Rôle de la ségrégation séquentielle pour la séparation de voix concurrentes

Etienne Gaudrain

N°45–2008 Année 2008

THÈSE

présentée devant

l'UNIVERSITÉ LYON 1 – CLAUDE BERNARD

pour l'obtention du

DIPLÔME DE DOCTORAT

(Arrêté du 6 août 2006)

Mention : Acoustique

présentée et soutenue publiquement le 10 avril 2008

par

Etienne GAUDRAIN

Rôle de la ségrégation séquentielle pour la séparation de voix concurrentes

Directeurs de thèse : Docteur Nicolas GRIMAULT
Professeur Jean-Christophe BÉRA

Jury : Jean-Luc Schwartz, Président
Roy D. Patterson, Rapporteur
Christian Lorenzi, Rapporteur
Marie-Annick Galland
Jean-Christophe Béra
Nicolas Grimault

UNIVERSITÉ CLAUDE BERNARD - LYON I

Président de l'Université	**M. le Professeur L. COLLET**
Vice-Président du Conseil Scientifique	M. le Professeur J.F. MORNEX
Vice-Président du Conseil d'Administration	M. le Professeur J. LIETO
Vice-Président du Conseil des Etudes et de la Vie Universitaire	M. le Professeur D. SIMON
Secrétaire Général	M. G. GAY

SECTEUR SANTÉ

Composantes

UFR de Médecine Lyon R.T.H. Laënnec	Directeur : M. le Professeur P. COCHAT
UFR de Médecine Lyon Grange-Blanche	Directeur : M. le Professeur X. MARTIN
UFR de Médecine Lyon-Nord	Directeur : M. le Professeur J. ETIENNE
UFR de Médecine Lyon-Sud	Directeur : M. le Professeur F.N. GILLY
UFR d'Odontologie	Directeur : M. O. ROBIN
Institut des Sciences Pharmaceutiques et Biologiques	Directeur : M. le Professeur F. LOCHER
Institut Techniques de Réadaptation	Directeur : M. le Professeur MATILLON
Département de Formation et Centre de Recherche en Biologie Humaine	Directeur : M. le Professeur P. FARGE

SECTEUR SCIENCES

Composantes

UFR de Physique	Directeur : Mme. le Professeur S. FLECK
UFR de Biologie	Directeur : M. le Professeur H. PINON
UFR de Mécanique	Directeur : M. le Professeur H. BEN HADID
UFR de Génie Electrique et des Procédés	Directeur : M. le Professeur G. CLERC
UFR Sciences de la Terre	Directeur : M. le Professeur P. HANTZPERGUE
UFR de Mathématiques	Directeur : M. le Professeur M. CHAMARIE
UFR d'Informatique	Directeur : M. le Professeur S. AKKOUCHE
UFR de Chimie Biochimie	Directeur : Mme. le Professeur H. PARROT
UFR STAPS	Directeur : M. C. COLLIGNON
Observatoire de Lyon	Directeur : M. le Professeur R. BACON
Institut des Sciences et des Techniques de l'Ingénieur de Lyon	Directeur : M. le Professeur J. LIETO
IUT A	Directeur : M. le Professeur M. C. COULET
IUT B	Directeur : M. le Professeur R. LAMARTINE
Institut de Science Financière et d'Assurances	Directeur : M. le Professeur J.C. AUGROS

Remerciements

Cette thèse a été financée par une bourse du programme Émergence de la Région Rhône-Alpes. Le groupement d'audioprothésistes Entendre y a aussi contribué activement et je tiens à remercier Xavier Debruille et Patrick Arthaud pour leur accueil et leur soutien. Je remercie aussi Lionel Collet pour m'avoir accueilli dans son laboratoire lors de mon DEA puis de ma thèse.

Je tiens tout particulièrement à remercier Nicolas Grimault et Jean-Christophe Béra pour leur encadrement actif et respectueux, et pour le plaisir que j'ai pris à travailler avec eux. Je remercie aussi Eric W. Healy pour son importante contribution à ce travail, ses conseils et ses encouragements.

Je remercie vivement Roy D. Patterson et Christian Lorenzi d'avoir accepté d'être les rapporteurs de cette thèse.

L'Équipe Technique de l'UMR 5020 m'a apporté tout son soutien dans le développement d'outils nécessaires à la réalisation de cette thèse. En particulier, je tiens à remercier Vincent Farget pour avoir mis à ma disposition les importantes infrastructures informatiques qu'il développe et maintient au laboratoire, et sans lesquelles nous serions réduits à utiliser du papier et un crayon ! Je tiens aussi à remercier Samuel Garcia pour sa sélection de puissants outils de traitements et ses nombreux conseils concernant la programmation, et plus généralement la méthode de travail (il se débrouille aussi très bien avec un rouleau de peinture).

Toute l'Équipe CAP, qui s'est formée et structurée au cours de ces dernières années, a fortement contribué à la qualité de cette thèse. Barbara Tillmann, Nicolas Grimault et Fabien Perrin ont permis de rendre ces trois ans extrêmement formateurs sur des plans aussi bien scientifiques, philosophiques, éthiques, qu'administratifs.

Puis je veux remercier tous les gens qui ont contribué à rendre ces trois ans agréables et productifs. D'abord les doctorants et post-doctorants de l'équipe CAP : Carine Signoret pour ses longues discussions sur l'EEG et son amitié ; Aymeric Devergie pour son soutien, son écoute et son amitié ; Frédéric Marmel pour son regard critique et son verbe vif ; Katrin Schulze, Lisianne Hoch et Bénédicte Poulin-Charronnat pour tout leur soutien et leur gentillesse. Merci aussi aux cowboys sans peur du Pavillon U : Idrick Akhoun et Mikaël Ménard. Et aussi aux obscurs de l'autre côté du miroir : Tristan Cenier, Agnès Savigner, Jane Plailly, Pascaline Aimé, Marion Richard, Johan Poncelet et Nicolas Torquet pour m'avoir gratifié de leur amitié. Un merci particulier à Samy Barkatt pour m'avoir fait rêver, et pour avoir partagé avec moi ses discussions sur "l'accord".

Merci encore à tous les *apparatchiks* des JJCAAS 2006 avec qui nous avons passé de si bon moments : Émilie Geissner, Aurélie Bidet-Caulet, Claire Grataloup, Caroline Jacquier, Vincent Koehl, Laurent Saby et Arnaud Trollé.

Je dois à mes parents, Joël et Evelyne, de m'avoir doté de la curiosité nécessaire au plaisir d'apprendre, et leur témoigne toute ma gratitude pour avoir été et être encore des *parents si acceptables*. Le chemin jusqu'au doctorat est souvent long, sinueux et semé d'embûches. Mais ça peut passer pour une petite promenade digestive si on a le meilleur des amis comme coach : sans Samuel Bousard, cette thèse en serait encore au stade larvaire, ou pire. Et puis cette expérience n'aurait pas été la même sans Olaf Gainville, qui doit en ce moment même écrire ses remerciements.

Enfin le plus grand des mercis à mon épouse, Emmanuelle, pour sa patience, et toutes mes excuses à ma fille, Hannah, pour l'avoir emmenée dans ce pays peuplé d'Anglais.

Cambridge, le 18 mars 2008.

Table des matières

Introduction

Il est rare que nous parlions dans le silence. La plupart du temps, nos paroles se mélangent à d'autres sons avant de parvenir aux oreilles de nos auditeurs. Pourtant, même lorsque ces sons perturbateurs sont très importants, les auditeurs peuvent parvenir à saisir le contenu du message. Or les caractéristiques de propagation des sons ne permettent pas d'expliquer ces performances d'intelligibilité. En effet, en enregistrant plusieurs locuteurs simultanés avec un seul microphone, on peut supprimer toute la nature directionnelle de la propagation et ne conserver que les fluctuations de pression instantanées en un point de l'espace. En écoutant cet enregistrement (à l'aide d'un casque), un auditeur pourra cependant distinguer les différents locuteurs et comprendre ce qu'ils disent. Ainsi, le système auditif est capable de distinguer deux sources acoustiques, même si elles sont virtuellement placées au même point.

La psychoacoustique est un outil efficace pour comprendre comment le système auditif parvient à classifier les sources sonores et à les séparer. Au cours des 50 dernières années, des recherches ont porté sur les paramètres utilisés par le système auditif pour séparer différentes sources. Une première approche a consisté simplement à faire varier certaines caractéristiques des sons et à observer leur influence sur la capacité des auditeurs à séparer plusieurs sources, ou plus particulièrement plusieurs locuteurs (section 1.1). Cependant ces études ne permettent pas de comprendre comment le système auditif exploite ces caractéristiques pour distinguer et séparer les différentes sources.

En abordant le problème par un biais plus fondamental, il est apparu que l'auditeur a la capacité de former des *flux auditifs*. C'est-à-dire qu'il peut assembler des parties du signal acoustique, concentrer son attention sur ces parties et ignorer le reste du signal. Il lui est même possible de faire passer

volontairement son attention d'un sous-ensemble du signal à un autre. Ce phénomène a d'abord été mis en évidence avec des sons simples de synthèse, ce qui a permis d'aboutir à la théorie de l'*analyse des scènes auditives* (section 1.1.3). Cette théorie a conduit à dégager deux familles de mécanismes impliqués dans la formation de flux auditifs : la ségrégation simultanée et la ségrégation séquentielle. En étudiant ces mécanismes, certaines caractéristiques acoustiques permettant à des sons d'être séparés ont pu être associées à des variables ou fonctions auditives. Ainsi, il a été montré que la sélectivité fréquentielle était un des facteurs déterminant pour la perception de la hauteur fondamentale (section 1.3). Outre la clarification de nos connaissances sur le fonctionnement de l'appareil auditif, ces études ont ouvert des pistes de réflexion pour améliorer la réhabilitation des malentendants, et plus particulièrement en milieu bruyant.

Depuis l'apparition de la théorie de l'analyse des scènes auditives, la ségrégation simultanée et la ségrégation séquentielle ont été étudiées séparément. Pourtant, quand plusieurs personnes parlent en même temps, les événements sonores ne sont ni purement simultanés, ni purement séquentiels. Dans les situations naturelles, si ces deux mécanismes sont réellement impliqués alors ils interagissent, ou tout du moins ils coexistent. Pour comprendre comment s'opère cette interaction, on peut faire converger l'étude de la ségrégation simultanée et l'étude de la ségrégation séquentielle vers l'étude de situations plus écologiques où deux locuteurs sont en compétition. Cependant, si la ségrégation simultanée a été largement étudiée avec des signaux de parole, la ségrégation séquentielle, quant à elle, n'a bénéficié que de très peu d'études impliquant de tels sons. Les conséquences que peuvent avoir les spécificités des signaux de parole sur la ségrégation séquentielle sont donc largement méconnues. Dans l'objectif de rapprocher la ségrégation simultanée et la ségrégation séquentielle de la perception de voix concurrentes, la première étape consiste donc à éclaircir le phénomène de ségrégation séquentielle pour des signaux de parole (section 1.2).

Les études présentées dans ce qui suit visent plus spécifiquement à clarifier le rôle de la hauteur dans la ségrégation séquentielle de voyelles. La hauteur fondamentale est ainsi la seule dimension acoustique a avoir été manipulée, mais les indices perceptifs disponibles pour les auditeurs ont varié selon les études. Ce sont donc les fonctions auditives employées pour la sé-

grégation séquentielle de voyelles différant par leur hauteur fondamentale qui ont été recherchées. Une méthodologie spécifique a d'abord été mise en place et validée (chapitre 2, Experiment 1). Le rôle de la sélectivité fréquentielle a ensuite été exploré à travers des simulations de lissage spectral (chapitre 2, Experiment 2) et d'implant cochléaire (chapitre 3). Le rôle des indices temporels de hauteur a été évalué (chapitre 3). La relation avec la perception de la parole dans le bruit a été étudiée pour produire une première estimation de l'implication de la ségrégation séquentielle dans la perception de voix concurrentes (chapitre 4). Enfin, la variabilité de sélectivité fréquentielle qui existe dans la population a été exploitée pour en observer l'effet sur la ségrégation séquentielle de séquences de voyelles (chapitres 4 et 5).

Le chapitre 1 présente l'essentiel des concepts abordés et discutés dans les chapitres suivants. Par choix, la bibliographie qui y est présentée n'est pas exhaustive. Il s'agit de références choisies soit pour leur aspect novateur (le premier article à avoir utilisé une méthode, ou mis en évidence un phénomène), soit pour leur pertinence (l'article qui illustre le mieux un phénomène), soit pour leur récence. Les chapitres 2 à 4 ont la forme d'articles et sont donc rédigés en anglais.

Chapitre 1

Perception de voix concurrentes et ségrégation séquentielle

Les sons que nous percevons proviennent généralement d'un ensemble de sources acoustiques. Les ondes acoustiques émises par chacune de ces sources se propagent jusqu'à l'entrée de l'oreille où elles se combinent pour former un signal sonore unique appelé *mixture* (Bregman, 1990). Cependant, l'auditeur ne perçoit généralement pas cette mixture comme un signal unique. De façon volontaire ou non, l'auditeur réalise une analyse et une décomposition du signal en objets sonores qui correspondent idéalement à chacune des sources acoustiques. La situation d'écoute la plus commune est une situation où l'auditeur cherche à comprendre ce qu'un locuteur dit alors que ses paroles sont noyées dans un bruit de fond. Il est généralement fait référence à ce genre de situation sous l'appellation *cocktail party* (Cherry, 1953).

Ce phénomène peut être étudié simplement en observant l'intelligibilité d'un message parlé présenté dans un bruit de fond. Cette approche très similaire à une situation écologique permet de déterminer les principales caractéristiques acoustiques qui permettent à un son d'être extrait de la mixture (*cf.* section 1.1). En revanche, elle ne permet pas d'observer les mécanismes impliqués dans cette tâche. Pour étudier plus en détails ces mécanismes, une autre approche est utilisée, théorisée sous l'appellation d'analyse des scènes auditives (ASA, *cf.* section 1.1.3).

1.1 Perception la parole dans le bruit

1.1.1 Cocktail party et situation expérimentale

La parole est composée de sons voisés (les voyelles et certaines consonnes) et de sons non voisés (consonnes). Les voyelles sont des sons complexes harmoniques relativement stationnaires dont la fréquence fondamentale est comprise entre 80 et 400 Hz environ, et dont l'enveloppe spectrale présente des pics appelés formants. Les positions spectrales des formants caractérisent les phonèmes et permettent, notamment, de distinguer les différentes voyelles qui composent la langue. Les éléments quasi-stationnaires de la parole contiennent d'autres dimensions caractéristiques que l'on peut regrouper sous l'appellation de *timbre*. Les consonnes, au contraire, sont plutôt caractérisées par leur nature transitoire.

Pour étudier la perception de voix concurrentes, le protocole généralement employé consiste à rapporter une cible (mot ou phrase) présentée simultanément à un masque. Cette procédure permet de mesurer une performance d'identification en fonction d'une variable, classiquement le rapport entre les niveaux de la cible et du masque exprimé en décibels (soit le rapport "signal sur bruit" ou SNR, pour *signal-to-noise ratio*). Cette mesure conduit à l'estimation d'un seuil de perception (SRT, pour *speech reception threshold*) qui est défini le plus souvent comme la valeur du SNR qui donne 50% de bonnes réponses. Le SRT ou les scores d'identification moyens sont classiquement utilisés comme mesure de performance de ségrégation de la cible dans le masque. La nature de la cible, mais surtout la nature du masque qui interfère avec la cible, influencent fortement les résultats observés.

Le masque le plus simple qui peut être envisagé est un bruit blanc (par exemple Hawkins et Stevens, 1950). Ce type de son est facile à générer et à manipuler mais ne correspond pas à la situation réelle la plus courante. En effet, dans une situation réelle, le bruit masquant est souvent produit par d'autres locuteurs ou par d'autres sources ayant des caractéristiques proches comme les instruments de musique. Le masque et la cible partagent alors un certain nombre de caractéristiques acoustiques comme la présence d'harmoniques, la présence de formants... La parole a un spectre relativement large bande mais dans lequel toutes les fréquences ne présentent pas la même énergie, comme illustré sur la figure 1.1 où le spectre à long terme de la parole

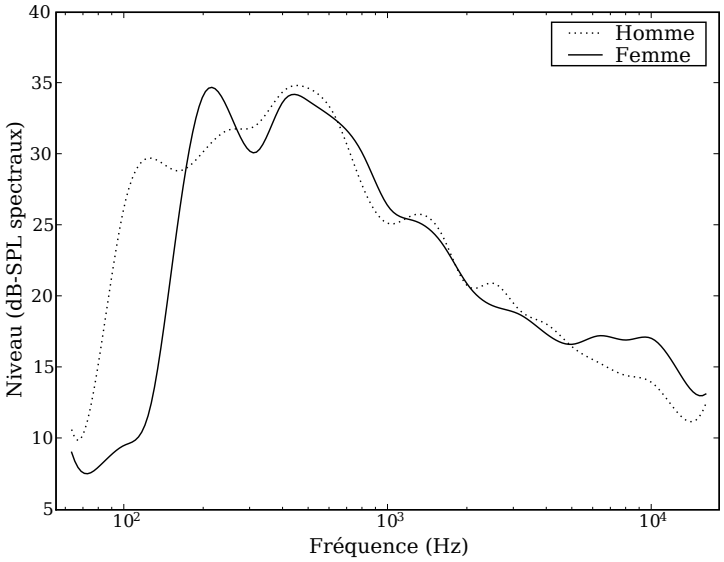

FIG. 1.1 – Spectres à long terme de la parole pour un homme (pointillés) et une femme (trait continu), moyennés sur 17 langues différentes. D'après Byrne *et al.* (1994).

est représenté. Dans une étude visant à comparer l'effet de différents types de masques sur la perception de la parole, Miller (1947) conclut que le bruit masquant le plus efficace est un bruit dont le spectre à long terme est celui de la parole (*speech-shaped noise*). L'auditeur peut tirer profit de "trous" dans le spectre du masque si ces trous interviennent à des plages de fréquence dans lesquelles la parole contient de l'énergie. Peters, Moore, et Baer (1998) ont observé, chez de jeunes normo-entendants, une amélioration des SRT d'environ 9 dB en ajoutant des trous spectraux d'une largeur équivalente à $2 \times \mathrm{ERB}_N$[1] dans un bruit façonné par la parole. Le contenu spectral du masque conditionne donc fortement l'intelligibilité de la cible. Cependant, contrairement à la parole qui présente d'importantes fluctuations temporelles d'amplitude, les bruits précédemment évoqués sont quasi-stationnaires. Pour produire un masque quasi-stationnaire avec un signal de parole, il faut ajouter un nombre

[1]Les ERB (pour *Equivalent Rectangular Bandwidth*) représentent la largeur des filtres auditifs (Glasberg et Moore, 1990). L'indice N indique qu'il s'agit de filtres standardisés pour les normo-entendants qui sont calculés par la formule : $\mathrm{ERB}_N = 10,8 \cdot 10^{-2}\ f + 24,7$.

