération considérables entre bandes de fréquence (Greenberg & Arai, 1998). Cette

[1] Un logatome est une séquence de parole courte et sans signification, comme par exemple /aka/.

[2] VCV : Séquence de parole présentant une structure Voyelle - Consonne - Voyelle.

[3] La variabilité des taux de performance en fonction de la tâche est certainement déterminée par le type d'entraînement auquel sont soumis les participants. Dans les expériences de Shannon et al. (1995), les participants avaient déjà entendu les phrases dans leur forme originale. La quantité de phonèmes prononcés doit certainement rendre plus faciles les processus d'appariement entre le signal perçu et la représentation linguistique stockée en mémoire. Plus que les indices temporels disponibles dans l'enveloppe globale, il est possible d'envisager qu'en

➥

désynchronisation constitue l'une des manifestations des phénomènes de réverbération qui sont en fait bien réels dans les signaux auxquels nous sommes confrontés dans la plupart des situations naturelles de communication[4] et consiste à introduire décalages de phase entre différentes bandes de fréquence du signal de parole. Greenberg & Arai (1998) montrent que ces ruptures de synchronie ne gênent pas considérablement l'identification de phrases pour un décalage de phase inférieur à 150 ms. Ce phénomène met en évidence l'existence d'une fenêtre d'intégration temporelle qui permettrait au système auditif de structurer un signal acoustique malgré les distorsions qui peuvent être appliquées au signal acoustique dans de nombreuses situations. Cette capacité d'intégration dans l'accès à des représentations linguistiques montre qu'il n'est pas nécessaire de dériver du signal acoustique une représentation spectrographique précise et que certains événements acoustiques peuvent être décalés dans le temps lorsqu'ils parviennent au système auditif périphérique sans pour autant déstabiliser les percepts qui en découlent.

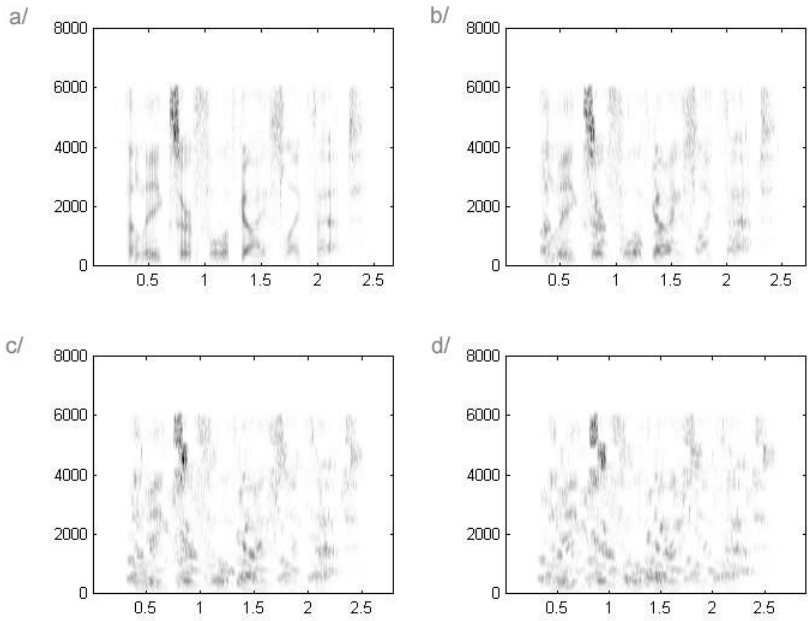

Figure 3 : Désynchronisation artificielle de la phrase 'The leagues are full of guys like that' (d'après Greenberg & Arai, 1998). Chaque spectrogramme correspond à une durée moyenne de désynchronisation de a/ 0 ms (signal original), b/ 40 ms, c/ 120 ms, d/ 220 ms. Les stimuli correspondants sont accessibles sur le site `http://www.ICSI.berkeley.edu/~steveng`

réalité les modulations d'amplitude de l'enveloppe sont traitées à l'intérieur de chaque bande de fréquence (cf. le rôle des cartes de Modulation d'Amplitude ou *Amplitude Modulation map*s, Greenberg & Araï, 1998).

[4] Ceci s'explique par le fait que, dans un environnement ouvert, chaque fréquence du spectre est renvoyée avec une vélocité différente par les divers objets qui sont présents (en raison de leur taille, de leur forme et des caractéristiques du matériau qui les constitue).

On peut donc affirmer que plusieurs types d'images auditives (évolution des composantes fréquentielles en fonction du temps, forme de l'enveloppe d'intensité, intégration temporelle sur des segments d'image auditive de type spectrographique) peuvent être pris en compte par les processus de traitement auditif pour l'accès à des représentations linguistiques.

1.2. L'appariement entre représentations auditives et phonétiques

Dès lors que l'une ou plusieurs de ces images auditives ont été générées par le système auditif périphérique, le système cognitif va devoir apparier chaque portion temporelle de ces images avec des représentations linguistiques abstraites (matrices de traits, phonèmes, diphones, syllabes). Or les signaux de parole que doit traiter le système auditif humain présentent une variabilité importante (Blumstein, 1986 ; Klatt, 1986). Il est possible de définir deux types de variabilité. Nous appelons variabilité intrinsèque les formes de variabilité qui affectent le signal de parole sans être déterminées par le contexte phonétique (hauteur de la voix, vitesse d'élocution). La variabilité extrinsèque correspond aux formes de variabilité qui sont déterminées par le contexte. Les problèmes posés par ces diverses sources de variabilité pourraient éventuellement trouver leur solution dans des procédures communes (par exemple l'analyse en ondelettes, cf. infra.).

1.2.1. Variabilité intrinsèque

L'une des principales difficultés posées par cette étape est liée à l'importante variabilité des productions possibles pour un même message linguistique. Cette variabilité est à la fois fréquentielle et temporelle. Du fait des différences dans la taille des cavités articulatoires des locuteurs (liées à leur âge, à leur sexe, etc.), leurs fréquences de résonance diffèrent d'un individu à l'autre. Les filtres qui vont amplifier ou atténuer certaines fréquences du signal source sont donc différents. Chaque locuteur génère ainsi des signaux de parole avec une hauteur de voix (fréquence fondamentale ou F_0) différente ; ce phénomène induit également une répartition variable des formants sur l'échelle des fréquences. Par ailleurs, un locuteur peut parler plus ou moins rapidement en fonction des situations et changer de vitesse d'élocution à l'intérieur d'un même énoncé. Cette variabilité dans l'organisation spectrale et temporelle des sons de parole n'entrave cependant pas la stabilité perceptive qui permet à un auditeur d'entendre des phonèmes formes stables.

L'une des méthodes qui ont été proposées afin de rendre compte de la capacité du système cognitif humain à gérer la variabilité du signal de parole consiste à effectuer une normalisation spectrale et / ou temporelle. Cette méthode repose sur le principe de l'appariement entre un

signal physique de forme variable et un référent. Par des méthodes mathématiques qui consistent à transformer la représentation spectrographique du signal à la fois dans le domaine spectral et temporel, on peut apparier des signaux de parole variant sur ces deux dimensions avec une représentation normalisée. Ce type de procédure pose cependant le problème des méthodes adéquates qui permettent d'*identifier* le signal référent pertinent. On trouve dans le domaine des travaux sur la vision des propositions alternatives telles les *cônes généralisés* ('generalized cones', Marr, 1982 p.223) ou les *géons* ('geometric ions', Biederman, 1987). Le principe consiste à utiliser un nombre restreint de formes (un *cône* chez Marr, diverses formes de base nommées *géons* chez Biederman) pour décrire ou définir une scène visuelle. Si l'on prend l'exemple des *cônes généralisés*, une forme conique peut être utilisée pour reproduire intégralement une scène visuelle par la combinaison d'un nombre considérable de ces unités. En faisant varier leur taille et leur disposition respectives, on peut générer une reproduction de la scène visuelle réelle. Il reste alors à identifier en mémoire le ou les objets qui correspondent à cette combinaison de cônes. Le problème de la normalisation -donc de l'identification adéquate de l'objet sur la base duquel on effectuera cette normalisation- ne se pose pas puisque la taille des cônes contribuant à l'image n'a d'importance que relative : c'est la configuration des cônes entre eux qui permet de récupérer l'objet en mémoire. Dans le domaine du traitement de la parole, des propositions similaires tendent à voir le jour actuellement avec l'utilisation des transformations en ondelettes ('wavelets transformation', Graps, 1995) comme une alternative à la Transformée de Fourier qui est utilisée actuellement pour aboutir à une représentation spectrographique, mais le recours à ces procédures se limite encore aux travaux effectués en Reconnaissance Automatique de la Parole, et rien n'a été proposé jusqu'à maintenant, comme l'avaient fait Marr (1982) ou Biederman (1987) pour la vision, dans le cadre de la description d'un modèle du fonctionnement cognitif appliqué au traitement de la parole.