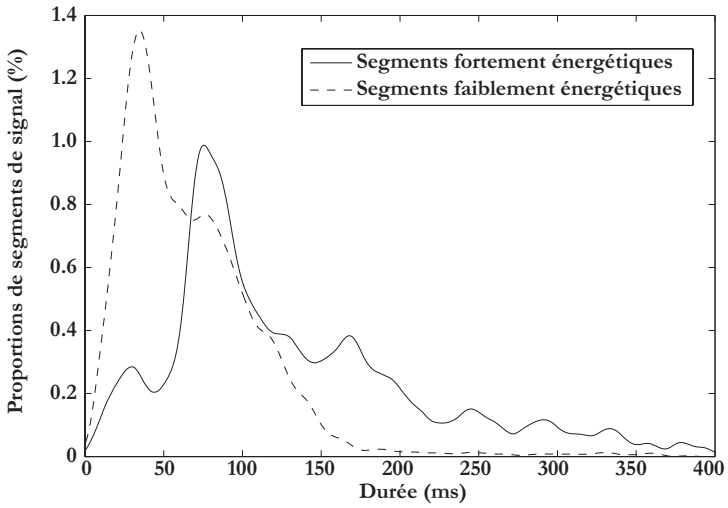

FIG. 1.2 – Distribution des durées des segments les plus énergétiques (en trait continu) et des moins énergétiques (trait tireté) dans la parole. Cette figure a été obtenue en extrayant l'enveloppe à 15 Hz d'un texte lu par un homme. Lorsque le niveau de cette enveloppe dépassait un seuil donné, le signal était considéré comme contenant de l'énergie. En dessous de ce seuil, le signal est considéré comme du silence. L'axe des ordonnées représente la proportion de portions de signal trouvés selon leur durée en abscisse. Le seuil utilisé était un seuil variable dérivé de l'enveloppe à 3 Hz du même signal. Le pic pour les segments énergétiques se situe autour de 76 ms.

suffisant de locuteurs. Il a en effet été observé que les performances d'identification d'une cible dans un masque constitué de plus de 4 locuteurs étaient similaires à celles obtenues pour un masque constitué d'un bruit façonné par la parole (Bronkhorst, 2000).

Les signaux de parole ne sont pas stationnaires et l'énergie qu'ils contiennent varie rapidement au cours du temps. En particulier, les voyelles sont relativement énergétiques tandis que les consonnes sont pour la plupart peu énergétiques. La figure 1.2 illustre ces alternances de portions de signal énergétiques et de portions peu énergétiques. Miller (1947) avait noté que la continuité temporelle du masque était une caractéristique importante dans l'évaluation du SRT. En hachant un bruit grâce à un interrupteur électronique, Miller (1947) avait observé que si le bruit n'était présent que 80% du temps, le SRT était amélioré d'environ 10 dB. Pour se placer dans une situation plus écologique, le masque peut être modulé en amplitude de façon

à reproduire les fluctuations temporelles de la parole. Peters *et al.* (1998) ont modulé l'amplitude d'un bruit façonné par la parole avec l'enveloppe temporelle d'une phrase. Cette enveloppe était calculée en évaluant l'amplitude[2] du signal avec une constante de temps de 10 ms. Ils ont observé que moduler ainsi le bruit améliorait le SRT de plus de 6 dB chez de jeunes normo-entendants. Il faut noter que dans ce cas, le niveau du bruit a été évalué comme le niveau moyen sur toute la durée du masque. Le niveau maximal atteint à un instant par ce bruit modulé était par conséquent plus grand que le niveau moyen du bruit stationnaire. Dans ces conditions, il est difficile de comparer ces SRT. Néanmoins, en utilisant un locuteur unique comme masque, Peters *et al.* (1998) ont obtenu un SRT de 2 dB meilleur qu'avec le bruit modulé en amplitude. Ce bruit façonné par la parole aussi bien spectralement que temporellement, rend donc relativement bien compte de l'effet de masquage produit par un seul locuteur.

Cependant, l'usage de bruits ou d'un mélange de plusieurs locuteurs pour masquer la cible ne permet pas d'étudier l'effet de différences temporelles ou spectrales fines sur la ségrégation perceptive de la cible. En particulier, les sons de parole étant largement voisés, leur structure spectrale à court terme est organisée en raies harmoniques. La superposition de deux sons voisés de hauteur fondamentale différente ne produit pas le même effet que la superposition d'un son voisé avec un bruit, même modulé en amplitude. Afin de mieux étudier ces phénomènes de masquage par un locuteur unique, il convient de prendre en compte la différence de fréquence fondamentale entre la cible et le masque.

1.1.2 Influence de la fréquence fondamentale

Brokx et Nooteboom (1982) ont spécifiquement étudié l'influence de la fréquence fondamentale (F_0) sur la séparation de deux voix concurrentes. Ces auteurs ont employé plusieurs minutes de parole prononcée par un locuteur (une histoire lue), et dont le F_0 a été ajusté à 100 Hz à l'aide d'un algorithme basé sur la prédiction linéaire. Ce flux de parole constituait le masque. Le même locuteur a par ailleurs prononcé des phrases constituées de mots monosyllabiques et ayant toutes la même syntaxe. Toutes les phrases

[2]le RMS pour *Root Mean Square*, soit $\sqrt{< x^2 >_t}$.

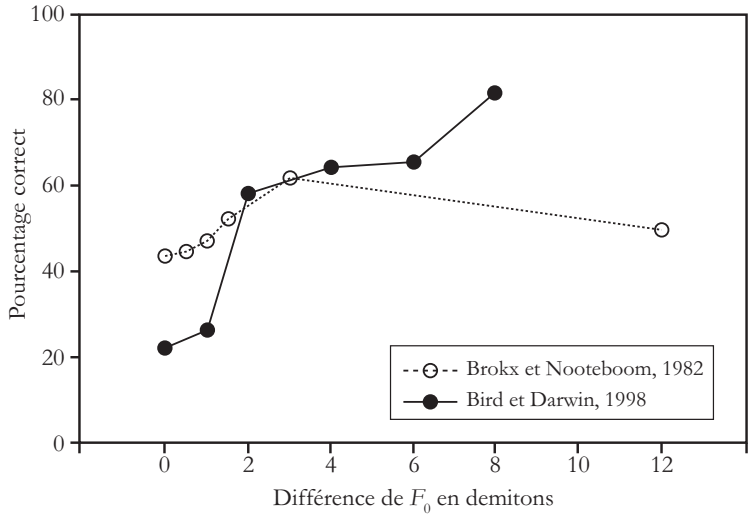

FIG. 1.3 – Pourcentage d'identification correcte en fonction de la différence de F_0 entre la cible et le masque. Les cercles et le trait tireté représentent les performances observées par Brokx et Nooteboom (1982). Les disques noirs et le trait plein représentent les performances observées par Bird et Darwin (1998). D'après Bird et Darwin (1998).

étaient sémantiquement incongrues. De la même façon que pour le masque, la hauteur fondamentale de ces phrases a été manipulée pour atteindre 100, 103, 106, 109, 120 et 200 Hz. Ces phrases constituaient la cible que les sujets devaient répéter. La figure 1.3 montre les résultats qui ont été obtenus. Les scores d'identification progressent de 40 à 60% lorsque la différence de F_0 entre la cible et le masque augmente de 0 à 3 demitons. Pour une différence d'une octave, le score est d'environ 50%. Les sujets tirent donc profit de l'augmentation de la différence de F_0 entre la cible et le masque. Brokx et Nooteboom (1982) ont aussi observé que ce profit perdurait quand la hauteur des voix n'était pas constante mais que l'intonation était préservée. La différence de F_0 est alors une différence moyenne. Ceci indique plus clairement que l'indice de hauteur fondamentale peut être exploité dans des situations écologiques.

Afin de préciser l'importance de cet effet, Bird et Darwin (1998) ont réalisé le même genre d'expérience, mais en utilisant comme masque un flux de parole ne contenant que peu de consonnes formant des ruptures abruptes du signal (plosives, fricatives). En effectuant cette manipulation, c'est-à-dire

en réduisant la proportion de sons non voisés dans le masque, ces auteurs espéraient amplifier l'effet de F_0 sur la séparation de voix concurrentes. Les résultats obtenus sont présentés dans la figure 1.3. Ils montrent un bénéfice lié à la différence de F_0 (ΔF_0) qui augmente quasi linéairement de 20 à 80% de bonnes réponses pour des ΔF_0 de 0 à 8 demitons. Ces études suggèrent donc, dans le cas où le masque est constitué d'un seul locuteur, un effet déterminant du différentiel de hauteur fondamental (ΔF_0) entre le masque et la cible.

Néanmoins, ces études ne permettent pas d'établir quels mécanismes cognitifs et sensoriels sont impliqués dans la ségrégation de voix concurrentes. Pour pouvoir étudier plus précisément ces mécanismes sous-jacents, une autre approche a été développée : l'analyse des scènes auditives.

1.1.3 Analyse des scènes auditives

Selon Bregman (1990), pour parvenir à séparer deux locuteurs, l'auditeur doit résoudre l'*analyse de la scène auditive* (ASA), c'est-à-dire qu'il doit séparer la mixture en *flux auditifs*. Un flux auditif peut être défini comme l'information auditive associée à la représentation que se fait un auditeur d'une source acoustique. Pour effectuer cette analyse, l'auditeur utilise des informations qui doivent caractériser les différentes sources acoustiques qu'il tente d'isoler. C'est la représentation perceptive de ces traits acoustiques que l'on appelle *indices* perceptifs. L'auditeur collecte donc des indices qui lui permettent de séparer la mixture en flux auditifs, chaque flux correspondant donc, en cas de succès total de l'analyse, à une source acoustique. Pour parvenir à effectuer cette séparation, Bregman propose une transposition des lois de la théorie de la forme (*gestalt-theory*) établies pour la vision : bonne forme, continuité, proximité, similitude, destin commun et clôture. Les dimensions physiques de l'espace dans lequel ces lois sont projetées restent cependant à déterminer.

Il faut noter que la notion de "séparation" convient bien lorsqu'il s'agit d'observer des sources acoustiques physiquement séparées. Mais chaque son émis par chaque source est lui-même constitué de plusieurs primitives qui doivent donc être groupées dans un seul flux auditif. Par exemple, un son complexe harmonique est composé de plusieurs harmoniques qui vont exciter des zones différentes de la cochlée. La séparation de voix concurrentes consiste

donc à la fois à grouper les éléments sonores provenant d'une même source, et à les séparer des éléments provenant d'autres sources. Dans le cas de la parole, les différentes harmoniques et les différents formants constituant les sons voisés doivent être groupés entre eux, mais séparés des harmoniques provenant d'un autre locuteur.

Bregman (1990) a fondé la théorie de l'ASA sur les bases de deux grandes classes de mécanismes : les mécanismes de ségrégation simultanée, qui traitent les événements sonores simultanés ; et les mécanismes de ségrégation séquentielle, qui traitent les événements sonores ne se recouvrant pas dans le temps. Bregman propose aussi une autre catégorisation des mécanismes. Lorsque le processus de groupement/séparation se fait en groupant les éléments qui partagent un même indice, on parle de processus *primitif*. Ce type de processus est généralement considéré comme *bottom-up*, c'est-à-dire que l'information permettant la ségrégation circule de la cochlée vers le cortex auditif primaire. Ces mécanismes devraient permettre d'exploiter de façon automatique les indices extraits par les voies auditives primaires. En revanche, quand le processus de ségrégation se fait en exploitant des connaissances induites par des régularités dans les indices perceptifs, on parle de mécanisme *schema-based*. Ce type de mécanisme est considéré comme *top-down* et doit permettre d'exploiter nos connaissances *a priori* des sons qui nous sont familiers. Cependant, il est probable que la plupart des mécanismes auditifs impliquent une boucle de contrôle ou d'adaptation. Il devient alors difficile d'opposer strictement des processus *bottom-up* à des processus *top-down*.

Lorsque deux phrases sont prononcées simultanément, on se trouve dans une situation intermédiaire où la plupart des événements sonores ne sont ni réellement simultanés ni réellement séquentiels. Les deux mécanismes de ségrégation, simultané et séquentiel, sont donc vraisemblablement impliqués dans la ségrégation de voix concurrentes et y interagissent certainement. Cependant, les études menées jusqu'à maintenant sur ces mécanismes de ségrégation sont encore loin de faire le lien entre ASA et perception de voix concurrentes. Autrement dit, il est encore difficile d'interpréter les résultats obtenus pour des tâches de ségrégation simultanée ou séquentielle en termes de séparation de voix concurrentes.

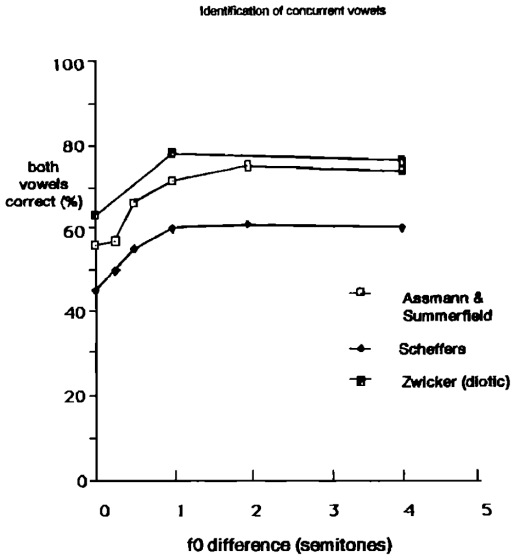

FIG. 1.4 – Capacité des sujets à identifier correctement deux voyelles simultanées en fonction de leur différence de F_0 pour trois études, dont Assmann et Summerfield (1990). Reproduit depuis Meddis et Hewitt (1992).

Ségrégation simultanée

La ségrégation simultanée a été largement étudiée au travers du paradigme des doubles voyelles (voir de Cheveigné, 1999, pour une revue), et a été longtemps considérée comme le mécanisme reflétant le mieux la séparation de voix concurrentes. Outre le rapport des intensités, les deux paramètres principaux pilotant ce phénomène sont la différence de F_0 entre les voyelles et la synchronisation des fronts montants (*onsets*) et descendants (*offsets*, voir par exemple Darwin, 1984). La figure 1.4 montre les résultats obtenus par Assmann et Summerfield (1990), ainsi que ceux de deux autres études, lorsqu'il est demandé aux participants de redonner les deux voyelles entendues. Dans cette étude, des voyelles de 200 ms étaient présentées simultanément. Les voyelles différaient par leur F_0 de 0 à 4 demitons. Les résultats montrent que le bénéfice maximal (18%) apporté par la différence de F_0 est atteint avant 2 demitons.

L'étude de la ségrégation simultanée a donné lieu à plusieurs modèles (par exemple Meddis et Hewitt, 1992), et a permis, en particulier, d'éclaircir nos connaissances sur les mécanismes de perception de la hauteur fondamentale (de Cheveigné, 2005). Cependant, l'implication de la ségrégation simultanée dans la perception de voix concurrentes ne semble pas manifeste. L'effet du ΔF_0 sur la séparation de voix concurrentes croît continûment de 0 à

8 demitons. Au contraire, Meddis et Hewitt (1992) ont établi leur modèle sur des données moyennes qui placent à 1 demiton l'effet maximal du ΔF_0. Cette différence d'échelle suggère que le mécanisme de ségrégation simultanée ne peut à lui seul permettre de quantifier les performances de ségrégation de voix concurrentes. Comme on le verra dans la section 1.3, l'étude de ce phénomène auprès des malentendants apporte un soutien supplémentaire à cette hypothèse. De nombreuses études ont permis d'observer le phénomène de ségrégation séquentielle pour des ΔF_0 allant de 0 à 12 demitons, suggérant ainsi que ce second mécanisme joue bien un rôle dans la séparation de voix concurrentes. Les investigations présentées dans cette thèse ont donc pour objet de compléter l'étude de ce mécanisme.

1.2 Ségrégation séquentielle

Le premier article à décrire le phénomène de ségrégation séquentielle est probablement celui de Miller et Heise (1950). Ces auteurs ont observé que l'alternance de deux sons qui différaient par leur fréquence pouvait produire des percepts différents selon la différence de fréquence ΔF. Lorsque ΔF était suffisamment petit, les sons semblaient venir d'une seule source. En revanche, lorsque ΔF était suffisamment grand, les sons semblaient venir de deux sources incohérentes. Miller et Heise ont appelé "*trill threshold*" (littéralement, "seuil de trille") le ΔF auquel la perception bascule d'une à deux sources. Ce phénomène de ségrégation séquentielle est plus communément appelé *streaming* pour illustrer que des événements sonores temporellement disjoints peuvent être regroupés dans un même flux auditif, comme illustré dans la figure 1.5. C'est le terme qui sera employé dans la suite de cette thèse. À la suite de Miller et Heise (1950), de nombreuses études impliquant des sons purs et des sons complexes harmoniques ont été réalisées. En particulier, de nombreuses propriétés du phénomène de streaming ont été explorées par van Noorden (1975) et Bregman (1990).

1.2.1 Streaming de sons purs

Plusieurs méthodes ont été développées par les chercheurs afin d'étudier le streaming. La méthode la plus communément employée se base sur le juge-

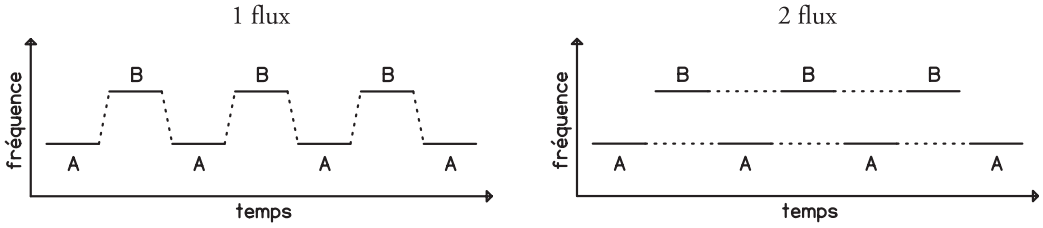

FIG. 1.5 – Deux séquences de sons A et B intercalés qui diffèrent par leur fréquence. *À gauche* : un seul flux est perçu, il y a fusion. *À droite* : deux flux sont perçus, il y a ségrégation.

ment subjectif du sujet. Elle consiste à construire des séquences de deux sons successifs : un son A et un son B. Ces sons sont assemblés en séquences sous la forme ABABAB comme dans la figure 1.5. Le sujet doit alors indiquer s'il perçoit un flux contenant les deux sons différents, ou deux flux distincts, l'un composé de A–A–A– (le symbole – représente un silence) et l'autre composé de –B–B–B.

Cohérence et fusion

Cependant, van Noorden (1975) a mis en évidence que cette méthode ne permettait pas de contrôler précisément la tâche réellement effectuée par le sujet et a mis au point une méthode qui améliorait ce contrôle. En utilisant des séquences de la forme ABA–ABA–, un rythme de "galop" est perçu quand la séquence est fusionnée en un seul flux. Au contraire, en cas de fission de la séquence, les deux flux perçus sont réguliers : A–A–A–A– et –B– – –B– –. En demandant au sujet d'essayer d'entendre le rythme de galop (1 flux) aussi longtemps qu'il le peut alors que la différence de fréquence entre les sons A et B augmente, on obtient un seuil au-delà duquel la perception d'un seul flux est impossible. Ce seuil est appelé seuil de cohérence temporelle (TCB) et il reflète une situation où la ségrégation est irrépressible. Ce seuil refletèrait des mécanismes primitifs. Si, au contraire, on demande au sujet d'essayer d'entendre le rythme régulier formé par la suite de sons A, tout en réduisant la différence de fréquence qui sépare les deux types de sons, on obtient un autre seuil au delà duquel il est impossible d'entendre deux flux. Ce seuil est appelé le seuil de fission (FB) et reflète une situation où la ségrégation est volontaire. Il reflèterait des mécanismes basés sur la reconnaissance de forme sonore

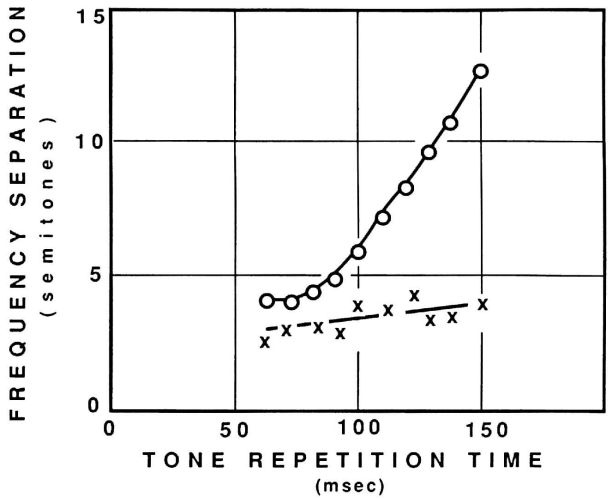

FIG. 1.6 – Seuil de cohérence temporelle (∘) et seuil de fission (×) en fonction du tempo. En-dessous du seuil de fission, un seul flux est toujours perçu. Au-dessus du seuil de cohérence temporelle, deux flux sont toujours perçus. Entre ces deux seuils, il y a incertitude que le sujet peut lever selon qu'il cherche à percevoir un flux ou deux. D'après van Noorden (1975), reproduit par Bregman (1990).

(*schema-based*). Van Noorden (1975) a exploré ces deux seuils pour différents tempos et a observé que le seuil de fission était relativement indépendant du tempo, tandis que le seuil de cohérence temporelle montrait une forte interaction entre fréquence et tempo, comme illustré dans la figure 1.6. La dépendance ou non d'un phénomène de streaming au tempo permet d'évaluer si la mesure concerne le seuil de cohérence temporelle ou le seuil de fission.

La mesure de ces seuils ne présente cependant qu'une vue limitée du phénomène. Le profil de la limite entre la perception d'un flux et de deux est en effet sujet à de nombreux paramètres. Réduire ce profil à une seule valeur de seuil ne permet pas de rendre compte de ces paramètres ni d'estimer comment le percept évolue dans le temps.

Construction des flux auditifs

Bregman (1978) a utilisé une méthode similaire pour mettre en évidence que l'effet de streaming nécessitait une phase de construction (*build-up*). Il a montré qu'il fallait qu'un sujet soit exposé environ 4 s à la séquence pour que les flux se forment. Ce temps relativement long, comparé aux durées carac-

téristiques des processus cognitifs, serait nécessaire à l'accumulation d'indices permettant la construction des flux. De la même façon, un certain temps est nécessaire à la dé-construction des flux. Le fait d'interrompre le signal pendant quelques secondes ne supprime pas entièrement l'effet de streaming. Ainsi, quand le signal reprend, il est possible d'observer un biais sur l'effet de streaming qui dépend de la durée du silence. L'existence de cette phase de construction (ou de dé-construction) est considérée comme une caractéristique du streaming et peut être utilisée pour distinguer un phénomène de simple discrimination, qui ne nécessite pas un tel temps de construction, d'un phénomène de streaming (voir aussi Anstis et Saida, 1985).

Récemment, il a été argumenté que la perception de flux auditifs pouvait donner lieu à des situations ambigües bistables (Pressnitzer et Hupé, 2006). Un certain temps est nécessaire pour qu'une séquence se sépare en deux flux auditifs, mais après un temps plus long, il est possible que la séquence soit perçue à nouveau comme un seul flux. Ce phénomène de bistabilité étant décrit comme partiellement irrépressible, on peut estimer qu'il existe toujours une possibilité que le nombre de flux perçus alterne, quelque soit la consigne donnée au sujet.

Perception de l'ordre

Une autre caractéristique importante du streaming est son effet sur la perception de l'ordre dans lequel sont présentés les éléments constitutifs de la séquence. Bregman et Campbell (1971) ont observé que lorsqu'une séquence donne lieu à la perception de deux flux distincts, le jugement de l'instant d'apparition d'un élément perçu dans un flux par rapport à un élément perçu dans l'autre flux devient impossible. C'est-à-dire que les positions temporelles relatives des éléments ne peuvent être perçues qu'au sein d'un même flux. La notion d'ordre temporel est donc perdue entre les flux perceptifs, mais conservée à l'intérieur de chaque flux.

1.2.2 Streaming de sons complexes

La plupart des sons de notre environnement, et notamment la parole, sont des sons complexes harmoniques. Ils ne peuvent donc pas être caractérisés uniquement par leur fréquence comme les sons purs. L'essentiel des caractéris-

tiques d'un son complexe harmonique peut être représenté par l'équation :

$$s(t) = \sum_k a_k \ \sin(2\pi \ k \cdot F_0 \ t + \phi_k)$$

Les différentes composantes, appelées harmoniques, ont chacune une amplitude a_k et une phase ϕ_k quelconque. En revanche, leur fréquence est un multiple entier de la fréquence fondamentale F_0. La perception de la hauteur fondamentale est plutôt définie par la perception d'au moins deux composantes consécutives que par la perception de la première harmonique ($k = 1$). Néanmoins, même lorsque la première harmonique n'est pas présente ($a_1 = 0$), la hauteur fondamentale peut être perçue, ce qui illustre le concept de hauteur virtuelle. Les a_k et ϕ_k définissent la forme temporelle de l'onde sonore. Ces paramètres sont généralement utilisés pour caractériser le timbre d'un son complexe harmonique, mais peuvent aussi agir sur la perception de la hauteur fondamentale.

Streaming sur la base de la hauteur fondamentale

Le streaming avec des sons complexes a d'abord été étudié avec des sons ayant des timbres similaires mais des hauteurs fondamentales différentes. Van Noorden (1975) a observé qu'une différence de hauteur fondamentale pouvait induire du streaming de la même façon qu'une différence de fréquence entre deux sons purs. La figure 1.7 montre les seuils de cohérence obtenus pour des sons complexes et des sons purs. Les sons complexes utilisés contenaient toutes les harmoniques dans la gamme audible, avec la même amplitude ($\forall k > 1, a_k = a_1$).

Cette étude ne permet pas de distinguer si le streaming observé résulte d'une différence de hauteur fondamentale perçue, ou s'il s'agit de streaming de sons purs résultant des différences de fréquence entre les harmoniques composant ces deux sons. Par exemple, l'effet de streaming pourrait être simplement dû au ΔF qui existe entre les premières harmoniques de chacun des sons A et B. En contrôlant indépendamment l'amplitude des harmoniques — c'est-à-dire le timbre — et la fréquence fondamentale, il est possible de mieux cerner ce problème.

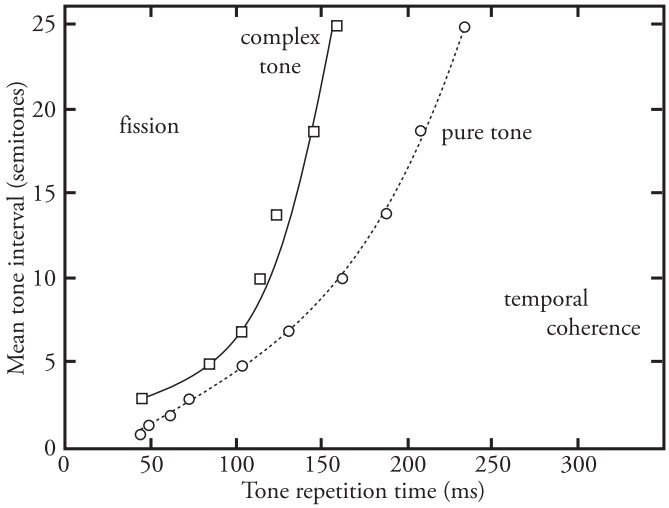

FIG. 1.7 – Seuil de cohérence temporelle pour des sons complexes harmoniques (carrés et trait continu) et des sons purs (cercles et trait tireté) en fonction du tempo. D'après van Noorden (1975).

Effet du timbre

Singh (1987) a utilisé des sons complexes constitués de 4 harmoniques d'égale amplitude. Ces sons pouvaient varier selon leur F_0, de 262 Hz à 1048 Hz, et selon leur timbre. Le timbre était manipulé en utilisant les harmoniques 1–4, 2–5, ..., ou 7–10. Les résultats montrent une interaction entre les effets de la hauteur et du timbre. Néanmoins, pour de petites différences de timbre et de hauteur, il semble que la hauteur virtuelle domine le timbre dans sa capacité à créer des flux auditifs. Le fait qu'il soit possible de distinguer les effets du timbre des effet de la hauteur fondamentale suggère qu'il existe bien un mécanisme de streaming basé sur la hauteur fondamentale.

Cette manipulation basique du timbre permet bien d'éclaircir le rôle de la fréquence fondamentale, mais ne permet pas d'examiner l'influence de variations de timbre plus écologiques. Bregman, Liao, et Levitan (1990) ont réalisé des expériences similaires à celle de Singh (1987), mais ont manipulé le timbre en déplaçant spectralement un formant unique. Ils ont observé que ce type de timbre pouvait aussi induire du streaming. Contrairement à Singh, ces auteurs ont suggéré que le timbre avait un effet plus puissant que la hauteur fondamentale, et que ces deux effets étaient indépendants plutôt

qu'en compétition. Un examen plus détaillé de cette étude est proposé au chapitre 2.

1.2.3 Streaming et perception de voix concurrentes

Les études décrites précédemment ont permis de cerner les caractéristiques importantes du streaming. Néanmoins, il est difficile d'évaluer l'implication de ce mécanisme dans la séparation de voix concurrentes lorsqu'il n'est étudié qu'avec des sons purs ou des sons complexes harmoniques. Il n'existe que très peu d'études de streaming utilisant des signaux de parole, et la plupart d'entre elles n'étaient pas initialement destinées à étudier ce phénomène.

Streaming sur la base de la structure formantique

De la même façon que des différences de timbre peuvent induire du streaming (Bregman *et al.*, 1990; Singh, 1987), des séquences de voyelles induisent aussi du streaming basé sur la structure formantique. Thomas, Hill, Carroll, et Garcia (1970) ont observé qu'il était impossible de donner dans l'ordre des séquences de 4 voyelles (/i ɛ a u/) lorsque la durée des voyelles était inférieure à 125 ms (tempo $>$ 8 voy/s). Lackner et Goldstein (1974) ont effectué une expérience similaire en utilisant des séquences composées de voyelles isolées (V) et de groupes consonne-voyelle (CV) de la forme CV-V-CV-V. Ils ont observé que les sujets ne parvenaient pas à percevoir ces séquences dans l'ordre et ont interprété leurs résultats en terme de streaming, les voyelles V constituant un flux, et les CV un autre flux.

Ce phénomène a été plus amplement décrit par Dorman, Cutting, et Raphael (1975). Ces auteurs ont supposé que les séquences de voyelles isolées avaient tendance à produire du streaming parce qu'elles n'étaient pas connectées par des transitions formantiques comme dans un flux de parole réel. Ils ont réalisé des séquences de 4 voyelles (/i ɔ u æ/) et ont manipulé les transitions formantiques entre ces voyelles, comme illustré sur la figure 1.8. Comme Thomas *et al.* (1970), ils ont demandé aux sujets de redonner les 4 voyelles dans l'ordre, les séquences étant présentées en boucle durant 30 s. Les résultats montrent que des discontinuités, essentiellement au niveau du premier formant, induisent la séparation de la séquence en deux flux et qu'il est

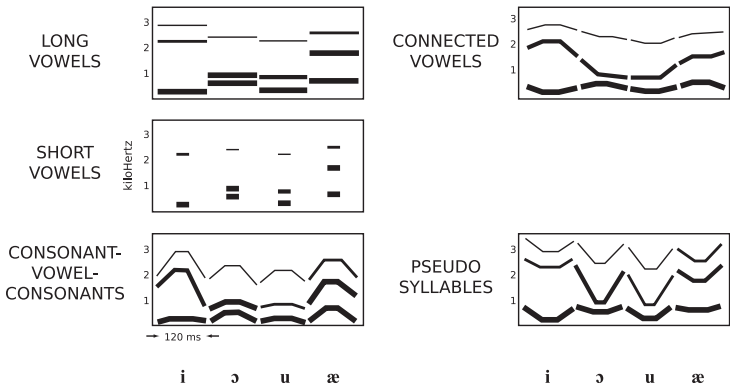

FIG. 1.8 – Représentation schématique des sonogrammes des différents types de séquences utilisés par Dorman *et al.* (1975).

impossible de donner les 4 voyelles dans l'ordre. En restaurant la continuité des formants, la ségrégation était fortement réduite et l'ordre des séquences était correctement perçu. Les sujets ont rapporté que deux des six séquences étaient particulièrement difficiles à percevoir dans l'ordre car elles se cassaient en deux groupes de deux voyelles. Il s'agissait des séquences /i æ u ɔ/ et /i ɔ u æ/ qui étaient perçues comme deux groupes sans relation : /i–u/ et /æ–ɔ/. Comme on peut le voir dans la figure 1.8, ces voyelles sont associées en paires selon la similarité de leur premier formant. Pour vérifier cette hypothèse, on peut réaliser un modèle simpliste en calculant une distance formantique entre deux voyelles successives comme la différence en demitons entre le premier formant d'une voyelle (F_i) et le premier formant de la voyelle suivante dans la séquence (F_{i+1}). La distance pour la séquence entière est définie comme la moyenne de ces distances par paire, en tenant compte du fait que la séquence est bouclée ($F_5 = F_1$) :

$$d^2 = \frac{1}{4} \sum_{i=1}^{4} \left(12 \cdot \frac{\ln(F_i) - \ln(F_{i+1})}{\ln(2)} \right)^2$$

Il est aussi possible de définir la même distance pour chacun des flux créé en cas de ségrégation, en supposant que ces flux sont constitués d'une voyelle sur deux :

$$e_1 = 12 \cdot \frac{\ln(F_1) - \ln(F_3)}{\ln(2)} \quad \text{, et} \quad e_2 = 12 \cdot \frac{\ln(F_2) - \ln(F_4)}{\ln(2)}$$

TAB. 1.1 – Distances formantiques (en demitons) des séquences employées par Dorman *et al.* (1975), et scores obtenus par les sujets dans la tâche de restitution des 4 voyelles dans l'ordre.

Séquence	d	e_1	e_2	r	score (%)
/i æ ɔ u/	0,36	5,69	6,69	17,17	86
/i æ u ɔ/	4,04	0,00	1,00	0,12	45
/i ɔ æ u/	0,36	6,69	5,69	17,17	84
/i ɔ u æ/	4,74	0,00	1,00	0,11	57
/i u æ ɔ/	4,03	6,69	5,69	1,54	60
/i u ɔ æ/	4,73	5,69	6,69	1,31	70

Il est enfin possible de calculer le rapport de ces distances pour évaluer s'il est plus facile de grouper les 4 voyelles dans un seul flux ou de les séparer en deux flux :

$$r = \frac{e_1 + e_2}{2d}$$

Si $r > 1$, alors la distance est plus grande dans le cas ségrégé (2 flux) que dans le cas fusionné (1 flux), et la séquence devrait donc plutôt être perçue comme 1 flux et les scores devraient être plus élevés. Au contraire, si $r < 1$, les scores devraient être plus faibles. Ces valeurs calculées pour les voyelles utilisées par Dorman *et al.* (1975) sont présentées dans la table 1.1. Bien que cette distance ne tienne pas compte de la différence de tempo entre les percepts 1 flux et 2 flux, les plus petits scores sont obtenus pour les plus petites valeurs de r. C'est-à-dire que, dans cette expérience, les sujets tendent bien à grouper les voyelles de façon à réduire la distance formantique sur le premier formant.

Dans une partie des travaux expérimentaux présentés dans cette thèse, nous avons cherché à maîtriser cette distance perceptive entre les structures formantiques. De Boer (2000) propose un modèle de calcul de la distance perceptive entre deux voyelles calculée à partir des 4 premiers formants. Le calcul de cette distance nécessite de calculer préalablement le *second formant*

effectif F_2', comme défini par Mantakas, Schwartz, et Escudier (1986) :

$$F_2' = \begin{cases} F_2 & \text{si } F_3 - F_2 > c \\ \frac{(2-w_1)F_2 + w_1 F_3}{2} & \text{si } F_3 - F_4 \leq c \text{ et } F_4 - F_2 > c \\ \frac{w_2 F_2 + (2-w_3)F_3}{2} - 1 & \text{si } F_4 - F_2 \leq c \text{ et } F_3 - F_2 < F_4 - F_3 \\ \frac{(2+w_2)F_3 - w_2 F_4}{2} - 1 & \text{si } F_4 - F_2 \leq c \text{ et } F_3 - F_2 \geq F_4 - F_3 \end{cases}$$

avec :

$$w_1 = \frac{c - (F_3 - F_2)}{c} \quad , \quad w_2 = \frac{(F_4 - F_3) - (F_3 - F_2)}{F_4 - F_2} \quad \text{, et} \quad c = 3{,}5 \text{ Bark}$$

Dans ces formules, toutes les fréquences sont exprimées en Barks. Le F_2' correspond à la moyenne pondérée des différents formants s'ils sont espacés d'une distance inférieure à la distance critique c. Finalement, la distance perceptive entre une voyelle a et une voyelle b est calculée comme :

$$D_{ab} = \sqrt{(F_{a1} - F_{b1})^2 + \lambda (F'_{a2} + F'_{b2})^2}$$

avec λ =0,3 (Schwartz, Boe, Vallee, et Abry, 1997). C'est cette distance que nous avons utilisée pour tenter de contrôler le streaming sur la base de la distance formantique. Des détails supplémentaires sont donnés au chapitre 2.

Streaming sur la base de la hauteur fondamentale

Si des voyelles présentant de grandes différences de formants se suivent, elles auront tendance à être placées dans des flux auditifs distincts. Cependant, si le premier et le second formant d'une même voyelle sont présentés dans des oreilles différentes, ils seront fusionnés pour reformer la voyelle complète, à condition qu'ils aient le même F_0 (Broadbent et Ladefoged, 1957). La hauteur fondamentale, comme la structure formantique, est donc un facteur de streaming.

Darwin et Bethell-Fox (1977) ont étudié la possible interaction entre l'effet de la hauteur et l'effet de la structure formantique sur la formation de flux auditifs dans des séquences de consonnes-voyelles. La structure formantique évoluait de façon continue tandis que le F_0 pouvait être constant ou alterner brutalement entre deux valeurs (101 et 178 Hz). Ces auteurs ont observé que ces discontinuités de hauteur conduisaient à la formation de flux audi-

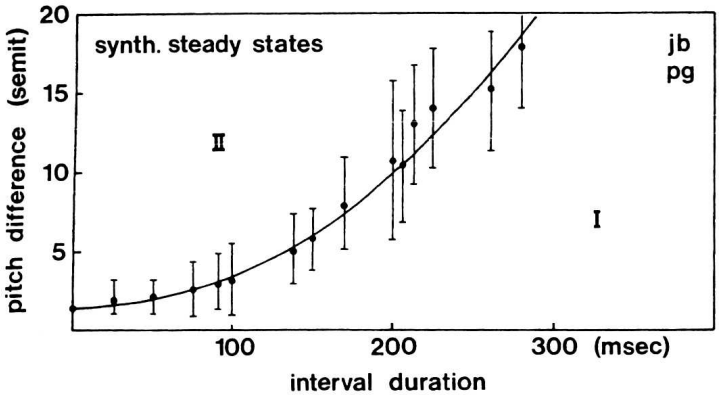

FIG. 1.9 – Différence de F_0 au seuil entre la perception d'un seul flux et de 2 flux en fonction du tempo, pour des séquences de 9 voyelles. Le tempo était manipulé en modifiant la durée de silence entre les voyelles de 100 ms (*interval duration*). D'après Nooteboom *et al.* (1978).

tifs, concluant que la continuité des formants et de la hauteur contribuaient largement à l'importante consistance du flux de parole provenant d'un locuteur donné.