1.2.2. Variabilité extrinsèque

Les difficultés posées par l'appariement entre image(s) auditive(s) et représentations abstraites proviennent également des contraintes inhérentes à la coproduction de segments proches. En effet, si l'on admet que les représentations linguistiques auxquelles devra être apparié le signal acoustique sont des phonèmes -donc des unités segmentales discrètes-, ce signal n'est pas, du fait de ce que l'on appelle la *coarticulation*, constitué d'une séquence discrète de segments. Ainsi, chaque unité identifiée par un auditeur est extraite d'une portion de signal dont les caractéristiques dépendent aussi de la prononciation des segments qui l'environnent (cf. Figure 4). Ce phénomène de coarticulation est lié au mode de production de la parole qui consiste à préparer les mouvements articulatoires correspondant à un segment alors même que

l'on est en train d'articuler le segment qui précède (effets d'anticipation). Les effets coarticulatoires se manifestent aussi par persévérance ; ainsi, nous générons des mouvements destinés à produire un son alors que l'on n'a pas encore terminé de produire le son précédent. Ceci pose deux énigmes essentielles qui sont intimement liées l'une à l'autre : la première est celle de la segmentation du signal en unités dont la taille correspond à un segment phonémique. Dans une séquence CV, ce phénomène de coarticulation induit la présence simultanée d'informations correspondant à la consonne et d'autres correspondant à la voyelle. Comment découper le signal en segments représentant chacun une unité puisque ce signal n'est pas discret ? La seconde est induite par le fait que ce caractère coarticulatoire des sons de parole génère une extrême variabilité dans la réalisation des sons en fonction de leur voisinage. Les formes acoustiques correspondant au phonème /g/ ne sont pas les mêmes selon qu'il est suivi d'un /i/ ou d'un /a/. Comment, dès lors, proposer des procédures d'appariement entre image auditive et représentation linguistique puisque les manifestations acoustiques des phonèmes sont si variables ?

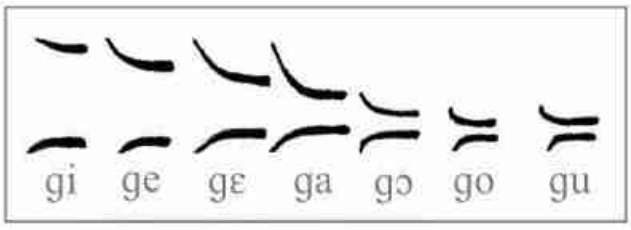

Figure 4 : Schéma représentant le phénomène de coarticulation. On peut voir l'influence des voyelles sur la prononciation du phonème /g/ (d'après Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967).

Outre les procédures d'appariement spectral et temporel d'un signal donné à une forme auditive référente, les études portant sur la compréhension de la parole doivent donc expliquer comment l'image auditive -une fois normalisée ou décrite à partir d'indices non-sensibles aux variations spectro-temporelles- peut être appariée à des représentations phonétiques. Pour cela, on a cherché à décrire un certain nombre de phénomènes stables dans le signal afin de pouvoir prédire, à partir d'un événement acoustique, le segment phonétique qui lui correspond. Dans cette optique, de nombreux travaux ont été effectués à partir des années cinquante dans lesquels on a tenté de mettre au jour des indices sur lesquels pourrait se fonder le système cognitif afin d'identifier avec certitude les différents phonèmes de la chaîne parlée. Ces indices constitueraient des invariants de relations acoustico-phonétiques qui permettraient de définir, avec une plus ou moins grande certitude, la relation entre forme acoustique et représentation linguistique. Les travaux réalisés dans ce domaine ont eu pour objectif de décrire des

caractéristiques invariables permettant de faire le lien entre une forme acoustique et un trait distinctif (voisement, mode d'articulation par exemple). Ils ont conduit à proposer plusieurs indices pour chaque type d'opposition phonémique. Par exemple, le trait de voisement est lié au délai qui sépare l'occlusion consonantique du début de vibration des cordes vocales mais aussi aux caractéristiques de l'enveloppe spectrale de la consonne (Summerfield & Haggard, 1977). Selon Stevens & Blumstein (1978), c'est la contribution conjuguée de la forme des transitions formantiques et du spectre à court terme de la consonne qui permet l'identification du mode d'articulation. Pour certaines valeurs de l'un des indices, il est nécessaire de disposer de l'autre indice pour identifier le mode d'articulation. Cette variabilité du signal de parole s'accompagne par ailleurs d'une redondance considérable des indices acoustiques (Stevens, Keyser, & Kawasaki, 1986). Il est donc possible, même si l'un des indices met le système en échec, d'aboutir à une identification correcte du phonème. La redondance des indices permet en effet d'appliquer des procédures probabilistes de prise de décision qui peuvent tolérer une certaine déviance par rapport à la forme phonétique idéale.

1.3. L'appariement entre représentations phonétiques et phonologiques

Indépendamment des difficultés générées par la variabilité du signal de parole en termes de distributions spectrale et temporelle des événements acoustiques pour la segmentation et l'identification phonétiques, les contraintes inhérentes aux langues posent des difficultés supplémentaires. Chaque langue est soumise à un ensemble de contraintes qui déterminent la réalisation effective des sons de parole (phénomènes d'allophonie, d'assimilation, contraintes phonotactiques portant sur les séquences de phonèmes admissibles). Supposons par exemple qu'un locuteur souhaite prononcer la phrase :

/ʒœpʀɑ̃ləmetʀoasizœʀpuʀaleʃeləmedəsɛ̃/

« Je prends le métro à six heures pour aller chez le médecin »

Si l'on admet que les mots sont représentés dans le lexique sous une forme phonologique abstraite correspondant à ce que les phonologues appellent la *représentation sous-jacente*, ce message a peu de chances d'être prononcé de manière aussi canonique. Au contraire, la phonologie et la phonétique montrent que de nombreuses contraintes vont transformer la forme canonique de ce message en une 'suite de segments phonétiques' qui, dans le cadre d'un modèle de compréhension de la parole, ne sont pas directement appariables avec les représentations phonologiques sous-jacentes. La plupart des locuteurs, en fonction bien sûr des conditions de production (lecture vs. parole spontanée), mais aussi du style de langue (soutenue ou pas), vont

aboutir à une forme phonétique qui différera considérablement de la représentation phonologique de départ. Cette séquence pourrait par exemple donner lieu à la suite phonétique suivante :

[ʃpʀɑ̃lmetʀoasizœʀpuʀaleʃelmɛtsɛ̃]

On notera la séquence initiale [ʃp] qui vient se substituer à la représentation phonologique /ʒœp/ par l'élision du /œ/ et l'assimilation de /ʒ/ à [ʃ]. Or il est couramment admis, malgré l'existence de positions alternatives qui fournissent des arguments convaincants[5] (cf. par exemple Goldinger, 1998), que l'unité de contact avec les représentations lexicales devrait être, dans un souci d'économie de traitement, la plus abstraite possible. En effet, si l'on parvient à représenter les mots stockés dans le lexique avec un nombre minimal d'unités de codage, on réalise une économie considérable par rapport à un stockage qui se ferait sous forme acoustique, auquel cas l'ensemble des exemplaires acoustiques de chaque mot devrait être stocké dans le lexique afin de pouvoir le reconnaître. Ainsi, si l'on admet que les mots doivent être représentés dans le lexique sous la forme la plus abstraite qui soit (c'est à dire la représentation phonologique sous-jacente proposée par la phonologie), l'existence de contraintes déterminant les modifications à apporter à la forme phonologique pour aboutir à une réalisation phonétique considérablement différente constitue un obstacle supplémentaire aux processus de compréhension de la parole, non seulement pour l'appariement entre représentations phonétiques et phonémiques (phénomènes d'allophonie nécessitant l'intervention de procédures spécifiques afin d'apparier les différentes variantes phonétiques d'un même phonème, par exemple les différents /r/ du français : [ʁ] et [r]) mais également pour la mise en correspondance des représentations phonémiques et phonologiques (assimilation du trait de voisement par exemple, modification de la qualité des voyelles en fonction de la structure syllabique, contraintes phonotactiques, …). Au-delà des problèmes posés au système perceptif pour l'appariement entre image auditive et représentation phonétique, les contraintes phonologiques introduisent donc des difficultés supplémentaires pour la compréhension de la parole. L'une des tâches du système de perception de la parole va donc consister, une fois le décodage acoustico-phonétique effectué, à apparier cette représentation phonétique avec une représentation phonologique adéquate afin d'être en mesure de contacter les représentations lexicales stockées en mémoire.