Nooteboom, Brokx, et de Rooij (1978) ont étudié de façon systématique l'effet d'une différence de F_0 sur la ségrégation perceptive de séquences de voyelles. Ces auteurs ont utilisé de courtes séquences de 9 voyelles synthétiques (100 ms) de la forme : /a u i a u i a u i/. Le F_0 de ces voyelles alternait entre 100 Hz et une valeur comprise entre 100 et 280 Hz. Les sujets avaient simplement pour tâche de rapporter s'ils entendaient un flux ou deux. Les résultats moyens pour deux sujets sont reproduits figure 1.9. La courbe de seuil tracée entre la perception 1 flux (I) et 2 flux (II) est similaire à celles obtenues par van Noorden (1975) pour le même type de tâche. Elle se situe entre le seuil de cohérence et le seuil de fission. Cependant, cette étude présente d'importants défauts. Tout d'abord, les séquences n'étaient constituées que de 9 voyelles et avaient donc une durée qui n'excédait pas 4 s. Ce délai est inférieur au délai nécessaire à la construction des flux (Bregman, 1978). Il est donc possible que l'observation ait été biaisée et qu'elle ait plus consisté en de la discrimination de hauteur qu'en un jugement du nombre de flux auditifs. Ensuite, la tâche utilisée était une tâche subjective dans laquelle ni l'attention ni l'effort du sujet n'étaient orientés vers la fusion, ou vers la ségrégation. Le

seuil mesuré, s'il s'agit bien d'un seuil de streaming, ne correspond donc ni au seuil de cohérence ni au seuil de fission. Il est alors difficile d'estimer ce que représente exactement ce seuil, puisqu'il était laissé la liberté au sujet de choisir s'il cherchait à percevoir plutôt un flux, ou plutôt deux flux. Enfin, seulement deux sujets ont participé à cette expérience. Cette étude constitue donc un excellent rapport préliminaire, mais ne constitue pas une base solide pour l'interprétation des résultats de streaming en terme de séparation de voix concurrentes.

Streaming sur la base de la longueur du tractus vocal

La longueur du tractus vocal caractérise la taille d'un locuteur. Ainsi, un auditeur parvient à distinguer une voix de femme d'une voix d'enfant alors que les hauteurs fondamentales peuvent être identiques. Il est possible d'observer l'effet de la modification de la longueur du tractus vocal sur l'enveloppe spectrale d'une voyelle. Pour un tractus vocal court, l'enveloppe spectrale sera étirée en fréquence, les formants étant déplacés proportionnellement vers les hautes fréquences. Au contraire, pour un tractus vocal plus long, l'enveloppe spectrale sera compressée selon l'axe des fréquences.

Tsuzaki, Takeshima, Irino, et Patterson (2007) ont observé qu'une différence de longueur du tractus vocal pouvait être utilisée pour séparer des séquences de voyelles en flux auditifs. Ils ont réalisé des séquences de 6 voyelles similaires à celles de Dorman *et al.* (1975), dans lesquelles la longueur du tractus vocal alternait entre deux valeurs. Ils ont observé que les scores d'identification des 6 voyelles dans l'ordre étaient réduits de plus de 20% quand la variation de longueur du tractus vocal passait de 0 à 1/2 octave (c'est-à-dire que l'enveloppe spectrale était multipliée par $\sqrt{2}$). Il est donc probable que l'information de taille du locuteur soit utilisée dans les situations de voix concurrentes pour discerner les différents locuteurs.

1.2.4 Théorie des canaux et modèles

Finalement, il apparaît qu'un grand nombre de paramètres acoustiques peuvent induire un phénomène de streaming. Une alternance de deux sons qui diffèrent selon un caractère acoustique pourra être combinée en flux auditifs dès lors que cette différence acoustique est suffisamment perceptible

(Moore et Gockel, 2002). La théorie des canaux périphériques (*Peripheral Channeling Theory*) propose une explication à certains des phénomènes de streaming rapportés dans la littérature, en s'appuyant sur l'organisation des voies auditives (Hartmann et Johnson, 1991). Cette théorie suppose que deux sons produisent deux flux auditifs s'ils excitent des populations de neurones différentes. Plus précisément, les auteurs de cette théorie associent cette séparation à la décomposition tonotopique effectuée dans la cochlée, et dont la structure semble conservée jusqu'au cortex auditif primaire. Il est donc considéré que si deux sons excitent des zones différentes de la cochlée, ils seront perçus dans deux flux différents. Au contraire, s'ils excitent les mêmes zones de la cochlée, ils seront intégrés dans un même flux. C'est donc la forme spectrale qui est prédominante dans cette théorie, puisqu'elle détermine la répartition de l'énergie dans la cochlée, et donc la probabilité de décharge des neurones du nerf auditif. Plus exactement, ce sont les diagrammes d'excitation (*excitation pattern*) de ces neurones qui déterminent si deux sons sont fusionnés ou ségrégés. Les spectres des sons et la résolution fréquentielle de l'oreille détermineraient donc l'état de ségrégation.

Cependant, comme il sera présenté dans la section 1.3, la résolution fréquentielle de l'oreille ne semble pas avoir autant d'importance que cette théorie lui confère. Ainsi, bien qu'elle ait inspiré certains modèles de streaming qui ont reproduit certaines données de la littérature, il existe d'autres modèles qui ne reposent pas exclusivement sur la représentation fréquentielle des sons dans le système auditif.

Modèles de streaming

La théorie des canaux a été implémentée au travers de deux modèles présentant de nombreux points communs (Beauvois et Meddis, 1996; McCabe et Denham, 1997). Beauvois et Meddis (1996) ont conçu un modèle destiné à reproduire l'organisation en flux de séquences de sons purs en prenant en compte la différence de fréquences entre les sons et le tempo de la séquence. L'architecture du modèle est présentée sur la figure 1.10. Le modèle comprend un premier étage qui représente le traitement effectué par la cochlée. Le signal subit une décomposition fréquentielle à l'aide d'un banc de filtres. Puis, à la sortie de chaque canal i, l'activité dans le nerf auditif est estimée, donnant une probabilité de décharge des neurones H_{it} en fonction du temps t et du canal

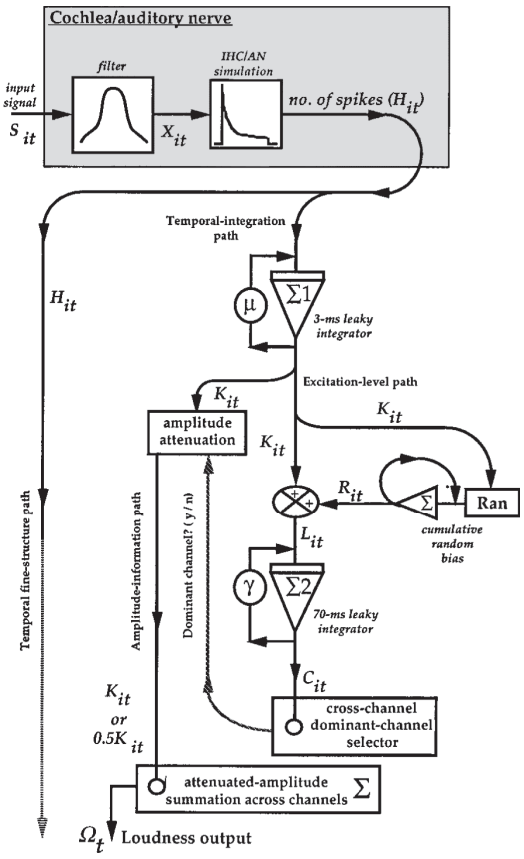

FIG. 1.10 – Architecture du modèle de streaming de Beauvois et Meddis (1996).

i. La sortie de chaque canal est alors divisée en deux parties : l'une destinée à l'intégration temporelle, l'autre destinée à préserver la structure fine du signal pour un usage ultérieur. L'activité H_{it} du canal i est ensuite transmise à un ensemble d'accumulateurs ayant différentes constantes de temps, pour finalement donner lieu au signal C_{it}, qui représente le niveau d'excitation de chaque canal. Un certain temps est nécessaire pour atteindre le niveau maximal d'excitation, et après l'arrêt de l'excitation acoustique du canal, il faut un certain temps pour qu'il retrouve son niveau au repos.

Si l'on soumet une séquence de deux sons purs ABABAB à ce modèle, et que les sons A et B sont dans le même canal, alors le niveau d'excitation du canal est déjà élevé quand le son B débute. Le niveau d'excitation de ce canal est donc relativement stable dans le temps et reste à un niveau élevé. Au contraire, lorsque les sons A et B sont dans des canaux différents, le canal du son B n'est pas pré-excité par le son A. Comme un délai est nécessaire

pour atteindre un niveau d'excitation important, le niveau d'excitation C_{it} de chacun de ces canaux sera instable et restera relativement faible. Il y a alors ségrégation. Ce modèle a reproduit avec succès les données collectées par Miller et Heise (1950), van Noorden (1975) et Bregman (1978). Il simule correctement l'effet du tempo grâce aux différentes constantes de temps des accumulateurs, et l'effet de ΔF grâce à la répartition en canaux. Il simule aussi le fait qu'un certain temps est nécessaire à la construction et à la déconstruction des flux.

Le modèle établi par McCabe et Denham (1997) présente une architecture différente. Dans ce modèle, deux modules de traitement sont connectés : un pour l'arrière plan et l'autre pour le premier plan de la scène auditive. Les sorties de ces modules sont ré-injectées en entrée des deux modules. Chaque module renforce sa propre analyse en s'appuyant sur le résultat précédent (simulant la cohérence temporelle). Chaque module reçoit aussi la sortie de l'autre module utilisée comme signal inhibiteur. Les processus s'excluent donc mutuellement. Ce modèle permet ainsi d'obtenir une image du schéma d'excitation de chacun des flux créés. Néanmoins, la décision d'exclure un élément d'un flux repose aussi sur sa proximité fréquentielle. De plus, comme dans le modèle de Beauvois et Meddis (1996), la structure fine du signal est ignorée.

Ces deux modèles sont relativement efficaces pour simuler la ségrégation engendrée par des séquences de sons purs. En revanche, ils ne permettent pas de traiter efficacement d'autres types de sons, comme par exemple les sons complexes harmoniques. D'autres modèles de ségrégation reposent moins fortement sur la théorie des canaux. Grossberg, Govindarajan, Wyse, et Cohen (2004) et Elhilali et Shamma (2007) ont ainsi développé des modèles plus généraux de ségrégation incluant ségrégation simultanée et ségrégation séquentielle.

Grossberg *et al.* (2004) font reposer leur modèle de ségrégation ART-STREAM sur le modèle de perception de la hauteur SPINET (Cohen, Grossberg, et Wyse, 1995). Ce modèle repose sur une analyse spectrale (crible harmonique) et une analyse de la structure fine du signal. Il prend donc en compte les phénomènes de phase décrits précédemment. En revanche, il ne permet pas d'extraire la hauteur de bruits modulés en amplitude. En sortie du modèle SPINET, le signal est donc représenté par son spectre de fréquence et son spectre de hauteur fondamentale. La cohérence temporelle

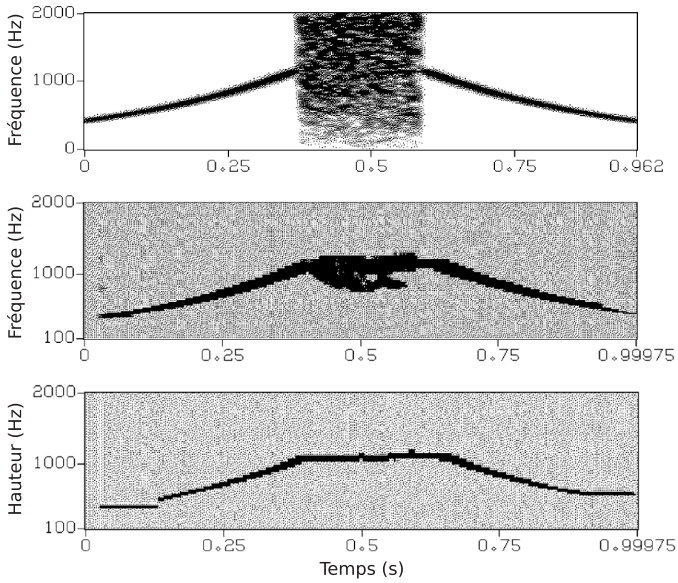

FIG. 1.11 – *En haut* : spectrogramme du son initial. Ce son produit une impression de hauteur continue malgré la présence de bruit autour de 0,5 s. *Au milieu* : la répartition énergétique dans la couche fréquentielle du modèle de séparation ARTSTREAM. La continuité est relativement bien simulée. *En bas* : la répartition énergétique dans la couche représentant la hauteur dans ARTSTREAM. La continuité de hauteur est parfaitement simulée. Reproduit de Grossberg *et al.* (2004).

est modélisée dans ARTSTREAM à partir de cette représentation, en utilisant un bouclage entre la représentation fréquentielle et la représentation de la hauteur. Si une hauteur fondamentale dominante résulte de l'analyse à un instant donné, certaines zones de la représentation fréquentielle seront inhibées. Ces zones prennent la forme d'un crible qui supprime pendant un certain temps le contenu spectral en dehors des harmoniques de la hauteur fondamentale dominante. Ceci permet de simuler les effets de continuité de perception de la hauteur, comme illustré figure 1.11. Ce modèle donne une importante rémanence à une hauteur perçue. Il permet ainsi de reproduire les effets de continuité de hauteur qui permettent notamment d'associer les éléments d'un flux perceptif dans les expériences de van Noorden (1975).

Le modèle décrit par Elhilali et Shamma (2007) prend un "spectrogramme auditif" (*auditory spectrogram*, Chi, Ru, et Shamma, 2005) en entrée. Un ex-

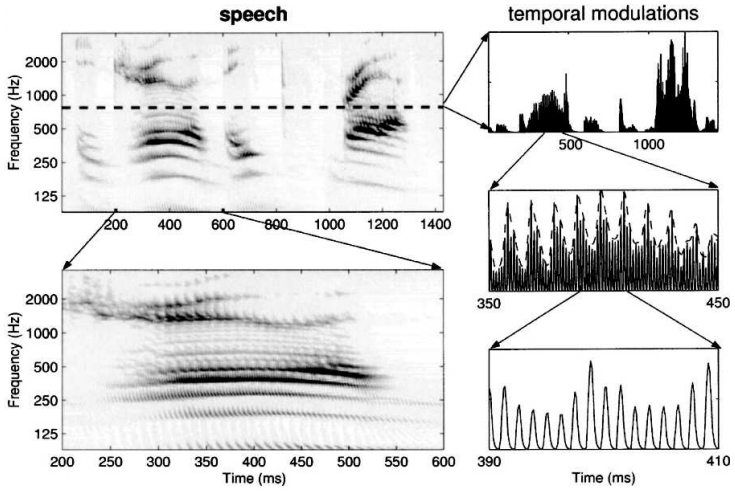

FIG. 1.12 – Spectrogramme auditif de la phrase "He drew a deep breath". Le trait tireté repère le canal à 750 Hz dont les modulations temporelles sont représentées dans la colonne de droite, à différentes échelles. À l'échelle la plus large (en haut), cette modulation permet de distinguer les différents segments syllabiques. À l'échelle intermédiaire (au milieu), la modulation reflète la fréquence fondamentale (environ 100 Hz). Enfin, à l'échelle la plus fine (en bas), la structure fine du canal est représentée. Reproduit de Chi *et al.* (2005).

emple de cette représentation spectro-temporelle pour un signal de parole est représenté figure 1.12. À ce stade, le modèle est relativement dépendant de la résolution fréquentielle du système auditif. Cependant, les aspects temporels du signal sont conservés jusqu'à la structure fine. Il n'y a donc pas réellement de prédominance de la représentation spectrale à ce stade du modèle. Ce spectrogramme auditif est ensuite analysé grâce à un ensemble de champs réceptifs spectro-temporels (*spectro-temporal receptive field*, STRF). Ces champs modélisent la réponse impulsionnelle temps-fréquence d'un neurone du cortex auditif primaire. L'analyse réalisée à l'aide de ces STRF est multi-résolution à la fois en temps et en fréquence. Le processus d'analyse donne donc lieu à une représentation à 4 dimensions : temps, fréquence, échelle temporelle, échelle spectrale. Le système étant multi-résolution, une perte de résolution en entrée n'affectera qu'une partie réduite de la représentation. Elhilali et Shamma (2007) ont postulé que la recherche de corrélations temporelles dans cet espace multidimensionnel représentait la cohérence temporelle de certaines composantes d'un signal sonore et permettait donc

de modéliser le phénomène de streaming. Une matrice de corrélations est obtenue et sert de masque pour pondérer le spectrogramme du son original. Ce modèle produit finalement deux spectrogrammes : celui du premier flux, et celui du second flux. L'analyse multi-résolution permet d'envisager un traitement efficace de tout type de sons. Par ailleurs, ce modèle multidimensionnel peut être aisément complété d'autres dimensions comme la hauteur fondamentale, la localisation, etc...

On l'a vu, dans les deux premiers modèles, la résolution fréquentielle joue un rôle déterminant pour le streaming, en accord avec la théorie des canaux. En revanche, dans ces deux derniers modèles, la résolution fréquentielle du système auditif, ou *sélectivité fréquentielle*, joue un rôle important mais non déterminant dans le streaming. Il est réputé que les malentendants souffrent de grandes difficultés de compréhension de la parole dans le bruit. La section suivante est consacrée à la perte auditive, et plus particulièrement à la perte de sélectivité fréquentielle que les prothèses ne peuvent réhabiliter.

1.3 Perte auditive et sélectivité fréquentielle

La perte partielle ou totale de l'audition est caractérisée par des dégradations de l'appareil auditif qui engendrent une dégradation de la plupart des fonctions auditives mesurables par de simples tâches psychoacoustiques (section 1.3.1). Ces dégradations entraînent au quotidien des difficultés extraordinaires à percevoir la parole dans le bruit (section 1.3.2). Si la ségrégation simultanée et la ségrégation séquentielle sont bien impliquées dans la perception de voix concurrentes, il est donc probable qu'elles soient altérées par une perte auditive (sections 1.3.3 et 1.3.4).

1.3.1 Altération des indices perceptifs

Les pertes auditives auxquelles il est fait référence ici sont des pertes neurosensorielles. C'est-à-dire qu'elles sont caractérisées par une disparition partielle ou totale des cellules ciliées externes, et éventuellement internes. Ce type de perte peut être causé par des syndromes génétiques, une surexposition au bruit ou simplement au vieillissement du système auditif (presbyacousie). Ce type de perte est généralement caractérisé par 3 effets clairement identi-

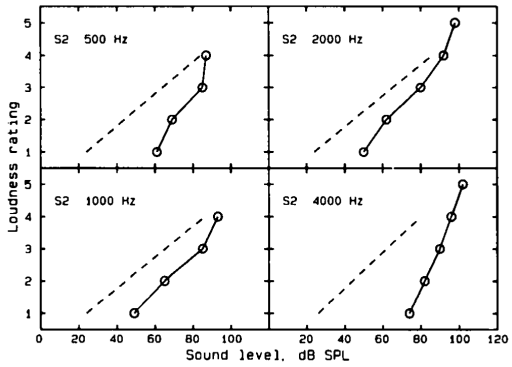

FIG. 1.13 – Jugement de la sonie d'un bruit filtré sur une demie octave en fonction de son intensité pour un sujet malentendant (cercles). Les traits tiretés représentent les résultats obtenus pour des sujets normo-entendants. Reproduit de Moore *et al.* (1992).

fiés : un effet de seuil, une perte de compression et une perte de sélectivité fréquentielle (Moore, 1998).

Seuil et compression

L'effet de seuil est mesuré par l'audiogramme tonal. La perte de compression modifie l'intensité perçue pour une intensité physique donnée. Comme le montrent les observations de Moore, Johnson, Clark, et Pluvinage (1992), reproduites figure 1.13, tant que l'intensité est inférieure à 90 dB SPL, les malentendants ont une impression de sonie plus faible que les normo-entendants. En revanche, au-delà de 90 dB SPL, la sonie chez les malentendants souffrant de dommages cochléaires est typiquement identique à celle des normo-entendants. Ceci peut s'expliquer par le fait que l'amplification du signal dû à l'activation des cellules ciliées externes n'est opérationnelle que pour des signaux d'intensité inférieure à 90 dB SPL. La disparition de ces cellules ciliées n'a donc que peu d'influence sur la perception de ces niveaux sonores. Il est important de noter que dans le cas d'une perte de transmission, l'écart entre la sonie chez le malentendant et le normo-entendant reste constant quelque soit l'intensité du stimulus.

Sélectivité fréquentielle

La perte de sélectivité fréquentielle correspond à une réduction de l'acuité spectrale. Ceci peut se représenter par la forme ou la largeur caractéristique des filtres auditifs. La figure 1.14 illustre le phénomène d'élargissement des filtres auditifs chez des sujets souffrant d'une perte auditive unilatérale (à gauche). Les filtres de l'oreille endommagée sont plus larges que ceux de l'oreille saine, conférant une bien moins bonne résolution à la représentation spectrale des sons dans le système auditif. La partie droite de la figure 1.14 représente la largeur de filtres auditifs (ERB) pour différentes fréquences, en

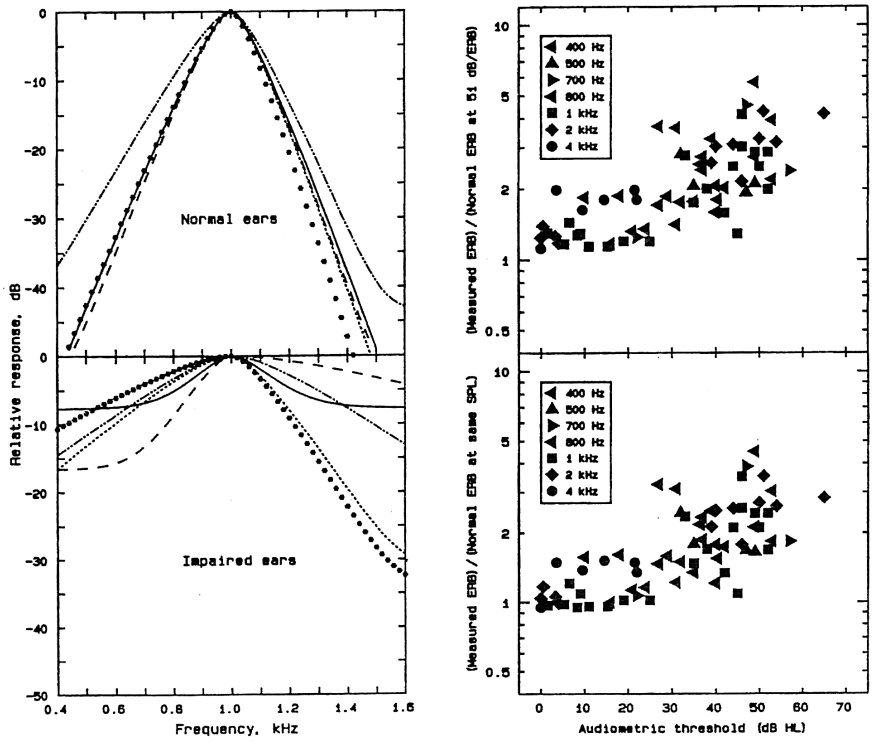

FIG. 1.14 – *À gauche* : forme des filtres auditifs à 1 kHz de l'oreille normale (en haut) et de l'oreille endommagée (en bas) de six sujets ayant une perte auditive unilatérale. Reproduit de Glasberg et Moore (1986). *À droite* : largeur du filtre auditif exprimée par l'ERB en fonction du seuil audiométrique. En haut les mesures de largeur de filtre sont comparées à l'ERB_N. En bas, elles sont comparées à des mesures chez de jeunes normo-entendants au même niveau sonore. Ces données ont été compilées à partir de plusieurs études. Reproduit de Moore (1998).

fonction du seuil audiométrique mesuré à cette même fréquence. La largeur des filtres est estimée sous la forme de l'ERB et représentée comme le rapport avec l'ERB_N (en haut) ou avec l'ERB mesurée chez de jeunes normo-entendants au même niveau sonore (en bas). La largeur du filtre auditif dépend en effet du niveau du signal employé pour effectuer la mesure. L'ERB_N est calculée pour un niveau de 51 dB/ERB (environ 30 dB/Hz). Il apparaît qu'il n'existe pas de relation simple entre le seuil et la largeur des filtres auditifs et qu'il existe une importante variabilité. Cependant, pour des seuils supérieurs à 30 dB HL, les largeurs de filtres tendent à augmenter avec le seuil. La perte auditive a donc un effet difficilement prédictible sur la réso-

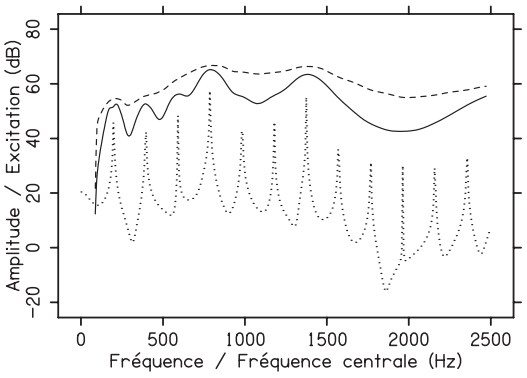

FIG. 1.15 – Spectre d'un /a/ à 196 Hz (pointillés). Simulation de la répartition du niveau d'excitation en fonction de la fréquence pour des filtres auditifs normaux (trait continu) ou élargis 3 fois (trait tireté).

lution fréquentielle. À ce jour, les prothèses auditives ne permettent que de restaurer l'effet de seuil et la perte de compression.

La sélectivité fréquentielle est importante pour la perception des sons complexes harmoniques. Si les filtres auditifs sont suffisamment étroits, les harmoniques qui composent le son complexe seront chacunes isolées dans les filtres auditifs qui ne se recouvrent pas. On parle alors d'harmoniques résolues. En revanche, si les filtres auditifs sont larges, les harmoniques pourront exciter des filtres auditifs très proches et interférer. Et on parle alors d'harmoniques non résolues. Dans ce dernier cas, la perception de la hauteur devient sensible aux relations de phases entre les harmoniques. La non résolution des harmoniques peut intervenir à haute fréquence chez les normo-entendants, ou à basse fréquence quand les filtres auditifs sont élargis, comme illustré dans la figure 1.15. Dans le cas normal, la forme de l'excitation montre des bosses correspondant aux premières harmoniques. Dans le cas élargi, seule la première harmonique produit une bosse.

Résolution temporelle

Enfin, il existe un effet, moins clair, de la perte auditive sur la résolution temporelle de l'oreille. La résolution temporelle est difficile à distinguer de la résolution fréquentielle puisque les deux sont liées. En effet, la complexité de la structure fine en sortie d'un filtre auditif dépend directement de la largeur de ce filtre (Moore et Carlyon, 2005). Par ailleurs, l'analyse des structures temporelles et fréquentielles est effectuée par des neurones qui présentent des caractéristiques spectro-temporelles (STRF). On peut néanmoins tenter de

distinguer précisément la résolution temporelle de la résolution fréquentielle. Comme on l'a vu dans le modèle de Chi *et al.* (2005), il existe au moins trois niveaux de résolution temporelle pertinents pour la perception de la parole. Un niveau très lent, à la fréquence des syllabes. Ce niveau est globalement peu affecté par une perte auditive. Puis un second niveau qui correspond à la modulation d'amplitude liée à la fréquence fondamentale. Une mesure de la perception des modulations d'amplitude est donnée par la fonction de transfert de modulation temporelle (*temporal modulation transfer function*, TMTF). Les TMTF sont généralement considérées comme préservées chez les malentendants (Moore, 1998). Enfin, le troisième niveau de résolution temporelle, le plus fin, contient la structure fine du signal. Les cellules ciliées émettent des impulsions nerveuses à la fréquence de cette structure fine. Cette fréquence est comprise dans la bande passante du filtre auditif correspondant, mais permet une estimation de la fréquence beaucoup plus précise grâce à la synchronisation des impulsions entre les différents neurones concernés par ce filtre auditif (*phase-locking*). Ceci s'illustre par le fait que les performances de discrimination de fréquences sont bien meilleures que ne peuvent le permettre les filtres auditifs seuls. Moore et Moore (2003) et Hopkins et Moore (2007) ont étudié la discrimination de sons complexes chez des sujets normo-entendants et des sujets malentendants. En utilisant différentes conditions de résolvabilité et différents types de sons complexes, ces auteurs sont arrivés à la conclusion que les normo-entendants utilisent la structure fine pour obtenir la hauteur fondamentale, même lorsque les harmoniques ne sont pas résolues. En revanche, les malentendants ne semblent pas exploiter ce type d'indice temporel, et se contentent de l'enveloppe temporelle du signal. Lorenzi *et al.* (2006) ont tenté d'observer l'importance et la dégradation de la structure fine pour la perception de la parole dans le bruit chez les malentendants. Les stimuli utilisés étaient des triplets VCV, où V était toujours /a/, et C était une consonne parmi /p, t, k, b, d, g, f, s, ʃ, v, z, j, m, n, r, l/. Ces triplets étaient décomposés en 16 bandes dans lesquelles étaient isolées la structure fine (TFS) d'une part et l'enveloppe temporelle d'autre part (E). Des sujets normo-entendants et malentendants, jeunes et âgés, devaient identifier les consonnes présentées. Les résultats sont reportés figure 1.16. Ces résultats démontrent que les normo-entendants atteignent de hauts niveaux d'identification avec l'enveloppe seule ou la structure fine seule

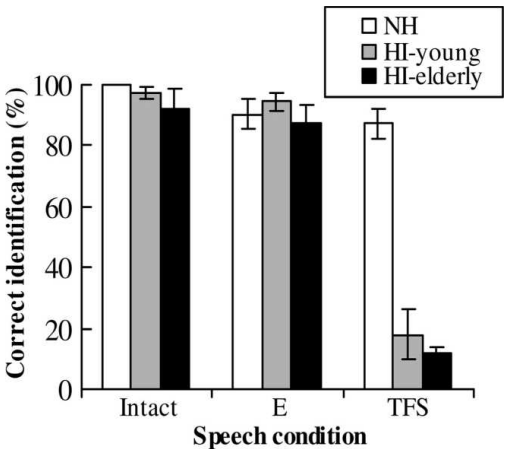

FIG. 1.16 – Pourcentage d'identification correcte des triplets VCV pour trois conditions : signal intact, enveloppe (E) et structure fine (TFS) pour chacun des groupes de sujets : normo-entendants (NH), malentendants jeunes (HI-young) et malentendants âgés (HI-elderly). Reproduit de Lorenzi *et al.* (2006).

du signal dans chaque bande de fréquence. En revanche, les malentendants, quelque soit leur âge, se sont révélés incapables de tirer profit de la structure fine pour effectuer l'identification. Cette deuxième étude semble confirmer que les malentendants auraient aussi un déficit de perception de la structure fine.

Résolutions temporelle et fréquentielle dans l'implant cochléaire

Dans le cas où la perte auditive est trop sévère, par exemple si trop de cellules ciliées externes et internes ont été endommagées, il est parfois possible de poser un implant cochléaire au patient. L'implant cochléaire stimule électriquement le nerf auditif par l'intermédiaire d'une vingtaine d'électrodes. La sélectivité fréquentielle est donc fortement dégradée puisque l'implanté ne perçoit plus qu'une vingtaine de bandes de fréquences différentes. La faible résolution fréquentielle autorisée par l'implant implique que les composantes de la plupart des sons complexes harmoniques de notre environnement sont non résolues. Seule l'enveloppe spectrale est grossièrement conservée, ainsi que l'enveloppe temporelle. La structure fine temporelle et la structure harmoniques sont supprimées. Les conséquences de l'implant en termes d'indices perceptifs et de performances auditives sont détaillées au chapitre 3.

1.3.2 Perception de voix concurrentes

La dégradation de chacun des indices perceptifs liés à la perte auditive (seuil et compression, sélectivité fréquentielle, indices temporels) engendre une réduction des performances de perception de la parole dans le bruit. Plusieurs études ont cherché à évaluer l'importance de certains indices perceptifs ou de certaines fonctions auditives en tentant de corréler ces grandeurs aux performances de perception de parole dans le bruit. Les études de Festen et Plomp (1983) et de Glasberg et Moore (1989) sont détaillées au chapitre 4. La principale conclusion de ces deux études est que la perception de la parole dans le silence est principalement pilotée par l'audibilité (et donc les seuils auditifs), tandis que la perception de la parole dans le bruit dépend essentiellement des mesures liées à la représentation fréquentielle des sons dans le système auditif. Une autre approche du problème consiste à utiliser des simulations. En ne simulant que certaines caractéristiques de la perte auditive, il est possible d'en évaluer les conséquences sur la perception de la parole dans le bruit, comme nous allons le voir dans les sections suivantes.

Seuil et compression

Moore et Glasberg (1993) ont utilisé une simulation pour observer l'effet de l'élévation des seuils auditifs et de la perte de compression sur la perception de paroles concurrentes. Les sujets étaient soumis à une tâche de perception de phrases en présence d'une voix concurrente. Quatre conditions ont été réalisées :

- R1 : signal original non transformé.
- R2 : perte plate moyenne avec un seuil à environ 50 dB HL.
- R3 : perte plate sévère avec un seuil à environ 67 dB HL.
- RX : perte croissante avec la fréquence allant de 33 dB HL pour 879 Hz à 67 dB HL pour 5837 Hz.

Une variante des trois dernières conditions a aussi été testée, notées R2+, R3+ et RX+. Dans cette variante, les signaux traités étaient ensuite amplifiés linéairement dans chaque canal fréquentiel en suivant les recommandations destinées au réglage des prothèses auditives (amplification NAL). Les résultats, figure 1.17, montrent que les effets de seuil et de recrutement de sonie réduisaient d'environ 20% les performances d'identification des sujets

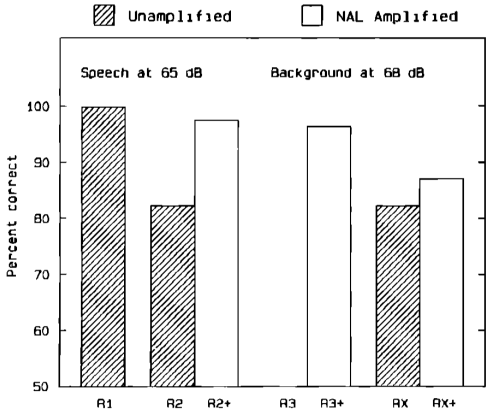

FIG. 1.17 – Comparaison des scores d'identification de la phrase cible à un rapport cible-masque de –3 dB, pour les différentes conditions. Aucun score n'est donné pour la condition R3 car la cible était inaudible pour cette condition. Reproduit de Moore et Glasberg (1993).

dans le cas d'une perte plate moyenne (R2) et dans le cas d'une perte plus importante en haute fréquence (RX). Dans les trois conditions dégradées, l'amplification (R2+, R3+ et RX+) permettait de restaurer l'audibilité et améliorait sensiblement les performances d'identification sans pour autant permettre d'atteindre les mêmes performances que dans la condition intacte (R1). D'autres études ont montré que cette même amplification linéaire était efficace pour restaurer la perception de la parole dans un masque stationnaire (par exemple Moore, Glasberg, et Vickers, 1995, cité par Moore, 1998).

Sélectivité fréquentielle

L'amplification réalisée dans les prothèses auditives permet donc de restaurer l'audibilité et l'intelligibilité si le signal de parole est masqué par un bruit stationnaire. Cependant, les malentendants, même équipés de prothèses auditives restaurant l'audibilité, souffrent d'un déficit d'intelligibilité de la parole dans le bruit qui ne peut donc venir des effets de seuil et de compression. Baer et Moore (1993) ont simulé une perte de sélectivité fréquentielle en effectuant un lissage spectral selon la procédure schématisée figure 1.18. Cet algorithme est décrit plus en détails au chapitre 2. Des phrases noyées dans du bruit, ayant le même spectre à long terme que les phrases, ont été traitées avec l'algorithme de lissage spectral, de façon à simuler un élargissement des filtres auditifs. Les résultats obtenus sont présentés figure 1.19. Ils montrent que le lissage spectral rend rapidement très difficile l'identification de la parole dans le bruit quand le rapport signal sur bruit devient défavorable. En

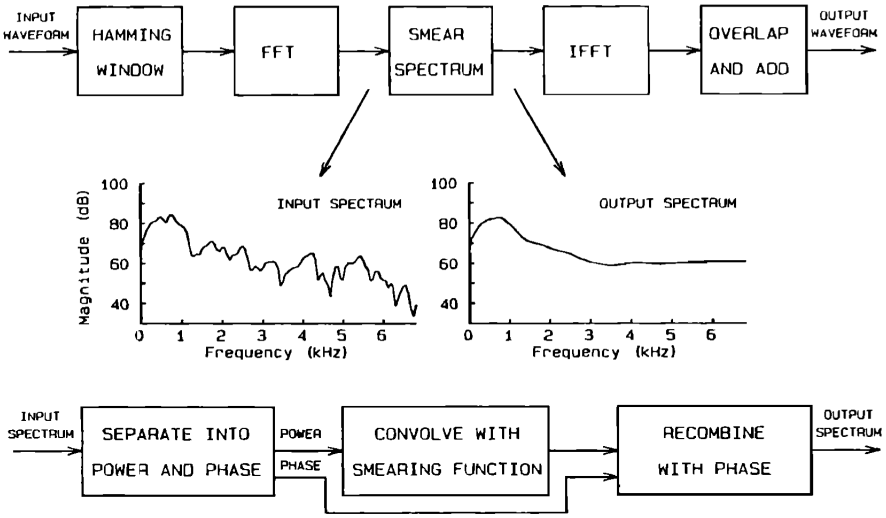

FIG. 1.18 – Représentation schématique du traitement utilisé pour réaliser le lissage spectral. La partie supérieure montre la séquence globale de traitements ainsi qu'un exemple de lissage spectral sur le spectre à court terme d'une fenêtre temporelle. La partie inférieure montre le détail du bloc "smear spectrum". Reproduit de Baer et Moore (1993).

particulier pour un SNR de –3 dB, la simulation de filtre 6 fois plus larges que la normale conduit à une chute de plus de 50% des scores d'identification.

Une autre méthode consiste à évaluer la capacité des sujets à profiter de trous spectraux dans un bruit stationnaire. Peters *et al.* (1998) ont réalisé des bruits spectralement façonnés par la parole et y ont ménagé des trous de différentes largeurs : 2, 3 et 4 ERB_N. L'insertion de ces trous réduisait les SRT des normo-entendants de 8.7, 12.3 et 14.9 dB, alors que cela ne les réduisait que de 2.3, 2.8 et 3.3 dB chez les malentendants. Même après amplification NAL, le bénéfice que les malentendants tiraient de ces trous spectraux ne dépassait pas 8.8 dB (dans la condition 4 ERB_N).

De façon encore plus évidente que pour les autres malentendants, les implantés cochléaires ont de grandes difficultés à percevoir la parole dans le bruit. Stickney, Zeng, Litovsky, et Assmann (2004) ont comparé les performances de sujets normo-entendants avec et sans simulation d'implant cochléaire, aux performances d'implantés réels pour différents types de masques. Ils ont observé que sans simulation d'implant, les normo-entendants dépassaient 80% de bonnes réponses dès que le SNR était supérieur à 5 dB.