[5] Et qui démontrent qu'il est probablement nécessaire d'adopter une position intermédiaire entre un modèle dans lequel le stockage se ferait intégralement sous forme abstraite et un autre dans lequel l'ensemble des exemplaires acoustiques possibles serait représenté.

2. Des connaissances qui influencent la perception du signal ?

Comme nous venons de le voir, l'accès à une représentation linguistique du signal de parole s'avère particulièrement complexe et passe par la mise en œuvre de nombreuses étapes fonctionnelles de traitement. Afin de rendre compte de la possibilité d'appariement entre forme auditive et représentations phonétiques ou phonologiques, de nombreux auteurs envisagent l'utilisation, à l'intérieur du système de reconnaissance de la parole, de connaissances de haut niveau qui rétroagiraient sur des niveaux de représentation moins élaborés (représentation segmentale du signal de parole notamment). Ces connaissances porteraient sur les caractéristiques qui constituent la langue et qui se développent au cours de l'acquisition : le lexique et les contraintes phonologiques.

2.1. Le recours à des connaissances lexicales

Afin de pouvoir identifier correctement les phonèmes dans le signal de parole, l'une des solutions qui ont été envisagées pour faciliter l'appariement d'une image auditive avec des représentations phonémiques ou phonologiques consiste à faire intervenir des procédures de rétroaction des niveaux lexicaux vers les niveaux de représentation prélexicaux. Ainsi, des connaissances de haut niveau faciliteraient la tâche du système d'identification phonémique en guidant les choix perceptifs dans les situations problématiques. Les partisans d'une approche interactive (notamment McClelland & Elman, 1986) se sont heurtés à un courant autonome dont les représentants (Cutler, Mehler, Norris, & Segui, 1987) affirmaient l'indépendance entre niveaux de traitement de bas niveau et représentations plus élaborées. Nous décrivons ici certaines des données expérimentales qui ont été présentées comme reflétant des preuves du recours à des rétroactions lexicales. Une rapide description des interprétations alternatives qui ont été proposées est présentée pour chacun des effets. Ceci nous conduira à mettre au jour l'un des problèmes essentiels dans l'étude du rôle de certaines caractéristiques des langues dans les processus perceptifs : la confusion entre diverses variables qu'il peut s'avérer difficile de contrôler du fait même des particularités de la langue. La conscience de cette confusion sera essentielle pour notre travail sur le rôle des contraintes phonologiques dans les processus de segmentation lexicale.

2.1.1. Les données expérimentales

Quatre types d'effets seront présentés ici qui ont donné lieu à une discussion sur l'existence de processus rétroagissant des représentations lexicales vers les représentations

phonémiques : l'effet du statut lexical (1) sur la catégorisation de phonèmes ambigus, (2) sur les temps de détection de phonèmes, (3) sur la restauration phonémique (Warren, 1970) et (4) sur la compensation perceptive de la coarticulation (Elman & McClelland, 1988).

2.1.1.1. L'effet du statut lexical sur la catégorisation phonémique

L'un des effets interprété comme une preuve de l'interaction lexique-phonèmes est celui, mis en évidence par Ganong (1980), du statut lexical sur la catégorisation de phonèmes ambigus. Cet auteur a présenté à des auditeurs des séquences de parole CVC constituées d'une consonne occlusive initiale qui variait sur un continuum de voisement (temps d'attaque vocale -en anglais VOT (Voice Onset Time)- qui distingue /d/ de /t/ par exemple). Ces phonèmes ambigus étaient présentés dans des séquences de type '$^{d}_{t}$VC'[6] qui pouvaient donner lieu à un mot pour une extrémité du continuum (/t/ précédant /æsk/ donne *task*, 'tâche') et à un non-mot pour l'autre extrémité (/d/ précédant /æsk/ donne le non-mot *dask*). Le contexte variait afin de contrebalancer la consonne qui donnait lieu à l'extrémité lexicale. Ainsi, dans la condition opposée, le phonème ambigu était présenté dans le contexte /æʃ/ qui donnait lieu à un mot après le phonème /d/ (*dash*, 'tiret') et à un non-mot après /t/ (/tæʃ/). Ganong met en évidence un déplacement de la frontière catégorielle qui s'exprime par une préférence, de la part des participants, pour des réponses en accord avec une interprétation lexicale de la séquence de phonèmes dans les parties médianes du continuum. Dans le contexte /æsk/, les auditeurs ont plutôt tendance à classer le phonème ambigu dans la catégorie /t/ alors que l'attitude inverse se manifeste dans le contexte /æʃ/ qui donne lieu à une proportion plus élevée de réponses /d/. Cet effet met en évidence une tendance à percevoir un signal de parole en faveur d'une interprétation lexicale plutôt que non lexicale.

Une interprétation *autonome* peut cependant en être donnée. Dans le modèle Race (Cutler & Norris, 1979), les décisions phonémiques peuvent se faire à partir de deux classes d'informations elles-mêmes dérivées de deux voies de traitement : une voie prélexicale et une voie lexicale. On peut simuler des effets lexicaux par la combinaison et l'intégration de ces deux classes d'informations lors de l'étape de 'prise de décision'. Ainsi, dans le cas de l'effet observé par Ganong (1980), la voie lexicale peut tout à fait influencer l'étape décisionnelle en

[6] Classiquement, dans ce type d'expériences, le phonème ambigu est noté /?/. Nous avons choisi de ne pas suivre ce 'standard' afin, d'une part de ne pas causer de confusion avec le caractère /ʔ/ (occlusive glottale non-voisée) de l'International Phonetics Association (IPA) mais aussi pour permettre au lecteur de garder à l'esprit les 2 éléments extrêmes du continuum considéré. Ainsi, le phonème ambigu correspondant à un continuum acoustique qui va du phonème /d/ au phonème /t/ est noté /$^{d}_{t}$/.

introduisant un biais à donner une réponse phonémique conforme aux informations dérivées du traitement lexical. La tâche des sujets consistant à classifier un son dans une catégorie phonémique, la voie non-lexicale fournit une information purement acoustique sur laquelle peut se fonder le sujet pour apparier ce son avec une représentation phonétique et effectuer une catégorisation[7]. Parallèlement, les informations fournies par le signal permettent d'engendrer un traitement lexical. Au moment de la prise de décision, le sujet dispose de deux types d'informations pour donner sa réponse : une information acoustico-phonétique qui peut ne pas suffire à effectuer la catégorisation pour les items centraux du continuum et une information lexicale qui favorise nécessairement -d'autant plus que l'information acoustique n'est pas 'fiable'- l'une des deux possibilités étant donné qu'une seule des deux extrémités correspond à un mot de la langue. Dans ce modèle, c'est cette information qui va influencer la réponse du sujet et induire un effet lexical ; mais à aucun moment du traitement proprement dit, il n'y a action des niveaux de représentations lexicaux sur les niveaux de représentation phonémiques. L'interprétation des résultats obtenus avec la tâche d'identification phonémique a donné lieu à une quantité importante de travaux qui ne permettent pas de trancher entre ces deux interprétations (cf. Pitt & Samuel, 1993 pour une revue).

2.1.1.2. L'effet du statut lexical sur la détection de phonèmes

Une preuve du recours aux connaissances lexicales dans les processus d'identification phonémique a également été entrevue dans certains effets que l'on a pu mettre en évidence avec la tâche de détection de phonèmes. Dans ces expériences, les auditeurs devaient détecter un phonème qui pouvait apparaître dans diverses positions (au début, au milieu, à la fin) de stimuli auditifs lexicaux ou non. L'idée essentielle est que si les niveaux de traitement lexicaux rétroagissent vers les étapes sublexicales de représentation, on devrait observer des effets différents en fonction du statut lexical et du taux d'avancement de l'identification lexicale. C'est effectivement ce qui a été observé dans plusieurs expériences. On a notamment pu mettre en évidence un effet de la position du phonème à détecter qui diffère selon le statut lexical des items porteurs plurisyllabiques. Dans des mots plurisyllabiques, on n'observe pas d'effet lexical sur les temps de détection de phonèmes lorsque le phonème-cible est en position initiale (Foss & Blank,

[7] On pourra objecter qu'il est nécessaire de disposer d'informations sublexicales afin de rendre possible l'accès lexical. On considèrera en fait l'information linguistique déterminant l'accès à cette représentation lexicale comme une information différente de celle dérivée des traitements réalisés par la voie sublexicale (par exemple, des matrices de traits ou de spectres à court terme pourraient générer l'accès lexical sans donner lieu à une identification phonémique proprement dite, …). On peut trouver dans Marslen-Wilson & Warren (1994) un exemple de modèle utilisant des matrices de traits non-discrètes pour l'appariement avec le lexique. Les mêmes informations sublexicales pourraient servir à développer des représentations phonémiques sans que celles-ci aient pour autant une influence sur l'identification des mots.

1980 ; Segui, Frauenfelder, & Mehler, 1981). Par contre, un effet lexical émerge lorsque le phonème-cible apparaît en fin de plurisyllabe (Cutler, Mehler et al., 1987 ; Frauenfelder, Segui, & Dijkstra, 1990) ; les phonèmes sont alors détectés plus rapidement dans des mots que dans des non-mots.

Ces effets pourraient également être prédits par un modèle purement autonome de la tâche de détection de phonèmes dans lequel l'information lexicale permettrait de fournir des informations supplémentaires pour la prise de décision et rendre par conséquent plus rapide la génération de la réponse du sujet. On observe cependant que ces effets lexicaux augmentent progressivement au fur et à mesure que le phonème-cible est localisé vers la fin du mot. On n'observe pas, contrairement à ce que l'on aurait pu prédire à partir d'un modèle dans lequel aucun processus de propagation de l'activation lexicale n'est implémenté, une absence totale d'effet avant le *Point d'Unicité* (PU) suivie d'une brusque émergence après le PU (Frauenfelder et al., 1990). Il y a au contraire une évolution progressive de ces processus rétroactifs qui est corrélative des phénomènes d'activation lexicale et qui n'est pas prédite par un modèle purement symbolique tel que Race (Cutler & Norris, 1979) dans lequel l'information lexicale n'est disponible qu'un fois le mot identifié.

2.1.1.3. Effets lexicaux sur la restauration phonémique

Des effets lexicaux émergent également dans la tâche de restauration phonémique. Le phénomène de restauration phonémique a été mis en évidence par Warren (1970). La tâche consiste à présenter à des sujets deux types de stimuli : des séquences de parole dans lesquelles l'un des segments a été remplacé par du bruit et d'autres dans lesquelles on a ajouté du bruit au segment, celui-ci restant indemne. Si l'on demande aux auditeurs de décider pour chaque stimulus si le segment bruité a été remplacé ou s'il est indemne, ils éprouvent du mal à donner une réponse correcte. Cette difficulté induit les auditeurs à percevoir un segment de parole qui a été remplacé par du bruit comme relativement indemne. Les sujets réagissent comme si le bruit ne faisait que recouvrir le signal de parole alors que celui-ci est en réalité absent. Ce phénomène repose en partie sur des processus ascendants de traitement de l'information que l'on peut observer dans les traitements psychoacoustiques (cf. la continuité illusoire ; Bregman, Colantonio, & Ahad, 1999). Un effet lexical a cependant été mis en évidence par Samuel (1996) qui se réfère à la Théorie de la Détection du Signal (TDS ; Green & Swets, 1966). Si l'on demande à un sujet de réaliser la tâche décrite précédemment, on peut observer un phénomène de restauration phonémique qui consiste à distinguer difficilement les stimuli intacts des stimuli dans lesquels le signal a été remplacé par du bruit. Ceci se manifeste par une tendance à

percevoir le segment remplacé par du bruit comme intact. Cet effet peut se mesurer avec une variable dérivée de la TDS : le *d'*. Cette variable est considérée comme un indice de la discriminabilité perceptive entre conditions expérimentales. Samuel (1996) met en évidence un phénomène de restauration phonémique pour les deux types de stimuli (mots ou non-mots) ; mais ce phénomène s'accentue pour les mots : la valeur du *d'* est moins élevée pour les mots que pour les non-mots. En d'autres termes, les sujets éprouvent plus de difficultés à distinguer les stimuli intacts des stimuli dans lesquels le signal a été supprimé lorsqu'ils identifient des mots que lorsqu'ils traitent des stimuli qui n'ont pas de représentation spécifique dans le lexique. Cette diminution, dans les mots par rapport aux non-mots, de la discriminabilité perceptive entre un segment auquel on a *ajouté* du bruit et un segment qui a été *remplacé* par du bruit a été interprétée par Samuel (1996) comme un indice d'une rétroaction du lexique vers les représentations phonémiques. Là encore, on peut objecter que ce phénomène ne reflète pas nécessairement une rétroaction lexico-phonémique. Le problème se situe ici au niveau de la signification -et surtout de l'interprétation- du *d'*. Dans le cadre de la TDS, le *d'* est un indice de discriminabilité perceptive. Un effet en termes de *d'* représente une différence dans les processus perceptifs mais pas dans des processus stratégiques (qui se refléteraient plutôt dans la variable β). Si l'on se réfère à un modèle dans lequel les effets lexicaux sont simulés par une intégration des informations sublexicales et lexicales, les informations lexicales peuvent contribuer à influencer la prise de décision, mais ceci ne coïncide pas nécessairement avec le développement d'une stratégie de réponse[8]. Les effets observés sur la variable *d'* peuvent donc tout à fait refléter un processus d'intégration d'informations et se refléter dans les valeurs de la variable *d'*. On ne peut par conséquent pas inférer de ces modifications du *d'* que ces effets lexicaux sont liés à une rétroaction lexique-phonème qui induirait une modification du codage même de l'information phonémique, plutôt qu'à une étape d'intégration des deux classes d'informations qui seraient codées indépendamment l'une de l'autre.

Plus récemment, Samuel (1997) a mis en évidence la possibilité, dans une procédure d'adaptation sélective, de générer un effet lexical dans le phénomène de restauration phonémique. La procédure d'adaptation consiste à soumettre des auditeurs à l'écoute prolongée de séquences de parole dans lesquelles l'un des phonèmes est prononcé de manière répétée (par

[8] On utilisera donc ici le terme 'intégratif' pour désigner des processus d'intégration des diverses informations disponibles à la sortie du traitement -intégration qui peut se faire par pondération des différentes catégories d'information ou par sélection de l'information jugée la plus pertinente- ; et on les distinguera des processus 'stratégiques' ou 'décisionnels' qui, dans le cadre de la théorie de la détection du signal, correspondraient à l'élaboration (consciente ou non) d'une *ligne de conduite* dans le choix des réponses afin d'accomplir au mieux la tâche avec le moins d'erreurs possibles (ce qui correspondrait alors au critère β).

exemple [p]). Au bout d'un certain temps, on fait écouter aux auditeurs des stimuli qui contiennent un phonème ambigu situé sur un continuum acoustique (imaginons que ce continuum corresponde à une modification du temps d'attaque vocale et que, par conséquent, le phonème opposé à celui qui était présenté dans la phase d'adaptation soit [b]). Dans une seconde phase, on présente des stimuli ambigus qui varient sur un continuum [$^{b}_{p}$]. Le phonème [$^{b}_{p}$] peut être interprété soit comme celui qui s'est répété au cours de la phase d'adaptation ([p]), soit comme celui qui correspond à l'autre extrémité du continuum ([b]). On observe une tendance, suite à la phase d'adaptation, à interpréter le phonème ambigu comme faisant partie de la classe opposée à celle à laquelle les auditeurs ont été soumis au préalable (Eimas & Corbit, 1973). Une saturation des récepteurs qui interviendrait à différents niveaux de traitement (sensoriel, phonétique, phonémique ; Samuel & Kat, 1996) se mettrait en place qui induirait un déplacement de la frontière perceptive. Dans la situation imaginée par Samuel (1997), les sujets sont soumis dans une première phase à une procédure d'adaptation phonémique. Ils entendent, au cours de cette période, des stimuli dans lesquels c'est toujours le même phonème qui est détérioré (par remplacement ou superposition). Ces stimuli peuvent correspondre à des mots ou à des non-mots. L'idée avancée par Samuel (1997) est que, si les représentations lexicales rétroagissent effectivement sur les niveaux de représentation prélexicaux en modifiant l'activation de ces derniers, l'effet lexical qui accentue le phénomène de restauration phonémique devrait induire un autre phénomène : une adaptation phonémique sélective à l'une des extrémités du continuum, en l'occurrence celle qui donne lieu à un mot de la langue. On devrait alors observer, dans la seconde phase de l'expérience, une tendance à percevoir le phonème ambigu dans la catégorie opposée à celle du phonème qui avait été remplacé par du bruit dans la première phase. Effectivement, les résultats mettent en évidence une tendance, après une période d'adaptation au cours de laquelle les auditeurs entendaient des mots dans lesquels l'un des phonèmes était remplacé par du bruit, à favoriser le phonème opposé sur le continuum. Selon Samuel (1997), ceci met en évidence l'action effective du lexique sur les représentations phonémiques prélexicales par une modification de l'activation des récepteurs phonémiques. On peut cependant, là encore, affirmer qu'un modèle autonome peut rendre compte de cet effet. La question se pose notamment de savoir à quel niveau se situent précisément le phénomène d'adaptation phonémique et l'influence lexicale observés par Samuel (1997). On peut se demander si les représentations phonémiques sur lesquelles sont censées agir les informations lexicales sont réellement des représentations précoces recevant des activations de niveaux de représentation plus élaborés ou si, au contraire, ces représentations phonémiques ne sont pas