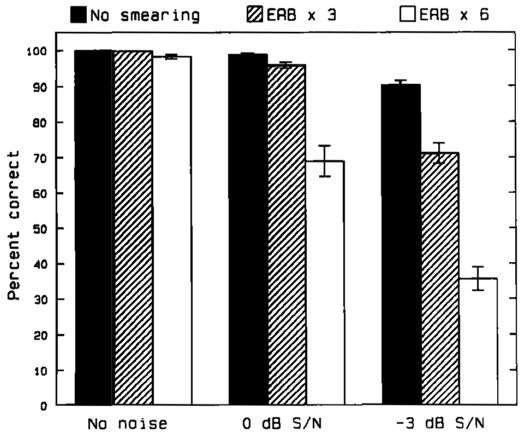

FIG. 1.19 – Scores d'identification des phrases sans lissage spectrale, et avec une simulation de filtres auditifs élargis 3 fois et 6 fois, en fonction du rapport signal sur bruit. Reproduit de Baer et Moore (1993).

Avec une simulation d'implant à 8 canaux, les mêmes normo-entendants atteignaient 80% de bonnes réponses pour un SNR supérieur à 15 dB. Enfin, les implantés, équipés d'implants à 22 ou 24 électrodes, atteignaient difficilement 50%, même pour un SNR de 20 dB. En particulier, dans le cas où la cible était une voix d'homme et le masque une voix de femme, les normo-entendants ont obtenu des scores constants à environ 90% de bonnes réponses quelque soit le SNR, tandis qu'avec la simulation d'implant ils n'atteignaient pas 80%, même au SNR le plus favorable. Les implantés ne profitaient que très faiblement de la différence entre ces deux locuteurs, et n'atteignaient pas 50% d'identification correcte. Ces résultats montrent que la médiocre résolution spectrale et la dégradation de la structure fine dues à l'implant réduisent fortement la capacité de perception de la parole dans le bruit. En particulier, ces résultats soulignent les difficultés des implantés à bénéficier d'une différence de F_0. Néanmoins, la pauvreté de la représentation spectrale fournie par l'implant au système auditif ne justifie pas à elle seule les faibles scores des implantés puisque les normo-entendants ont obtenus des scores légèrement supérieurs avec la simulation d'implant, alors que la résolution fréquentielle dans cette simulation était moins bonne que dans les implants réels utilisés dans cette étude.

Indices temporels

En plus de l'effet de trous spectraux, Peters *et al.* (1998) ont aussi observé ce que devenaient les performances de normo- et malentendants lorsque le

bruit masquant était modulé en amplitude. La modulation d'amplitude en reprenant l'enveloppe temporelle (l'amplitude sur des fenêtres de 10 ms) de phrases permettait d'améliorer les SRT de 6,2 dB chez les normo-entendants, et de 2,7 dB chez les malentendants (4,2 dB avec l'amplification NAL). Ce résultat indique donc que les malentendants profitent moins de la modulation temporelle du signal, bien que leurs TMTF soient similaires à ceux des normo-entendants.

Lorenzi *et al.* (2006) ont mis en évidence que les malentendants présentaient des difficultés pour utiliser la structure fine temporelle de la parole. Pour vérifier si ce déficit pouvait être relié à la capacité des malentendants à percevoir de la parole dans un bruit modulé en amplitude, ces auteurs ont mesuré l'effet de l'adjonction d'un bruit de fond modulé sur les scores d'identification de syllabes (*masking release*). Ils ont observé que les scores d'identification des syllabes à partir de leur structure fine seule (TFS) étaient bien corrélés à l'effet de l'adjonction d'un bruit modulé sur les scores d'identification des syllabes. Ceci suggère que la dégradation de la perception de la structure fine serait responsable des difficultés des malentendants à profiter des trous temporels présents dans un masque.

1.3.3 Ségrégation simultanée

Summers et Leek (1998) ont comparé les performances de normo-entendants et de malentendants dans différentes tâches : une tâche de discrimination de F_0 portant sur des voyelles, une tâche de ségrégation simultanée avec des doubles voyelles, et une tâche de perception de phrases concurrentes différant par leur F_0. Ils ont trouvé que les seuils de discrimination de F_0 avaient tendance à être plus grands pour les malentendants, bien que l'effet de groupe n'ait pas été significatif. Dans la tâche de doubles voyelles, les malentendants se sont révélés incapables de tirer profit de la différence de F_0 entre les voyelles. Les scores des normo-entendants augmentaient de 20% lorsque la différence de F_0 passait de 0 à 4 demitons (l'essentiel du bénéfice ayant lieu entre 0 et 1 demiton), tandis que les scores des malentendants n'augmentaient que de 8%. Enfin, dans la tâche de perception de phrases concurrentes, les normo-entendants ont aussi montré en moyenne un profit légèrement plus grand de la différence de F_0 que les malentendants, mais cette différence ne s'est pas révélée significative. Les sujets normo-entendants

semblaient donc profiter plus des différences de F_0 dans les trois tâches que les malentendants. Cependant, après avoir observé les résultats individuels, Summers et Leek (1998) ont conclu que les capacités des sujets à tirer profit de la différence de F_0 dans les deux tâches impliquant de la parole n'étaient que faiblement reliées. Comme indiqué précédemment (section 1.1.3), la ségrégation simultanée ne permet que de prédire faiblement les performances de séparations de voix concurrentes.

Qin et Oxenham (2005) ont observé que des sujets normo-entendants avaient d'importantes difficultés à profiter d'une différence de F_0 dans une tâche de doubles voyelles, lorsque les stimuli étaient soumis à une simulation d'implant cochléaire. Les implantés semblent ne pas pouvoir tirer profit de différences de périodicité de l'enveloppe temporelle (comme indice de hauteur) pour séparer des sons (Carlyon, Long, Deeks, et McKay, 2007).

1.3.4 Ségrégation séquentielle

Sons purs

Grose et Hall (1996) ont observé, à travers plusieurs tâches, que les malentendants souffrant d'une perte neurosensorielle avaient besoin d'une différence de fréquences plus importante que les normo-entendants pour séparer des séquences de sons purs. Rose et Moore (1997) ont cherché à valider le modèle de Beauvois et Meddis (1996), et par là de vérifier la théorie des canaux (Hartmann et Johnson, 1991). Selon cette théorie, la sélectivité fréquentielle détermine à quel point deux sons purs provoqueront des formes d'excitation différentes. Le seuil de fission exprimé comme une différence de fréquences sur une échelle d'ERB devrait donc être constant. Pour tester cette hypothèse, Rose et Moore (1997) ont mesuré le seuil de fission chez des sujets normo-entendants et malentendants. Ils ont utilisé le numéro d'ERB_N (Glasberg et Moore, 1990) pour exprimer la fréquence des sons purs :

$$E = 21,4 \ \log_{10}(4,37.10^{-3} \ F + 1)$$

et ont ainsi pu exprimer le seuil de fission comme une différence ΔE. Le seuil de fission était mesuré en faisant varier la fréquence du son B dans un motif ABA–ABA–. Dans un cas, les sons B étaient plus graves que les sons A, et leur fréquence augmentait (condition I) ; dans l'autre, les sons B étaient plus

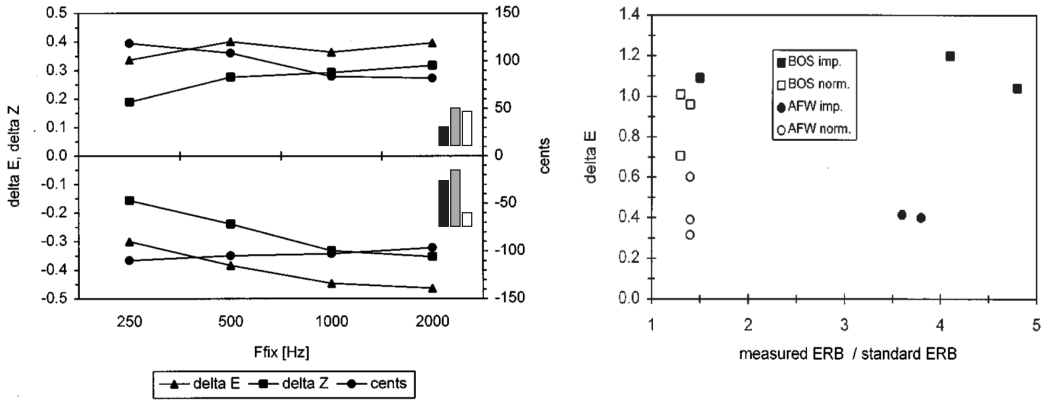

FIG. 1.20 – *À gauche* : Seuils de fission moyens en fonction de la fréquence du son A. Le cadre supérieur correspond à la condition I, le cadre inférieur à la condition II. Les triangles représentent le seuil de fission sous la forme ΔE, les carrés le représentent sous la forme d'une différence en Barks ΔZ, et les ronds le représentent en centièmes de demitons. Les histogrammes montrent graphiquement les différences (en valeur absolue) entre le seuil de fission mesuré à 250 Hz et celui mesuré à 2000 Hz pour les trois représentations du seuil. *À droite* : seuil de fission ΔE en fonction de l'ERB mesurée pour deux sujets ayant des pertes unilatérales. Les symboles pleins représentent les oreilles endommagées, et les symboles vides représentent les oreilles saines. D'après Rose et Moore (1997).

aigus que les sons A, et leur fréquence diminuait (condition II). Les résultats obtenus pour les normo-entendants sont présentés figure 1.20, à gauche. Ces résultats montrent que le seuil de fission exprimé en demitons est le plus constant en moyenne sur les deux conditions, puis vient la représentation ΔE et enfin celle en Barks ΔZ. Néanmoins, toutes ces expressions du seuil de fission présentent peu de variations avec la fréquence du son A, et sont donc relativement en accord avec la théorie des canaux.

En testant les deux oreilles de sujets présentant une perte unilatérale, Rose et Moore (1997) n'ont pas observé de différence systématique entre l'oreille endommagée et l'oreille saine. La figure 1.20, à droite, montre les seuils de fission sous la forme ΔE en fonction de l'ERB mesurée sous la forme $\mathrm{ERB}/\mathrm{ERB}_N$. Ce nuage de points montre qu'il n'y a pas de relation claire entre la sélectivité fréquentielle et le seuil de fission pour des sons purs. Ce résultat remet en cause la théorie des canaux.

Dans une étude ultérieure, Rose et Moore (2005) ont comparé le seuil de fission (FB) avec le seuil de discrimination pour des sons purs (FDL), tous deux exprimés sur une échelle d'ERB$_N$. Ils ont observé que, pour des normo-entendants, le rapport FB/FDL était quasiment constant à 8 jusqu'à 2000 Hz, puis qu'il décroissait pour atteindre 1 à 8000 Hz. Pour les malentendants, ce rapport variait de 1 à 40 et n'était pas systématiquement relié à la perte auditive. Ce résultat indique que la discriminabilité de deux sons purs successifs ne contribue que partiellement à la définition du seuil de fission.

En reprenant la méthode utilisée par Rose et Moore (1997), Mackersie, Prida, et Stiles (2001) ont observé que les sujets malentendants qui avaient les moins bons seuils de fission étaient aussi ceux qui avaient le plus de difficultés à séparer deux voix concurrentes. Hong et Turner (2006) ont effectué la même observation chez les implantés cochléaires. Il semble que les implantés puissent former des flux auditifs à partir de stimuli séquentiels en utilisant la périodicité de l'enveloppe temporelle ou la dimension tonotopique (Chatterjee, Sarampalis, et Oba, 2006).

Sons complexes

Pour des sons complexes harmoniques, la sélectivité fréquentielle détermine si les harmoniques peuvent être résolues ou non. Malgré l'importance de ce phénomène sur la perception de la hauteur, la résolvabilité des harmoniques ne semblent pas influencer le seuil de fission (Vliegen et Oxenham, 1999). Les indices temporels seuls semblent donc pouvoir induire du streaming volontaire. En revanche, Vliegen, Moore, et Oxenham (1999) ont conclu que l'information spectrale était dominante pour la formation irrépressible de flux auditifs, bien que l'information temporelle puisse aussi être utilisée dans une moins large mesure.

Dans une tâche subjective (streaming ni volontaire ni automatique), l'importance du degré de résolvabilité pour la ségrégation séquentielle peut être mis en évidence (Grimault, Micheyl, Carlyon, Arthaud, et Collet, 2000). Il est possible de manipuler la résolvabilité des harmoniques en filtrant un son complexe dans différents intervalles de fréquence. Grimault *et al.* (2000) ont utilisé trois régions LOW (125–625 Hz), MID (1375–1875 Hz) et HIGH (3900–5400), dans lesquelles un son complexe de 250 Hz était résolu, partiellement résolu et non résolu. Les sujets devaient simplement indiquer le nombre de

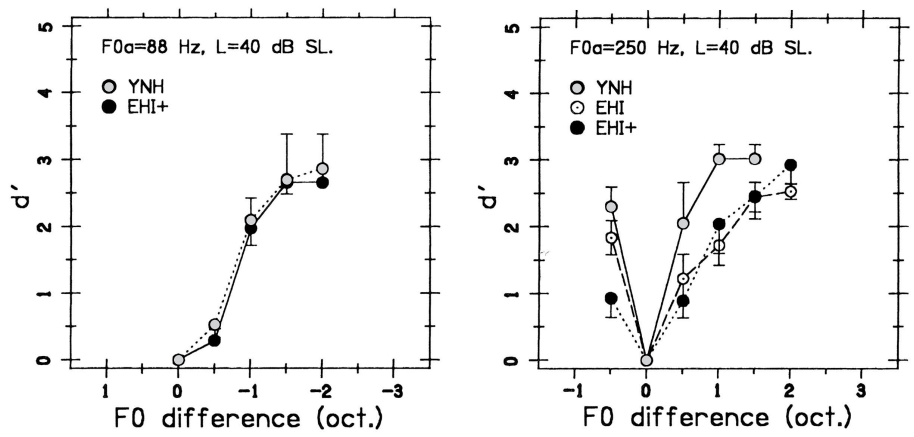

FIG. 1.21 – Scores de streaming en fonction de la différence de F_0 dans la région MID, pour des normo-entendants (YNH), des malentendants âgés ayant une audition normale pour leur âge (EHI), et des malentendants âgés ayant une perte supplémentaire (EHI+). Le cadre de gauche montre les résultats pour un son complexe de F_0 =88 Hz. Le cadre de droite montre les résultats pour un son complexe de F_0 =250 Hz. D'après Grimault *et al.* (2001).

flux entendus. Les résultats montrent qu'il fallait des ΔF_0 de 0,5, 1 et 1,5 demitons dans les conditions LOW, MID et HIGH respectivement pour que la séquence conduise à la perception de 2 flux dans plus de 90% des cas. Ces résultats suggèrent donc une influence de la résolvabilité des harmoniques sur les performances de streaming de sons complexes harmoniques.

Grimault, Micheyl, Carlyon, Arthaud, et Collet (2001) ont reproduit une expérience similaire en impliquant des malentendants. Les résultats, présentés figure 1.21, montrent que les malentendants ont des performances de streaming inférieures à celles des normo-entendants lorsque les sons complexes harmoniques sont résolus uniquement pour ces derniers (cadre de droite, F_0 =250 Hz). En revanche, lorsque les sons sont non-résolus pour les deux groupes de sujets, les performances de streaming sont identiques. Ces résultats peuvent sembler conformes à la théorie des canaux dans la mesure où la résolvabilité des harmoniques, et donc la précision du diagramme d'excitation, détermine les performances de streaming. Néanmoins, dans les cas où les harmoniques sont non-résolues, la théorie des canaux prédirait qu'il ne peut y avoir ségrégation, contrairement à ce qui a été observé.

Lorsque les harmoniques sont non-résolues, elles interfèrent dans les filtres auditifs et la forme temporelle de l'excitation à la sortie du filtre auditif devient sensible aux relations de phase entre harmoniques. En utilisant une tâche de jugement de rythme permettant d'évaluer le streaming irrépressible, Roberts, Glasberg, et Moore (2002) ont montré qu'il était possible d'observer du streaming basé uniquement sur des différences dans la relation de phase. Ces différences de relation de phase n'induisent aucune différence dans le diagramme d'excitation, et ces résultats sont donc en contradiction flagrante avec la théorie des canaux. Les malentendants semblent d'ailleurs aussi sensibles à ces indices de ségrégation (Stainsby, Moore, et Glasberg, 2004a). La relation de phase entre les harmoniques modifie l'enveloppe temporelle du signal. Celle-ci est périodiquement modulée au F_0, et la manipulation de la relation de phase modifie notamment la profondeur de cette modulation. La relation de phase détermine aussi la forme de la structure fine temporelle.

Grimault, Bacon, et Micheyl (2002) ont réalisé des bruits large bande modulés en amplitude par une sinusoïde (SAMN). Ils ont mesuré le seuil de cohérence et le seuil de fission pour ces bruits, chez des normo-entendants, pour des séquences de bruits modulés ne différant que par leur fréquence de modulation. Ces sons ne différaient donc ni par leur contenu spectral, ni par leur structure fine. Le seuil de cohérence trouvé était de 12,4 demitons, tandis que le seuil de fission était de 9 demitons. Les résultats de cette étude indiquent clairement que la fréquence de modulation peut constituer à elle seule un indice provoquant du streaming.

1.4 Conclusion

Plusieurs études indiquent donc que la ségrégation séquentielle est fortement impliquée dans la séparation de voix concurrentes. En particulier, le profit que des sujets normo-entendants et malentendants peuvent tirer d'une différence de hauteur pour les aider à séparer deux locuteurs semble mieux représenté par leurs performances de ségrégation séquentielle que par leurs performances de ségrégation simultanée. Cependant, les indices utilisés pour bénéficier de cette différence de hauteur restent méconnus, aussi bien dans le cas de la séparation de voix concurrentes que pour la ségrégation séquentielle de sons purs et de sons complexes harmoniques. Les indices spectraux de

hauteur ainsi que la périodicité de l'enveloppe temporelle, ou encore la structure fine temporelle pourraient être employés comme l'indiquent plusieurs études. Cependant, ces indices ont été étudiés de façon isolée avec des sons très simples. La complexité des signaux de parole pourrait donc conduire à des résultats très différents.

Les travaux réalisés dans cette thèse visent à éclaircir les indices employés pour bénéficier d'une différence de F_0 entre deux signaux de parole, dans une tâche de ségrégation séquentielle. Nous avons aussi cherché à évaluer dans quelle mesure ces indices étaient pertinents pour la séparation de voix concurrentes. Dans la première étude, présentée au chapitre suivant, nous avons mis en place un paradigme, dérivé de celui employé par Dorman *et al.* (1975), destiné à produire une mesure objective de streaming irrépressible pour des signaux de parole différant par leur F_0. Nous avons utilisé des voyelles quasi-stationnaires comme premier compromis entre les sons complexes harmoniques, employés dans plusieurs études présentées dans ce chapitre, et des signaux de parole plus complexes, comportant des phases transitoires plus importantes. Dans une première expérience, nous avons vérifié que ce paradigme produisait bien la mesure escomptée en s'appuyant sur les caractéristiques connues du streaming. Dans une seconde expérience, nous avons fait une première évaluation de l'importance des indices spectraux en dégradant artificiellement la sélectivité fréquentielle à l'aide d'un algorithme de lissage spectral. Le chapitre 3 est ensuite consacré aux indices temporels d'enveloppe. Pour étudier ces indices, nous avons utilisé un vocodeur généralement employé pour la simulation d'implant cochléaire. Cette seconde étude fournie donc aussi un premier rapport sur la possibilité, pour les implantés, d'utiliser la périodicité de l'enveloppe temporelle pour exploiter la hauteur de la parole dans une tâche de ségrégation. Au chapitre 4, nous avons exploré la variabilité de performances observée pour notre tâche, en l'associant à d'autres mesures psychoacoustiques, en particulier la sélectivité fréquentielle. Enfin l'étude présentée au chapitre 5 visait à l'investigation du streaming de séquences de voyelles chez les malentendants.