calculées parallèlement aux informations lexicales, auquel cas une interprétation similaire à celle qui peut être proposée pour les effets observés dans la tâche de catégorisation phonémique est envisageable. Si représentations lexicales et sublexicales sont calculées par le biais de deux voies de traitement différentes, le phénomène d'activation des unités phonémiques par les unités lexicales que l'on observe dans le phénomène de restauration phonémique peut refléter un processus post-lexical lié à une intégration des deux types d'informations. L'adaptation induite par la restauration phonémique constituerait alors, elle aussi, un processus décisionnel dans lequel les récepteurs phonémiques auraient tendance à voir leur seuil d'activation s'élever en raison de cette influence qu'à le lexique sur les phonèmes lors de l'étape d'intégration des informations.

2.1.1.4. Effet lexical dans la compensation de la coarticulation

Finalement, l'effet essentiel sur lequel repose l'affirmation d'une boucle rétroactive du lexique vers les représentations phonémiques est celui mis en évidence par Elman & McClelland (1988). Cet effet repose sur un processus nécessairement pré-lexical et mis en évidence par (Mann & Repp, 1981) : la compensation perceptive des phénomènes coarticulatoires (*compensation-for-coarticulation*). En anglais, un phonème ambigu [$^{t}_{k}$] a tendance à être perçu [t] s'il est prononcé après [ʃ] alors qu'il est plutôt perçu [k] s'il est prononcé après [s]. Ce phénomène de compensation perceptive trouve son explication dans le partage de la place d'articulation des occlusives [t] et [k] et, respectivement, des fricatives [s] et [ʃ], partage qui induirait à compenser cette communauté de place d'articulation par une tendance à percevoir le phonème suivant dans la catégorie opposée. Les auditeurs utiliseraient donc une connaissance qu'ils auraient intégrée sur les phénomènes coarticulatoires afin d'influencer le processus d'appariement entre image auditive et représentation phonétique. Elman & McClelland (1988) utilisent ce phénomène dans des séquences lexicales qui se terminent par un phonème ambigu provenant d'un continuum [s]-[ʃ], ce phonème étant lui-même suivi d'un autre phonème présentant une ambiguïté entre [t] et [k]. Les sujets pouvaient par exemple entendre la séquence [fulɪ$^{s}_{ʃ}$] suivie d'un mot commençant par [$^{t}_{k}$]. Si le mot *foolish* provoque une modification de l'activation du phonème /ʃ/ par un processus de rétroaction du lexique vers une représentation prélexicale, on devrait alors mettre en évidence un phénomène de compensation perceptive induisant les auditeurs à percevoir de manière privilégiée le phonème ambigu [$^{t}_{k}$] comme un représentant de la catégorie /t/. Au contraire, si le biais qui était observé dans les expériences de catégorisation phonémique (Ganong, 1980) est localisé au niveau de l'extraction d'une

information phonémique à partir d'une représentation lexicale, la voie prélexicale de traitement devrait fournir des informations non-biaisées du phonème ambigu [$^{s}_{ʃ}$]. On ne devrait alors observer aucune préférence pour l'un des deux phonèmes /t/ ou /k/ en ce qui concerne le second phonème ambigu. Or les auteurs observent au contraire un déplacement de la frontière catégorielle du son [$^{t}_{k}$] en faveur d'une compensation perceptive de la coarticulation. Lorsque les auditeurs traitent la séquence /fulɪ$^{s}_{ʃ}{}^{t}_{k}$/ composée d'une suite de deux phonèmes ambigus, l'interprétation lexicale favorise un percept /ʃ/ pour le premier phonème, ceci influence alors la perception du second phonème dans la catégorie /t/. Les informations lexicales sont donc en mesure de favoriser l'augmentation d'activation du phonème /ʃ/ qui, à son tour, modifie l'activation du phonème /t/. Il existe cependant, à nouveau, une possibilité d'interprétation alternative. En anglais, en effet, il apparaît que la séquence /ɪʃ/ est plus fréquente que la séquence /ɪs/. On peut alors envisager que l'influence exercée par le contexte [ɪ] sur la catégorisation du phonème ambigu [$^{s}_{ʃ}$] est purement prélexicale et liée à des phénomènes probabilistes (cf. Saffran, Newport, & Aslin, 1996; Vitevitch & Luce, 1999 ; ces travaux seront présentés plus en détails dans le Chapitre 3) ; auquel cas l'effet qui est observé dans cette expérience pourrait s'expliquer par des processus qui sont tous localisés à un niveau pré-lexical. On n'aurait alors aucune nécessité de postuler une rétroaction lexique-phonèmes dans les processus de perception de la parole : deux processus prélexicaux se succéderaient, le premier consistant à favoriser l'interprétation du [$^{s}_{ʃ}$] dans la catégorie [ʃ] du fait même de la plus grande probabilité de [ɪʃ] en anglais ; le second -conséquence du biais probabiliste à percevoir [ʃ]- consistant à faire appel à des 'connaissances articulatoires' pour favoriser la perception d'un [t] dans le continuum [$^{t}_{k}$].

2.1.2. Un modèle autonome des phénomènes d'interaction ?

Concernant l'existence de processus rétroactifs permettant aux représentations lexicales d'influencer le codage même des représentations prélexicales, la question reste en suspens. Dans les quatre exemples qui précèdent, nous avons vu qu'un modèle autonome ne pouvait être exclu pour rendre compte des données considérées jusqu'ici comme des preuves de l'interaction entre niveaux de traitement plus ou moins élaborés. Il reste cependant des résultats que ne peut prédire le modèle RACE (Cutler & Norris, 1979).

2.1.2.1. Le problème des effets ‘lexicaux’ dans des non-mots

Un certain nombre d’auteurs ont mis en évidence l’existence d’effets ‘lexicaux’ dans des expériences dans lesquelles les stimuli étaient tous des non-mots. Connine, Titone, Deelman, & Blasko (1997), utilisant une tâche de détection de phonème dans des non-mots, observent des temps de réaction plus rapides dans les non-mots qui ressemblent à des mots (‘gabinet’ vs. ‘cabinet’ par exemple) que dans des non-mots qui ne ressemblent à aucun mot. Newman, Sawusch, & Luce (1997), quant à eux, ont répliqué l’expérience de Ganong (1980) en n’utilisant que des non-mots aux deux extrémités des continua. Les paires de non-mots différaient sur le plan du nombre de voisins lexicaux (i.e. de la densité de voisinage). L’une des extrémités constituait un non-mot qui avait beaucoup de voisins lexicaux (par exemple /gaɪs/ ou /kaɪp/) alors que l’autre avait peu de voisins lexicaux (par exemple /kaɪs/ ou /gaɪp/). Ils observent un phénomène similaire à celui observé par Ganong (1980) : il y a déplacement de la frontière catégorielle vers la catégorie phonémique qui donne lieu à une faible densité de voisinage. En d’autres termes, les auditeurs ont tendance à donner une réponse qui favorise le phonème correspondant au non-mot qui a la plus forte densité de voisinage.

Ces effets mettent en évidence l’existence de processus d’accès au lexique (ou plus précisément de propagation de l’activation lexicale) au cours du traitement de stimuli de parole qui n’ont pas de représentation dans le lexique. Ces processus influencent l’interprétation qui peut être donnée, par les participants, de stimuli phonétiques ambigus. Un modèle comme Race (Cutler & Norris, 1979) ne peut pas prédire ce type d’effets. Dans ce modèle, les non-mots sont uniquement traités par le biais de la voie prélexicale. Comme ils n’ont pas de représentations dans le lexique et que Race n’implémente pas de processus de propagation de l’activation lexicale qui pourrait être induite par des portions de non-mots qui existent dans des mots de la langue, cette voie lexicale ne peut pas influencer -même au cours de l’étape décisionnelle- les réponses des sujets lorsqu’ils traitent des non-mots. Un modèle comme TRACE (McClelland & Elman, 1986) pourrait au contraire prédire cet effet en raison de la mise en œuvre de procédures de propagation progressive de l’activation. Il n’est cependant pas exclu que les effets liés à la densité lexicale des non-mots (Newman et al., 1997) soient interprétables en termes pré-lexicaux plutôt que rétroactifs. En effet, il existe une corrélation forte entre le nombre de voisins lexicaux d’une chaîne de phonèmes et la fréquence des diphones qui la constituent (Vitevitch, Luce, Pisoni, & Auer, 1999; Vitevitch & Luce, 1999).

2.1.2.2. Le modèle MERGE

Norris, McQueen, & Cutler (2000) proposent un modèle qui permet de rendre compte des effets de compétition lexicale observés dans des non-mots tout en n'implémentant aucune boucle de rétroaction du lexique vers les phonèmes. Ce modèle est un Réseau Récurrent Simple (Simple Recurrent Network, SRN ; Jordan, 1986) qui, contrairement à TRACE (McClelland & Elman, 1986) dans lequel les connexions sont définies *a priori* par l'expérimentateur (*Interactive Activation and Competition model, IAC*) subit une phase d'apprentissage au cours de laquelle il régule lui-même le poids des différentes connexions entre les nœuds du réseau. On peut alors parler de *Réseau Neuronal* à proprement parler en raison des capacités du modèle à apprendre : c'est l'organisation des poids des différentes connexions qui détermine le comportement du modèle à l'issue de la phase d'apprentissage. Les réseaux récurrents simples (SRN), du fait de l'implémentation de procédures de récurrence locales -des unités cachées (*Hidden Units*) vers les unités d'entrée- autorisent l'insertion d'un retard temporel dans le retour de l'information et permettent ainsi d'aboutir à une représentation correcte des aspects séquentiels (et temporels) des informations traitées (Jordan, 1986 ; Elman, 1990) mais peuvent aussi, grâce à l'étape d'apprentissage, développer des représentations sensibles aux régularités séquentielles des formes rencontrées (cf. Chapitre 2 pour une présentation de cet aspect). Ces modèles n'utilisent que des connexions ascendantes[9] ; aucune rétroaction d'un niveau de représentation élaboré vers un niveau moins élaboré n'est possible. Le modèle MERGE est en fait le *cousin* connexionniste de Race (Cutler & Norris, 1979). Dans le même esprit, deux types d'informations[10] sont disponibles à la sortie du modèle (lexicales et sublexicales) ; et la même étape d'intégration de ces deux classes d'informations est implémentée. Ce qui diffère est en fait la possibilité qu'a le modèle de propager progressivement l'information lexicale vers les nœuds phonémiques -qui rappelons-le sont à considérer ici comme reflétant des représentations de nature différente mais de niveau similaire, les deux processus se déroulant en parallèle- sans attendre l'aboutissement du processus d'accès lexical.

Norris et al. (2000) montrent que ce modèle est en mesure de simuler toutes les données qui étaient présentées comme des preuves de l'existence d'une rétroaction du lexique sur les phonèmes. Il peut en outre, et contrairement à Race (Cutler & Norris, 1979), reproduire les effets de voisinage lexical observés dans des non-mots. Si ce modèle peut reproduire des effets d'interaction, c'est que les informations lexicale et sub-lexicale sont représentées comme deux

[9] La boucle de récurrence ne constitue pas une boucle de rétroaction proprement dite puisqu'elle a pour fonction de donner une mémoire temporelle au modèle et pas de moduler le codage effectué à un autre niveau du traitement.
[10] On parle de *Multiple Output models* (Boland & Cutler, 1996).

classes équivalentes d'unités de sortie. Ce sont donc les procédures décisionnelles -ou d'intégration des informations- (comme dans Race, Cutler & Norris, 1979) qui permettent l'influence de l'une sur l'autre. Du fait de sa nature connexionniste, il implémente également des processus de propagation progressive de l'activation des diverses unités. Il est ainsi en mesure de reproduire tous les effets dont nous avons parlé précédemment sans introduire de boucle rétroactive.

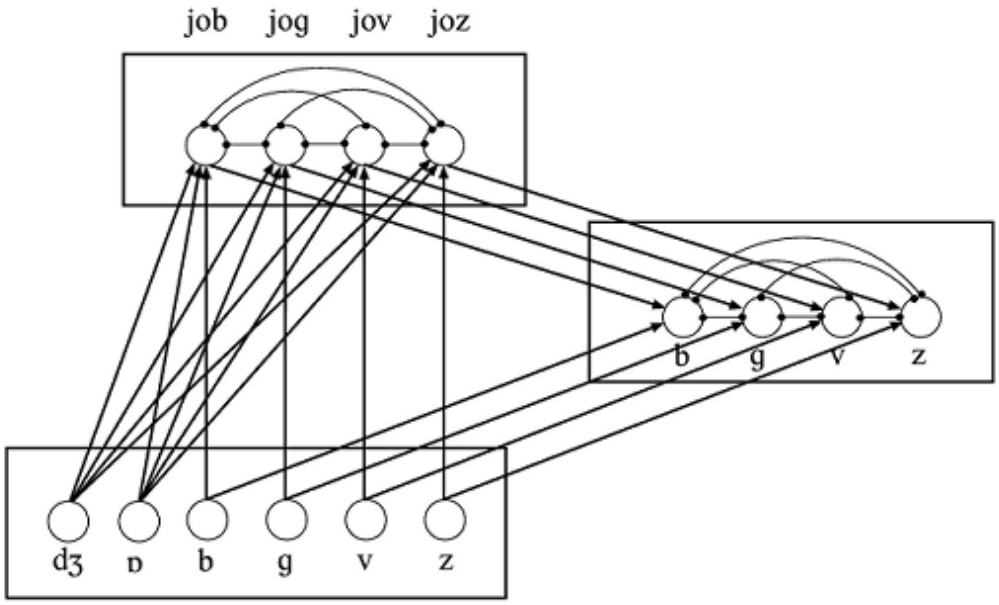

Figure 5 : Le modèle MERGE (Norris et al., 2000). Deux catégories d'informations sont disponibles à la sortie du traitement : la prise de décision (le niveau d'activation des nœuds de la couche de droite) peut aussi bien reposer sur le niveau d'activation des nœuds phonémiques (en bas) que sur celui des nœuds lexicaux (en haut). Le type de représentation utilisé ne permet pas de faire apparaître les boucles de récurrence.

Une question reste cependant posée : celle du statut des unités cachées dans l'émergence des représentations que développe le modèle (cf. Massaro, 1988 pour une critique de la fonction des unités cachées dans les modèles du fonctionnement cognitif). En effet, une Analyse en Composantes Principales effectuée par Norris et al. (2000) a consisté à essayer d'identifier quelles représentations émergent des unités cachées (*Hidden Layer Units*) suite à l'apprentissage. La question est importante puisqu'on pourrait imaginer que ces unités cachées, dont le nom indique qu'elles sont inaccessibles au processus de prise de décision, codent indépendamment les informations lexicales et phonémiques. Certaines unités constitueraient alors des unités lexicales alors que d'autres représenteraient des unités phonémiques. Les phénomènes d'interaction lexique-phonème qui -selon l'interprétation initiale des auteurs- émergeraient en raison de l'existence de procédures d'intégration de ces deux classes d'informations au niveau décisionnel, pourraient alors se produire avant la prise de décision (donc dans le cadre des processus de

codage de l'information). Ceci reviendrait simplement à implémenter des procédures rétroactives d'une manière différente de celle que l'on trouve dans des modèles d'activation interactive comme TRACE (McClelland & Elman, 1986). L'analyse effectuée par les auteurs montre au contraire que chaque unité cachée code à la fois des informations lexicales et sublexicales. C'est donc la combinaison d'activation de plusieurs unités cachées qui détermine l'activation des unités de sortie (phonémiques ou lexicales). Les auteurs en concluent que rien ne permet de dire qu'il y a eu interaction entre unités au niveau du codage ; c'est au contraire au niveau de l'intégration des informations que se situe l'interaction.