Chapitre 2

Effet du lissage spectral sur la ségrégation perceptive de séquences de voyelles

Dans cette première étude, une tâche de perception de l'ordre sur des séquences de 6 voyelles a été développée afin d'obtenir une estimation objective de l'état de ségrégation. Différentes mesures ont été réalisées afin de s'assurer que ce paradigme permettait effectivement de rendre compte du phénomène de streaming irrépressible. Les résultats obtenus donnent le profil du seuil de cohérence temporelle pour un tempo donné, et montrent qu'une différence de hauteur fondamentale permet bien de créer des flux auditifs sur des voyelles.

Dans une seconde expérience, la résolution fréquentielle du signal a été artificiellement manipulée à l'aide d'un algorithme de lissage spectral. Les résultats montrent que le lissage spectral améliorait significativement la capacité des sujets à donner les 6 voyelles dans l'ordre. Ces résultats indiquent que le lissage spectral provoque un déficit de ségrégation. Le profil du seuil de cohérence indique de plus que la ségrégation sur la base de la hauteur, et la ségrégation sur la base de la structure formantique sont toutes deux altérées par le lissage spectral. L'implication de ces résultats pour la perception de voix concurrentes chez les malentendants est discutée.

Cet article a été accepté pour publication dans Hearing Research *le 10 mai 2007.*

Available online at www.sciencedirect.com
ScienceDirect
Hearing Research 231 (2007) 32–41
Hearing Research
www.elsevier.com/locate/heares

Research paper

Effect of spectral smearing on the perceptual segregation of vowel sequences ☆

Etienne Gaudrain [a], Nicolas Grimault [a,*], Eric W. Healy [b], Jean-Christophe Béra [c]

[a] Neurosciences Sensorielles, Comportement, Cognition, CNRS UMR 5020, Université Claude Bernard, Lyon 1, France
[b] Speech Psychoacoustics Laboratory, Department of Communication Sciences and Disorders, University of South Carolina, Columbia, 29208, USA
[c] Inserm U556, Lyon, France

Received 1 September 2006; received in revised form 30 April 2007; accepted 10 May 2007
Available online 21 May 2007

Abstract

Although segregation of both simultaneous and sequential speech items may be involved in the reception of speech in noisy environments, research on the latter is relatively sparse. Further, previous studies examining the ability of hearing-impaired listeners to form distinct auditory streams have produced mixed results. Finally, there is little work investigating streaming in cochlear implant recipients, who also have poor frequency resolution. The present study focused on the mechanisms involved in the segregation of vowel sequences and potential limitations to segregation associated with poor frequency resolution. An objective temporal-order paradigm was employed in which listeners reported the order of constituent vowels within a sequence. In Experiment 1, it was found that fundamental frequency based mechanisms contribute to segregation. In Experiment 2, reduced frequency tuning often associated with hearing impairment was simulated in normal-hearing listeners. In that experiment, it was found that spectral smearing of the vowels increased accurate identification of their order, presumably by reducing the tendency to form separate auditory streams. These experiments suggest that a reduction in spectral resolution may result in a reduced ability to form separate auditory streams, which may contribute to the difficulties of hearing-impaired listeners, and probably cochlear implant recipients as well, in multi-talker cocktail-party situations.

Keywords: Streaming; Vowel sequences; Hearing impairment; Spectral smearing

1. Introduction

Bregman (1990) suggested that auditory scene analysis involves the ability to decompose a sound mixture into percepts corresponding to various acoustic sources. The mechanisms involved in this analysis have been described in the literature in terms of stream segregation. Following Bregman's classification, two mechanisms are usually described: one related to simultaneous sounds, the other related to sequential sounds. Using these mechanisms, most people are able to focus on a single talker and understand what is being said despite the presence of competing signals. Unfortunately, this ability may be diminished in hearing-impaired (HI) listeners and cochlear-implant (CI) recipients.

It seems obvious that segregation of simultaneously-occurring sounds is involved in the ability to understand speech in noisy backgrounds. However, Mackersie et al. (2001) demonstrated a relationship between the speech reception threshold (SRT) and the fusion threshold (as defined by Rose and Moore, 1997) suggesting that the reception of speech in noise may also be related to segregation of sequentially-occurring sounds (streaming). In fact, the weak relationship found between the SRT and simultaneous segregation in HI listeners (Summers and Leek,

☆ Portions of this work were presented in "Segregation of vowel sequences by normal-hearing and hearing-impaired listeners," Paper presented at the XII International Symposium on Audiological Medicine, Lyon, France, March 2005.

* Corresponding author. Tel.: +33 4 37 28 74 89; fax: +33 4 37 28 76 01.
E-mail address: ngrimault@olfac.univ-lyon1.fr (N. Grimault).

0378-5955/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.heares.2007.05.001

1998) raises the possibility that sequential segregation may be an even better predictor of speech in noise reception than simultaneous segregation.

Simultaneous segregation has been examined by a number of investigators using both non-speech and speech stimuli (for review, see de Cheveigné, 1999). In contrast, sequential segregation has primarily been investigated using pure or complex tones (for review, see Moore and Gockel, 2002). Thus, although both segregation mechanisms are potentially involved in the recognition of speech in noisy backgrounds, and despite the fact that both mechanisms are potentially impaired by reductions in frequency selectivity, empirical examinations of sequential segregation of speech are relatively sparse.

1.1. Segregation with reduced spectral cues

Although mixed, there is some evidence that listeners with HI have reduced stream segregation abilities. Grose and Hall (1996) employed a pair of tasks and found that listeners with cochlear hearing loss generally required a greater frequency separation for segregation of sequential pure tones. However, in contrast to this work and the theory of Hartmann and Johnson (1991), Rose and Moore (1997) found no consistent difference between ears of unilaterally-impaired listeners in the frequency difference required to segregate pure tones.

Recent work has also suggested that segregation may be impaired in CI users. Qin and Oxenham (2005) found that normal hearing (NH) listeners exposed to speech-vocoder simulations of a CI were unable to benefit from fundamental frequency (F_0) differences in concurrent-vowel identification. Carlyon et al. (2007) reported that CI users were unable to benefit from temporal pitch differences between channels to separate concurrent sounds. Cooper and Roberts (2007) employed pure-tone stimuli and also found little evidence of stream segregation in CI users. However, other investigators have observed that some users are able to perceptually segregate stimuli (e.g. Chatterjee et al., 2006). The importance of segregation in normal communication was highlighted by Hong and Turner (2006), who found that CI users who performed better on a streaming task also performed better on speech recognition in noise.

This possible reduction in segregation by HI and CI listeners could be related to reduced frequency specificity. Both Arehart et al. (1997) and Summers and Leek (1998) found that HI individuals benefited less from F_0 differences across voices in concurrent-vowel identification tasks. Similarly, Qin and Oxenham (2005) suggested that limited concurrent-vowel performance in their CI vocoder simulations was likely due to the limited spectral representation. However, Rose and Moore (2005) found large variability in the ratio between frequency discrimination in HI listeners and the frequency difference required to segregate pure tones. Thus, the effects of reduced frequency selectivity in the segregation of speech signals remain unclear.

Evidence from experiments involving flat-spectrum complex tones suggests more strongly that frequency tuning may play a role in segregation by showing that resolvability of harmonics is an important cue for segregation. Vliegen and Oxenham (1999), Vliegen et al. (1999), Grimault et al. (2000, 2001), Roberts et al. (2002) and Stainsby et al. (2004a,b) all employed complex tone sequences filtered to restricted spectral regions. For a given F_0, harmonics were generally resolved in low-frequency conditions, and unresolved in high-frequency conditions (except in Roberts et al., 2002, where they were always unresolved). It was generally found that streaming was weakened, though not absent, when components were unresolved. Thus, the difficulties of HI listeners in cocktail party situations may potentially be related to a loss of resolvability that would impair streaming mechanisms.

However, the complex-tone stimuli used in these experiments differ substantially from speech. These stimuli were restricted in frequency, and components were generally either resolved or unresolved. In ecological situations, HI listeners may resolve the lower portions of the broadband signal, but not the higher portions, and this resolution may change over time as a result of F_0 fluctuation. Also, these studies employed flat-spectrum complex tones lacking the formant structure that may affect streaming performance (e.g., Dorman et al., 1975; Singh, 1987; Bregman et al., 1990). Finally, speech may benefit from specific schema-driven mechanisms as suggested by Bregman (1990) and Remez et al. (1994).

1.2. Streaming with speech stimuli

Various acoustic cues can induce sequential segregation. For complex tone sequences (as a first approximation of speech), streaming seems to be influenced by two main factors: pitch and timbre (Bregman et al., 1990; Singh, 1987; Singh and Bregman, 1997). However, the timbre variations applied to the non-speech stimuli have involved elimination of harmonics or, at best, spectral shaping using a single formant. Thus, the influence of multi-formant timbre in streaming of speech is unclear.

There is limited work employing speech signals. Following many studies involving the perception of temporal order (Hirsh, 1959; Warren et al., 1969; Thomas et al., 1970; Lackner and Goldstein, 1974), Dorman et al. (1975) examined the influence of formant differences on streaming using four-item vowel sequences. The authors employed sequences of items having a constant F_0 and found that the ability to perceive the items in the correct order was dependent upon the sequence being perceived as a single auditory stream (see also Bregman and Campbell, 1971). It was concluded that, in the absence of formant transitions, vowel sequences of constant pitch could induce stream segregation. Darwin and Bethell-Fox (1977) observed that streaming can also occur with formant transitions if abrupt discontinuities exist in the pitch contour.

More recently, Bregman et al. (1990) examined the relative importance of, and possible interaction between, streaming based on pitch and streaming based on a single formant (spectral peak). Sequences consisted of four complex tones (A, B, C, and D) that differed in F_0 and/or frequency of the spectral peak. The sequences started with a looped pattern AB-- (where the symbol '-' represents a silent gap). After 20 repetitions, tones C and D were added to form the pattern ABCD. Subjects were then asked to judge on a five point scale (1 = hard, 5 = easy) how easily they could hear the standard pair AB in the pattern ABCD. Spectral peak positions and F_0s were manipulated independently. On some trials, the two tones comprising the standard were similar in F_0; on such trials, F_0 was the tested factor and spectral peak position the interfering factor. On other trials, the two tones in the standard were similar in spectral peak position; on such trials, spectral peak position was the tested factor and F_0 the interfering factor. Fundamental frequencies ranged from 128 to 277 Hz, and spectral peak positions ranged from 1000 to 2161 Hz. If the tested factor was dominant for segregation, tones A and B should have been segregated from tones C and D and the task should have been judged easy by the subject. In this subjective measure of streaming, the authors concluded that spectral peak position affected streaming more strongly than F_0. In a second experiment, the effect of spectral peak sharpness on streaming was evaluated. Triangularly-shaped peaks two octaves in width were employed. The height of the triangle (in relative dB) defined peak sharpness. It was found that broadening of the peak tended to weaken the streaming effect, further suggesting a relationship between tuning and streaming.

Nooteboom et al. (1978) may have provided the only systematic investigation of the effect of pitch on the segregation of sequences of vowels. The authors employed short sequences of nine synthesized vowels using the pattern /aᵘiaᵘiaᵘi/. The F_0 alternated between 100 Hz and another fixed value between 100 and 280 Hz. It was found that, for realistic speech rates (from 3 to 10 vowels/s), an F_0 difference between approximately two and five semitones produced segregation. However, that early study had substantial limitations. First, because the sequences were brief (from 1 to 4 s) and presented only once, the streaming effect was not stabilized when the subjects issued their response (cf. Bregman, 1978). Second, a subjective measure of streaming was employed in which subjects simply reported hearing one or two voices. Finally, only two subjects were examined. As a consequence, streaming with vowel sequences deserves further examination, both under normal conditions, and under conditions of reduced frequency tuning.

1.3. Rationale

Although sequential segregation of speech sounds plays a potentially important role in the reception of speech in noise, it has not been well studied. Further, the influence of reduced frequency selectivity on this ability to form separate auditory streams is not well understood.

In the present experiments, streaming was observed through an objective method based on the perception of temporal order. Looped sequences of vowels were presented to subjects who were required to identify the correct order of occurrence. Accurate identification is assumed possible only if the items form a single auditory stream. Because attention in this task is directed against streaming, the observed segregation is only that which cannot be suppressed, generally referred as *automatic* or *obligatory streaming*. Obligatory streaming relates to *primitive* mechanisms that should be dependent upon presentation rate (van Noorden, 1975; Bregman, 1990). In Experiment 1, the role of presentation rate and F_0 differences across vowels were investigated. In a second experiment, the influence of spectral smearing of speech sounds on the formation of separate auditory streams was assessed. These conditions provide information concerning the influence of broadened auditory tuning on the ability to segregate sequential speech signals.

2. Experiment 1: Intact vowel sequences

The purpose of this experiment was to examine the formation of separate auditory streams with sequential vowel stimuli using an objective method. Conditions in which pitch was held constant and items varied only in formant structure were employed, as were conditions in which alternate items had different F_0 values. It is worth noting that, unlike previous studies involving complex tones, the present experiment required identification of constituent items, ensuring recognition of the speech at least at a phonemic level.

2.1. Subjects

Ten young NH listeners aged 20–27 years (mean 23.5) participated in this experiment. All were native speakers of French, and all had pure-tone audiometric thresholds below 15 dB HL at octave frequencies between 250 and 4000 Hz. All were paid an hourly wage for participation, and none had participated in similar experiments previously.

2.2. Stimuli

Six French vowels /a e ɪ ɔ ʊ y/ were generated using a cascade-resonance synthesizer (Klatt, 1980) at 10 different fundamental frequencies (100, 110, 121, 134, 147, 162, 178, 196, 216, and 238 Hz). Durations of 135 and 175 ms were selected to be close to those associated with natural speech. Each vowel onset and offset was smoothed with a 10 ms cosine ramp. Vowels were chosen based on their extreme positions within the vowel space defined by the first and second formants. The center frequency and bandwidth values of formants are presented in Table 1. All vowels were adjusted to have the same RMS power.

Table 1
Values of formant frequencies and bandwidths for French vowels adapted from Tessier (2001)

Vowel	F_1 (Δ_{f1})	F_2 (Δ_{f2})	F_3 (Δ_{f3})
a	750 (75)	1344 (60)	2510 (84)
e	370 (55)	1900 (74)	2700 (100)
I	250 (55)	2000 (50)	3000 (120)
ɔ	380 (53)	850 (63)	2460 (70)
ʊ	244 (60)	750 (70)	2000 (100)
y	224 (74)	1728 (80)	2069 (83)

F_i is the center frequency of the ith formant in Hertz. The bandwidth Δ_{fi} is given in Hertz, between braces. For all vowels, $F_4 = 3300$ Hz ($\Delta_{f4} = 250$ Hz), $F_5 = 3850$ Hz ($\Delta_{f5} = 300$ Hz), and $F_6 = 4900$ Hz ($\Delta_{f6} = 1000$ Hz).

The vowels were organized into sequences containing the six items (see Fig. 1). The F_0 of successive items alternated, so that three items were at $F_{0(1)}$ and the alternate three were at $F_{0(2)}$. $F_{0(1)}$ was always 100 Hz, and $F_{0(2)}$ ranged up to 238 Hz and was constant for a given sequence. Half the sequences started with $F_{0(1)}$, and the other half started with $F_{0(2)}$. The sequences were presented in recycling fashion, so the number of different possible sequences was 6!/6, or 120. One hundred of the 120 possible permutations were randomly selected for inclusion. For each of the 10 F_0 conditions at each presentation rate, 10 sequences were randomly selected (without replacement) that differed only in the order of the vowels. The order of items comprising each sequence was determined independently for the two presentation rates. The sequences were built by concatenating vowels with no silent gap, so that the steady state portions of the vowels were separated by the two 10 ms ramps. No additional fade-in was applied to the sequences. The 16 bit, 44.1 kHz sequences were generated with MATLAB.

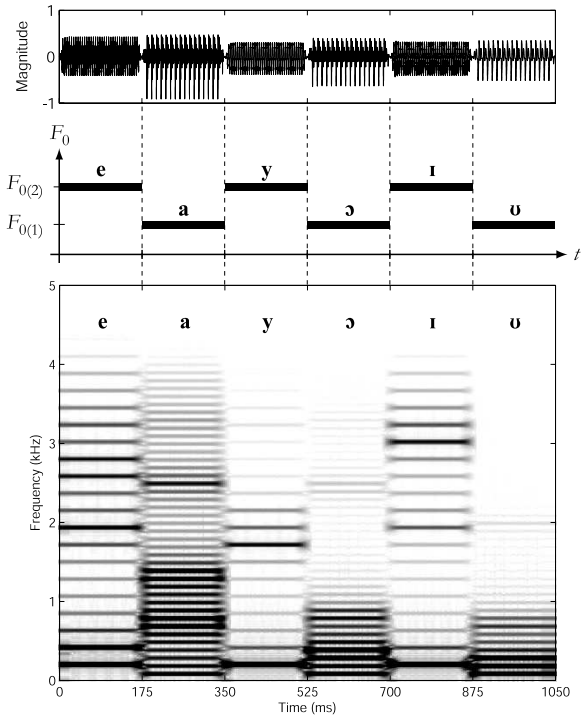

Fig. 1. *Upper panel*: Waveform of a sequence (/e a y ɔ I ʊ/), $F_{0(1)} = 100$ Hz, $F_{0(2)} = 216$ Hz, where each vowel is 175 ms in duration. *Middle panel*: A schematic representation of the F_0 pattern. *Lower panel*: Sonogram of the sequence.

2.3. Procedure

A preliminary identification task ensured that the individual vowels were easily identifiable. The six vowels were presented individually in random order at F_0s of 100, 110, 147 and 238 Hz, with 10 repetitions for a total of 240 presentations. Identification was found to be over 99% accurate.

In each of two subsequent sessions, subjects heard two blocks consisting of 100 sequences each. One block consisted of high-rate sequences (135 ms/vowel, 7.4 vowels/s) and one block consisted of low-rate sequences (175 ms/vowel, 5.7 vowels/s). Five subjects heard the high-rate block prior to the low-rate block, and the five remaining subjects heard the opposite order. Presentation order of sequences within block was randomized for each listener. The stimuli were presented diotically via a Digigram VxPocket 440 soundcard and Sennheiser HD-250 Linear II headphones in a double-walled sound booth. The level of the steady-state portions of the signal was calibrated to 85 dB SPL in an artificial ear (Larson Davis AEC101 and 824).

Each block began with presentation of the isolated vowels followed by presentation of two sample sequences with feedback. The subjects were then instructed to report the correct order of appearance of the six vowels in each sequence. This led to an ACROSS score that reflects the proportion of responses in which the six vowels were identified in the correct order, allowing circular permutations. When they judged this task impossible, subjects were instructed to report the order of the vowels within each stream one after the other. This led to a WITHIN score that reflects the proportion of responses in which the vowels comprising each F_0 group were reported in the correct order, allowing circular permutations within each group. For example, for sequence /e a y ɔ ɪ ʊ/, the response /e a y ɔ ɪ ʊ/ would increase the ACROSS score, and the response /e y ɪ a ɔ ʊ/ or /a ɔ ʊ e y ɪ/ would increase the WITHIN score.