Il faut toutefois noter que, s'il n'y a effectivement pas indépendance des représentations de la couche cachée (*Hidden Layer*) -ce qui implique qu'il n'est pas possible d'expliquer cette *interaction* lexique-phonème par le biais de processus interactifs qui prendraient place entre des unités fonctionnellement *et* structurellement indépendantes- le problème n'est pas pour autant résolu. Si chaque unité cachée a codé simultanément de l'information phonémique *et* de l'information lexicale, ceci les ayant conduit à développer des représentations *hybrides*, ces deux classes d'information ne sont par conséquent pas indépendantes l'une de l'autre. Cette absence d'indépendance est alors bien réelle avant même l'accès à l'étape d'intégration des informations, donc au cours du codage même des diverses représentations. La question qui se pose alors est de savoir si cette caractéristique du modèle est à envisager -en se référant à la typologie de Marr (1982)- comme une particularité *algorithmique* qui serait nécessaire pour être en mesure de simuler des effets d'interaction dans un modèle purement ascendant -auquel cas la non-indépendance des représentations qui se développent dans la couche cachée ne serait qu'une manière particulière de développer dans un programme informatique des processus ascendants aptes à simuler des phénomènes d'interaction, ou si c'est au contraire une caractéristique *computationnelle* qui serait alors à concevoir comme partie intégrante des processus permettant au système simulé, le système cognitif, de développer des effets de ce type. Si l'on admet une interprétation purement algorithmique de la couche cachée, on est en droit d'affirmer qu'un modèle ascendant tel que MERGE peut simuler des phénomènes rétroactifs.

Indépendamment de cette question, il reste que les SRN constituent une classe de modèles particulièrement intéressante dans le cadre du problème qui nous intéresse ici -le rôle des contraintes phonologiques séquentielles dans les processus de segmentation de la parole- car ils sont en mesure d'acquérir des informations sur les régularités d'occurrence des séquences (Elman, 1990) qu'ils reçoivent en entrée au cours de l'apprentissage, ce qui n'est pas le cas dans les modèles IAC comme TRACE (McClelland & Elman, 1986). Or les contraintes phonologiques

qui portent sur les régularités des séquences de phonèmes dans la langue pourraient tout à fait être acquises par le biais de processus de cet ordre (Brent, 1996).

2.2. Le recours à des connaissances pré-lexicales

Outre le rôle éventuel des connaissances lexicales dans les processus d'identification phonémique qui, si elles sont aujourd'hui avérées peuvent être modélisées aussi bien par des phénomènes rétroactifs que par des processus d'intégration d'informations, d'autres auteurs ont cherché à étudier le rôle éventuel des connaissances concernant les régularités phonologiques de la langue dans des processus d'identification.

2.2.1. Les données expérimentales

Nous décrivons ici trois groupes d'études qui ont abouti à la conclusion d'une utilisation, par le système cognitif, de connaissances concernant les régularités phonologiques de la langue dans les processus d'identification phonémique. Nous restreignons les études mentionnées à celles qui ont pour objet des contraintes séquentielles portant sur la production de phonèmes adjacents dans la chaîne de parole : (1) la compensation perceptive de la coarticulation (Mann & Repp, 1981), (2) le rôle des contraintes phonotactiques dans la catégorisation de phonèmes ambigus (Massaro & Cohen, 1983), et (3) le rôle des contraintes phonotactiques dans l'identification de phonèmes acoustiquement non-ambigus (Hallé, Segui, Frauenfelder, & Meunier, 1998).

2.2.1.1. La compensation perceptive de la coarticulation

Nous avons déjà présenté rapidement ce résultat, utilisé comme point de départ de l'étude réalisée par Elman & McClelland (1988), dans la partie consacrée au rôle du lexique dans l'identification phonémique (cf. Section 2.1.1.4). Il a été mis en évidence (Mann & Repp, 1981) un rôle des connaissances liées aux contraintes coarticulatoires sur les processus d'identification de phonèmes ambigus. Cet effet se manifeste par une tendance à percevoir un phonème ambigu [$^{t}_{k}$] comme représentant de la catégorie [t] s'il est prononcé après [ʃ] alors qu'il est plutôt perçu [k] s'il est prononcé après [s]. Ceci s'expliquerait par l'utilisation, de la part des auditeurs, de connaissances concernant le partage du lieu d'articulation dans la production de ces sons de parole, [t] et [s] partageant le trait [+dental] alors que [k] et [ʃ] partagent le trait [+palatal]. La coarticulation induit en effet le système de production de la parole à 'assimiler' certains phonèmes à d'autres. Ici, la production du son dental [s] entraîne les articulateurs à produire le

son suivant avec des caractéristiques qui lui sont proches -en l'occurrence le lieu d'articulation- en influençant le contrôle des gestes articulatoires afin de diminuer la distance -exprimée en termes de traits distinctifs- entre ce son et le précédent. Dans la production d'un signal naturel, ces contraintes influencent les caractéristiques acoustiques des sons de parole. Certains sons risquent alors de présenter des caractéristiques qui ne déterminent pas exactement leur représentation phonologique sous-jacente. Le système d'identification phonémique pourrait alors les interpréter comme des phonèmes différents de ceux que souhaitait transmettre le locuteur. Lorsque le système perçoit une séquence de phonèmes qui sont relativement proches, il serait utile de pouvoir compenser cet effet de la coarticulation. On voit donc en quoi ce type de procédures pourrait être utile au processus d'appariement entre forme phonétique de surface et représentation phonologique sous-jacente. L'effet mis en évidence par Mann & Repp (1981) illustre l'aptitude du système de perception de la parole à avoir recours à des connaissances qu'il aurait intégrées concernant les processus de production pour modifier les représentations qui sont dérivées du signal acoustique et parvenir à apparier cette image auditive avec les représentations phonologiques adéquates.

Ces effets mettent en évidence l'utilisation, dans les processus d'identification d'un signal de parole, de contraintes qui régissent le contrôle articulatoire pour la réalisation acoustique des sons. Un ensemble de régularités phonologiques des langues (dites contraintes phonotactiques) intervient également dans les processus de production du signal de parole. Ces contraintes sont dépendantes des langues et intégrées par les locuteurs au cours du processus d'acquisition. Elles définissent un ensemble de 'règles' qui régissent la possibilité (ou l'impossibilité) qu'ont les locuteurs d'une langue donnée de produire certaines séquences de phonèmes dans un signal de parole. Une discussion plus approfondie de la notion de *contraintes phonotactiques* sera proposée dans le Chapitre 4. Cette classe de contraintes constitue le cœur du problème que nous étudions dans le cadre de notre travail. Nous présentons ici les travaux qui ont été réalisés afin d'évaluer le rôle de ces contraintes dans les processus d'identification de la parole sachant que notre étude porte plus spécifiquement sur la question de la *segmentation du signal de parole en mots* ; problématique qui sera abordée dans le chapitre suivant.

2.2.1.2. Le rôle des contraintes phonotactiques dans l'identification de phonèmes ambigus

Si l'on présente à des auditeurs une séquence de parole sans signification dans laquelle l'un des phonèmes est ambigu, ceux-ci ont tendance à donner du son ambigu une interprétation qui préserve les contraintes phonotactiques de la langue. Ce résultat a été mis en évidence par Massaro & Cohen (1983). Dans une tâche de catégorisation phonémique, des locuteurs anglais

entendaient des logatomes CCV dans lesquels le phonème médian variait sur un continuum [r]-[l]. Dans chacun des groupes de stimuli, l'une des extrémités du continuum constituait un groupe de consonnes phonotactiquement illégal en début de mot en anglais (par exemple */tla/) alors que l'autre extrémité était parfaitement attestée dans la langue (/tra/) en position initiale de mot. En fonction de la catégorie dans laquelle les participants classaient le phonème ambigu, la séquence de consonnes initiale donnait par conséquent lieu soit à une séquence phonotactiquement illégale dans la langue (*/sra/, */dla/ ou */tla/), soit à une séquence légale (/sla/, /dra/ ou /tra/). Les auteurs observent une tendance des participants à favoriser une interprétation légale de la séquence de phonèmes ; ceci se manifestant par un déplacement de la frontière catégorielle. Dans les séquences /s[r]a/, les participants ont tendance à percevoir plus souvent [r] comme un représentant de la catégorie /l/ alors qu'ils le perçoivent plutôt /r/ dans /t[r]a/ et /d[r]a/. Les auditeurs seraient donc en mesure d'utiliser une certaine catégorie de connaissances ayant trait aux régularités phonologiques de leur langue pour modifier les représentations qui peuvent être construites à partir d'une analyse acoustique du signal de parole.

Pitt (1998) soulève cependant la question, déjà évoquée par Massaro & Cohen (1983), de la cause effective de ce déplacement de la frontière catégorielle dans les résultats précédents. En effet, alors que les séquences */tl/ et */dl/ constituent toutes deux des suites de phonèmes phonotactiquement illégales en anglais, les résultats obtenus par Massaro & Cohen (1983) montrent un déplacement plus important de la frontière catégorielle pour les réponses données aux séquences [t[r]] par rapport à celles observées pour les stimuli [d[r]]. Or il apparaît qu'en anglais, la différence de fréquence -en position médiane de mot- entre les séquences /tl/ et /tr/ est plus importante qu'entre les séquences /dl/ et /dr/. Ces effets pourraient alors s'expliquer plus par l'intervention de processus probabilistes que par des *connaissances* fondées sur des règles ou des contraintes distinguant deux catégories de séquences : légales et illégales. Selon cette interprétation, l'émergence d'un déplacement plus important de la frontière pour les paires /tl/-/tr/ que pour les paires /dl/-/dr/ serait lié à un décalage dans les fréquences relatives des groupes de consonnes. Pitt (1998) réplique l'expérience réalisée par Massaro & Cohen (1983) afin d'estimer la contribution relative des sources d'information fréquentielles et phonologiques dans les effets observés. Pour cela, il compare quatre continua dans lesquels les rapports, entre les deux extrémités, en termes de légalité phonotactique ou de fréquence sont en partie orthogonaux. Il compare ainsi les continua utilisés par Massaro & Cohen (1983) à deux autres continua dont

les extrémités constituent toutes deux des séquences légales mais présentent une différence dans leurs fréquences d'occurrence (/gr/ vs. /gl/, /gr/ étant plus fréquent -quelle que soit sa position dans les mots- que /gl/) ou pas de différence de fréquence du tout (/br/ vs. /bl/). Cette dernière condition fournit une ligne de base pour estimer la localisation de la frontière entre /r/ et /l/. L'étude d'un éventuel déplacement de la frontière dans le contexte /g/ permet d'estimer le rôle joué par la fréquence indépendamment de la légalité phonotactique. Avec ce contrôle, Pitt (1998) met en évidence un rôle effectif mais faible de la fréquence…

2.2.1.3. Le rôle des contraintes phonotactiques dans l'identification de phonèmes non-ambigus

Abordant une problématique similaire à celle de Massaro & Cohen (1983), Hallé, Segui, Frauenfelder & Meunier (1998) observent une modification dans la perception de séquences illégales *mais* non-ambiguës. Les auteurs comparent, chez des locuteurs français, le traitement de séquences phonotactiquement illégales en début de mot dans la langue (/tl/, /dl/) à leur contrepartie légale (/tr/, /dr/). Contrairement aux stimuli utilisés dans les expériences de catégorisation phonémique (Massaro & Cohen, 1983 ; Pitt, 1998), la consonne liquide médiane des stimuli utilisés n'est pas ambiguë. Par ailleurs, des différences phonétiques caractérisent le phonème /r/, lequel se réalise avec une articulation uvulaire ([ʁ]) en français alors qu'il est alvéolaire ([r]) en anglais. De ce fait, les phonèmes /r/ et /l/ sont, sur le plan acoustique, plus distincts dans la langue française qu'en anglais. Les auteurs prédisent par conséquent un effet plus probable de la légalité sur le percept attaché à la première consonne du groupe qu'à celui de la consonne liquide. Les stimuli utilisés sont des bisyllabes commençant par une séquence de consonnes illégale (/tl/ dans /tlabdo/) ou légale (/tr/ dans /trabdo/). Dans les deux premières expériences, la tâche des sujets consiste respectivement à retranscrire ce qu'ils pensent avoir entendu ou à choisir parmi deux possibilités celle qui leur semble correcte. Les résultats montrent que les sujets tendent effectivement à percevoir les séquences illégales comme des séquences légales. Lorsqu'elle est suivie d'un /l/, l'occlusive dentale initiale (respectivement /t/ et /d/) est dans la majorité des cas perçue comme une occlusive vélaire (respectivement /k/ et /g/). Les participants ne donnent que très rarement une interprétation autre que vélaire (par exemple les bilabiales /p/ et /b/) de la consonne initiale. Dans une tâche de dévoilement progressif (gating, Grosjean, 1985), il apparaît que la consonne initiale est correctement identifiée avant l'occurrence de la liquide (que celle-ci soit un /l/ ou un /r/) ; ce n'est qu'une fois la liquide identifiée que les sujets commencent à privilégier les réponses vélaires par rapport aux

réponses dentales lorsque la contrainte phonotactique n'est pas respectée. Les auteurs en concluent que les locuteurs utilisent des connaissances sur les régularités phonotactiques de leur langue pour 'filtrer' le signal acoustique et aboutir à un percept légal dans leur langue.

En reprenant les problèmes posés par l'étude de Massaro & Cohen (1983), on peut à nouveau mettre en doute le rôle de connaissances sur des contraintes permettant de distinguer deux catégories de séquences -légales et illégales- et se demander si des informations probabilistes ne suffiraient pas à rendre compte de ces effets. Du fait, certainement, de leur illégalité phonotactique, les suites dentale + liquide sont non seulement -presque- inexistantes en début de mot mais également, si l'on calcule leur fréquence d'apparition sans prendre en compte leur position dans les mots, beaucoup plus rares que les suites vélaire + liquide (cf. Chapitre 4 pour une analyse distributionnelle des groupes de deux consonnes dans un lexique français). Il est par conséquent possible que cette tendance à percevoir les occlusives dentales comme des vélaires lorsqu'elles sont suivies d'un /l/ puisse s'expliquer par l'utilisation, dans le cadre du système de décodage acoustico-phonétique proprement dit ou au niveau de l'intégration des diverses informations disponibles, de statistiques lexicales sur les probabilités d'occurrence des séquences phonémiques. L'une des difficultés posées par cette interprétation est que les sujets donnent presque exclusivement des réponses vélaires et très peu de réponses bilabiales. Dans le cadre d'une interprétation statistique, ces deux types de séquences ne présentant pas de contraste clair quant à leur fréquence d'occurrence dans le lexique, devraient se retrouver à part égale dans les réponses des sujets. Or on observe au contraire une dissymétrie marquée en faveur des réponses vélaires. L'interprétation phonotactique ne fournit cependant pas une solution plus claire à cette dissymétrie. Les séquences /kl/ et /pl/ sont tout aussi légales l'une que l'autre. Il n'existe par conséquent pas de justification phonotactique au déséquilibre observé. Par contre, on peut tout à fait envisager une proximité acoustique plus importante (qui pourrait se mesurer à l'aide d'une matrice de confusions dans une tâche perceptive) entre /t/ et /k/ qu'entre /t/ et /p/ ; ceci pourrait expliquer la forme des résultats et s'appliquer aussi bien dans le cadre d'une interprétation phonotactique que statistique du biais observé dans les expériences de Hallé et al. (1998).

Résumé

Nous avons vu, dans la première partie de ce chapitre, les diverses difficultés posées par la modélisation de l'appariement entre un signal

acoustique et des représentations linguistiques stockées en mémoire. Dans la seconde partie, nous nous sommes attaché à présenter un certain nombre d'études dans lesquelles la question de l'utilisation de connaissances de haut niveau (lexicales ou phonologiques) pour influencer le codage de l'information auditive s'est posée. Il apparaît clairement que la réponse à cette question n'est pas disponible à l'heure actuelle. Seules des convictions personnelles permettent aujourd'hui de se situer d'un côté ou de l'autre de la 'frontière théorique'. Cette analyse des données disponibles nous a cependant permis d'évoquer la difficulté que l'on peut rencontrer à tester certaines hypothèses en rapport avec le traitement du langage, difficulté en grande partie associée au lien étroit qui peut exister entre des variables telles que la densité de voisinage lexical et la fréquence des diphones (rôle du lexique) ou la légalité phonotactique et la fréquence (rôle des contraintes phonotactiques). Dans le chapitre suivant, nous abordons de manière plus précise la question centrale de notre travail : le rôle éventuel des contraintes phonologiques séquentielles de la langue dans la segmentation du signal de parole en mots en présentant deux approches complémentaires des processus de segmentation lexicale.