Subjects provided their response, starting with any vowel, using a computer mouse interface while listening to the repeating sequence. At the start of the auditory stimulus, the response screen displayed "Listen" for 5 s, then "Answer" and six columns each containing the six different vowels with radio buttons. The subject had to check one radio button in each column. When all buttons had been checked, a "Submit" button appeared. The stimuli stopped when the subject was satisfied with the response and clicked this button. The subject was locked-out from responding

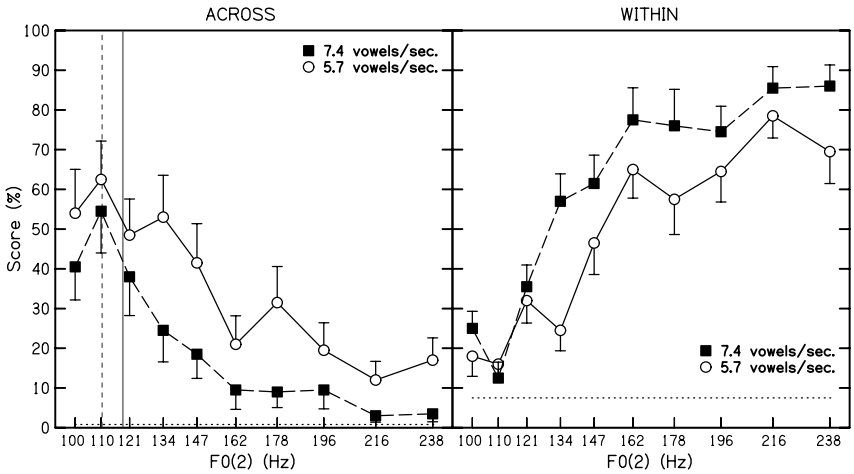

Fig. 2. *Left panel*: ACROSS scores (accurate identifications of the order of items in a six-vowel sequence) expressed in percent as a function of the fundamental frequency of alternate items. Filled squares (■) represent group mean scores for high speech-rate sequences (7.4 vowels/s). Open circles (○) represent group mean scores for low speech-rate sequences (5.7 vowels/s). Error bars represent standard deviations. Chance level is plotted with a horizontal dotted line. The vertical lines indicate the approximate locations at which subjects from Nooteboom et al. (1978) reported hearing a single voice (on the left side of the lines) or two voices (on the right side) for a speech-rate fixed at 7.4 vowels/s (dashed vertical line) or at 5.7 vowels/s (solid vertical line). *Right panel*: WITHIN scores (accurate identifications of the order of vowels at each F_0) expressed in percent as a function of the fundamental frequency of alternate items. The legend is the same as in the left panel. Chance level is plotted with a horizontal dotted line.

during the initial five seconds of exposure to each sequence to allow the streaming effect to stabilize (Bregman, 1978). No feedback was provided. Each session lasted approximately 35 min. All experimental paradigms were formally approved by a local ethics committee (CCPPRB Léon Bérard).

2.4. Results

The number of sequences identified in the correct order (ACROSS score) was tallied for each individual across the two blocks, in each F_0 and rate condition, yielding a score that ranged between 0 and 20. Scores are expressed in percent, with 100% corresponding to a score of 20. The ACROSS score averaged across individuals is plotted as a function of $F_{0(2)}$ in Fig. 2 (left panel). Chance performance is 0.8%. As previously described in the literature, high scores can be interpreted as a tendency toward integration across the $F_{0(1)}$ and $F_{0(2)}$ items and a resistance to streaming. All subjects demonstrated decreasing scores with increasing F_0 difference for both presentation-rate conditions. For small F_0 differences, mean scores are approximately 50% accurate responses, whereas for an F_0 difference larger than one octave, mean scores fall to about 10%. A two-way ANOVA[1] using F_0 and presentation rate as repeated parameters indicated that the effect of F_0 was significant [$F(9,81) = 21.14$, $p < 0.001$] and that low-rate sequences were accurately identified more often than high presentation-rate sequences [$F(1,9) = 13.81$, $p < 0.01$]. The interaction was also significant [$F(9,81) = 2.21$, $p < 0.05$].

The number of sequences in which the vowels comprising each F_0 group were reported in the correct order (WITHIN score) was also tallied for each individual across the two blocks, in each F_0 and rate condition, yielding a score that ranged between 0 and 20 expressed in percent. The WITHIN score averaged across individuals is also plotted as a function of $F_{0(2)}$ in Fig. 2 (right panel). Chance level equals 7.5%. A two-way ANOVA[2] using F_0 and presentation rate as repeated parameters indicated that the effect of F_0 [$F(9,81) = 49.97$, $p < 0.0001$] and presentation rate [$F(1,9) = 10.58$, $p < 0.01$] were significant and interacted [$F(9,81) = 5.06$, $p < 0.0001$].

2.5. Discussion

The expected strong effect of F_0 was found. Analyzing the ACROSS scores, larger F_0 differences led to more streaming, which made order judgments more difficult. No discrepancy in the pattern of scores was observed near the octave relation across alternate items. The fact that listeners remain good at judging the relative order of items sharing the same F_0 (WITHIN score) at large F_0 differences strengthens the argument that stream segregation is the key factor driving the decline in ACROSS performance as F_0 difference increases. Streaming was presumably quite

[1] Identical analysis performed on rationalized arcsine transformed ACROSS scores (RAU, Studebaker, 1985): F_0 [$F(9,81) = 23.75$, $p < 0.001$], presentation rate [$F(1,9) = 20.66$, $p < 0.01$], interaction [$F(9,81) = 1.87$, $p = 0.07$].

[2] Identical analysis performed on RAU WITHIN scores: F_0 [$F(9,81) = 46.76$, $p < 0.0001$], presentation rate [$F(1,9) = 11.80$, $p < 0.01$], interaction [$F(9,81) = 5.09$, $p < 0.0001$].

strong in the high-rate conditions over 162 Hz, in which F_0 identification was especially poor and within F_0 identification was especially good. This floor effect presumably contributed to the interaction. However, as detailed further, other factors might also contribute to this interaction.

Approximations of the thresholds from Nooteboom et al. (1978) are also shown in Fig. 2. The vertical lines correspond to the points at which the percept changed from a single voice to two voices. They correspond to an F_0 difference for which the subjects in the current study accurately identified approximately 50% of the sequences presented. It is potentially interesting to note that the experimental paradigm used in Nooteboom et al. (1978), in which subjects reported the number of voices heard, may prevent a reliable estimation of streaming based on differences in item identity (formant structure) and not based on pitch differences because voices are often characterized by their pitches. It is possible that a sequence segregated on the basis of formant differences would have been reported as emanating from a single talker in Nooteboom et al. (1978).

The subjects were unable to accurately identify the order of items (they had low ACROSS scores) at high values of $F_{0(2)}$. In addition to illustrating the F_0 effect, this shows that subjects were unable to develop strategies to overcome the streaming effect. In particular, as sequences were not faded in slowly over time, it might have been thought that subjects could benefit from exposure to the sequence before streaming developed. However, had any such strategy been successfully used, subjects should have been able to accurately identify the order of items at high values of $F_{0(2)}$.

The objective task employed likely directed attention away from segregation, as segregation tended to prevent accurate performance. Moreover, the mechanisms underlying segregation were sensitive to presentation rate. Because the temporal coherence boundary depends strongly on the tempo of the sequence, while the fission boundary is relatively independent of this parameter (van Noorden, 1975; Bregman, 1990), the strong effect of presentation rate in this experiment suggests that the paradigm provides a reliable estimation of temporal coherence, i.e., primitive segregation. When two speakers are speaking concurrently, it seems reasonable to assume that these utterances will not be entirely simultaneous. The primitive mechanisms of sequential segregation based on F_0 may then contribute to the understanding of speech-in-speech.

Overall, the results from the current experiment compare well with early reports and indicate that differences in both pitch (F_0) and timbre appear to impair the perception of temporal relationships between vowels within a sequence, and are potentially important factors leading to sequential segregation of speech. The current study provides advantages over previous work, by providing an estimate of the influence of F_0 on streaming of vowel sequences using a larger number of subjects, an objective measurement of obligatory streaming, and a method that ensures that the stimuli are recognized as speech.

3. Experiment 2: Smeared vowel sequences (hearing-loss simulation)

The investigation of streaming in speech stimuli under conditions of reduced spectral cues is potentially important for understanding the difficulties encountered by HI (and CI) listeners in multi-talker cocktail party situations. If broadened auditory tuning and limited access to pitch and timbre cues produce less streaming, then the simulation of broadened auditory tuning in the current experiment should allow more accurate identification of items in the correct order, because the interfering effect of streaming is reduced. It is worth noting that, because any segregation deficit should lead to better performance in the current paradigm, the results cannot be attributed to intelligibility (vowel identification) impairment or to any increase in cognitive load resulting from the broadening of the stimuli.

The use of young NH subjects in the current experiment allows the elimination of many difficulties encountered when testing HI individuals. These listeners had homogeneous and sharp auditory tuning, and broadened tuning was simulated by spectral smearing of the acoustic stimuli (after Baer and Moore, 1993). This ensured similar cochlear resolution across subjects. Further, the use of NH subjects avoids effects of loudness recruitment and ensures preserved cochlear compression. Possible effects of advanced age are also eliminated. Moreover, the procedure reduces intersubject variability and strengthens the statistical power of the smearing effect by simulating the loss of cochlear resolution within instead of across individuals.

3.1. Subjects

Ten French-speaking listeners, aged 21–29 years (mean 23.9), participated. All had pure-tone thresholds of 15 dB HL or better at octave frequencies from 250 to 4000 Hz. None participated in Experiment 1 or had previously taken part in any other similar experiment.

3.2. Stimuli

The 100 sequences having the lowest perceptual distances (de Boer, 2000)[3] were selected from the 120 possible permutations. In an attempt to provide greater similarity in sequences comprising each condition, the sets were generated so that the mean perceptual distances for each F_0 condition were similar. One set of sequences was constructed

[3] For each sequence, the perceptual distance between vowels was calculated using formulas (2)–(5) presented in de Boer (2000). The distance between two vowels is the Euclidian distance in a two-dimensional space in which dimensions are first formant and *effective second formant* calculated as the weighted sum of the second to fourth formants. The perceptual distance for a given sequence was estimated as the sum of the distances that separate each contiguous pair of vowels.

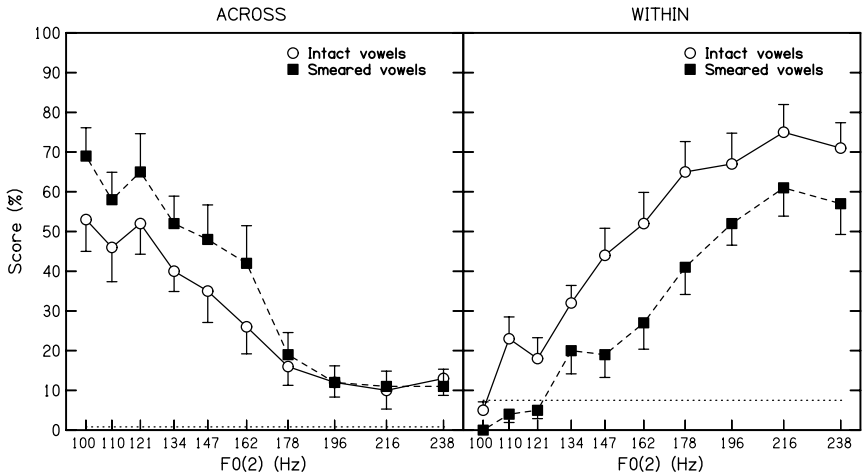

Fig. 3. *Left panel*: ACROSS scores (accurate identifications of the order of items in a six-vowel sequence) expressed in percent as a function of the fundamental frequency of alternate items. Results for sequences of spectrally-smeared vowels are plotted with filled symbols (■), and results for sequences of intact vowels are plotted with open symbols (○). Error bars represent standard deviations. Chance level is plotted with a horizontal dotted line. *Right panel*: WITHIN scores (accurate identifications of the order of vowels at each F_0) expressed in percent as a function of the fundamental frequency of alternate items. The legend is the same as in the left panel. Chance level is again plotted with a horizontal dotted line.

using the 175 ms vowels from Experiment 1, and a second set of sequences was created using spectrally-smeared vowels. The order of items within sequences comprising the Smeared and Intact conditions was identical.

The smeared vowels were generated by modifying the intact items using the algorithm of Baer and Moore (1993), with simulated auditory filters set to three times broader than normal. Although there is considerable variability in the relation between audiometric threshold and tuning, an auditory filter enlargement of three times normal would correspond to absolute thresholds from 30 to 60 dB higher than normal (Moore, 1998).

The technique of Baer and Moore (1993) involved first windowing the input signal using a Hamming window (8 ms) with an overlap (4 ms). For each time window, the spectrum was computed using a fast Fourier transform and smeared. The smearing process was performed by convolving the power spectrum with a smearing function. This smearing function was a bank of broadened, symmetrical, and normalized roex(p) filters (Patterson et al., 1982) simulating an impaired cochlea, multiplied by a bank of inverse normalized roex(p) filters simulating a normal cochlea. This evokes excitation patterns in a normal ear that resemble those that would be evoked in an impaired ear using unsmeared stimuli. Each smeared spectrum was then transformed to the time domain using an inverse fast Fourier transform. All time windows were then added using an overlap and add method to obtain the smeared output. The stimuli were processed using MATLAB.

3.3. Procedure

The session began with an identification test on the smeared vowels at 100, 147, and 238 Hz. Each vowel was repeated five times in each F_0 condition, resulting in a total of 90 presentations. Subjects repeated this test until identification of smeared vowels reached 94% accuracy (85/90). On average, the subjects needed 2.3 repetitions to reach this value. This identification test was then followed by two blocks of the streaming test as in Experiment 1. Each block was composed of 50 smeared and 50 intact sequences. Each sequence (each particular vowel order) appeared smeared in one block and intact in the other block. Half the subjects heard the smeared block first and the other half heard the intact block first, and presentation order of sequences within each block was randomized for each listener. The apparatus and other procedures were the same as those of Experiment 1.

3.4. Results

Group mean data are presented in Fig. 3. The results observed in this experiment for intact vowels are consistent with those of Experiment 1, including a starting point at approximately 55% accurate responses across F_0s at $F_{0(2)} = 100$ Hz, followed by a decrease in ACROSS scores to an asymptote around 10% accurate response. For eight of the 10 subjects, ACROSS scores for smeared sequences were greater than (again, reflecting less streaming), or equal to, ACROSS scores for intact sequences over all F_0 conditions. A two-way ANOVA[4] applied to the ACROSS scores involving F_0 and smearing condition as repeated parameters revealed a significant effect of smearing [$F(1,9) = 5.46$, $p = 0.04$] and F_0 [$F(9,81) = 34.21$, $p < 0.001$]. The

[4] Identical analysis on RAU ACROSS scores: Smearing [$F(1,9) = 4.90$, $p = 0.05$], F_0 [$F(9,81) = 38.60$, $p < 0.001$], interaction [$F(9,81) = 1.69$, $p = 0.10$].

interaction was not significant [$F(9,81) = 1.50$, $p = 0.16$]. Finally, a contrast between the matched F_0 conditions revealed higher scores in the Smeared condition [$p = 0.016$].

The WITHIN scores rose with increasing $F_{0(2)}$ in both the Intact and Smeared conditions. The results of an ANOVA[5] applied to the WITHIN scores were consistent with those for the ACROSS scores. The effect of smearing [$F(1,9) = 22.70$, $p < 0.005$] and F_0 [$F(9,81) = 50.44$, $p < 0.0001$] were significant, but did not significantly interact [$F(9,81) = 1.48$, $p = 0.17$].

3.5. Discussion

Identification of component order across F_0s was more accurate in conditions in which the vowels were smeared, relative to the intact conditions. Because poorer order identification performance is an indication of segregation, it may be concluded that segregation was impaired by spectral smearing of the vowel stimuli that simulated a typical broadening of auditory filters associated with cochlear hearing loss. It is worth noting that even under smearing conditions in which vowels were more difficult to identify, subjects produced better sequence-order identification scores. Since identical orders of items were employed for sequences across conditions, spectral smearing is the likely cause of the observed differences in segregation performance. This interpretation is strengthened by the detrimental effect of smearing upon order identification within F_0s (WITHIN scores).

4. General discussion

This study is directed toward clarifying the mechanisms that enable sequential streaming. Together, Experiments 1 and 2 showed that an F_0 difference is a strong cue for sequential segregation. This is consistent with the literature involving both complex tones (e.g., Singh, 1987; Bregman et al., 1990; Moore and Gockel, 2002) and vowels (Darwin and Bethell-Fox, 1977; Nooteboom et al., 1978). In particular, the present study strengthens the results of Nooteboom et al. (1978).

The strong effect of F_0 on segregation occurs despite the large timbre differences that exist across vowels. This suggests that this mechanism could also apply to everyday speech. Moreover, timbre is sometimes described in the literature as the strongest cue for streaming (Bregman, 1990; Bregman et al., 1990). It appears that timbre differences across vowels are not strong enough to lead to full sequential segregation, as the current task remains possible even with no formant transitions between vowels. This could be due to the presence of six formants instead of one in previous experiments aimed at determining the effect of timbre upon segregation. This suggests that single formant stimuli may not accurately represent segregation of speech. It is noteworthy that in ecological situations, formant transitions may hinder formant-based streaming, but pitch-based segregation still occurs (Darwin and Bethell-Fox, 1977).

In Experiment 1, average ACROSS scores were below 65% even in the most favorable conditions. One interpretation is that the order judgments were simply difficult and independent of streaming. However, this interpretation is not consistent with the results from Experiment 2 in which ACROSS scores at matched F_0 condition were higher in the degraded (smeared) condition than in the Intact condition. It is unlikely a degraded condition would lead to a better order judgment. This raises the alternative interpretation that order judgments were hindered by some streaming, even in the matched F_0 conditions.

Support for this interpretation comes from Dorman et al. (1975) who found that vowel sequences with constant pitch and no formant transitions could be perceived as segregated. Indeed, the ACROSS scores in the lowest $F_{0(2)}$ conditions compare well to those found by Dorman et al. (about 60% correct) even if direct comparison between studies is difficult because of differences in the experimental paradigms.

Support for this interpretation also comes from a subsequent analysis in which a correlation was found between the perceptual distance between vowels (de Boer, 2000)[3] and average ACROSS scores at $F_{0(2)} = 100$ Hz: The greater the perceptual distance across the $F_{0(1)}$ and $F_{0(2)}$ items, the lower the score (i.e., greater tendency toward streaming). Correlations were $r(8) = .79$, $p = 0.01$ for the low rate and $r(8) = .68$, $p = 0.03$ for the high rate. Although it may be most pronounced in conditions in which F_0 values were similar, streaming related to formant structure would tend to decrease ACROSS scores in all F_0 conditions.

Additional support for some streaming in the absence of an F_0 difference comes from Bregman et al. (1990) who attempted to determinate the relative importance to segregation of formant peak separation and F_0 separation. To estimate the effect of peak differences for vowels in the current study, the vowel distance[3] in Barks was used. Mean vowel distance for a given sequence varied from 2.46 to 4.45 Barks. Such formant differences would correspond to peaks at 1471 and 1979 Hz relative to a peak at 1000 Hz in Bregman et al. Interpolating Bregman et al.'s data for those peak values in his Table 3, yields mean clarity scores of 7.2 and 8.8. This suggests that, for the entire range of vowel distances used in the current study, the sequences should have been segregated, making the main task very difficult. This result supports the idea that the approximately 50% errors in the matched-F_0 conditions could be attributed to segregation based on the different vowel items having different formant structures. However, while Bregman et al. found almost no influence of F_0 when considered as an interfering factor (Table 3), F_0 had a large influence on segregation performance in the current study. This may be somewhat surprising given that the presentation

[5] Identical analysis on RAU WITHIN scores: Smearing [$F(1, 9) = 26.56$, $p < 0.001$], F_0 [$F(9,81) = 62.03$, $p < 0.0001$], interaction [$F(9,81) = 0.96$, $p = 0.48$].

rate of successive sounds is even faster in Bregman et al. (100 ms) than in the current study.

The discrepancy between the current results and those of Bregman et al. may be attributable to at least two sources. Perhaps the most probable involves methodological differences: Bregman et al. used a subjective judgment that encouraged perceptual segregation. In the current study, an objective measurement of streaming was employed that required fusion. The second possibility involves the use of vowels containing multiple formants in the current study compared to the single-formant stimuli employed by Bregman et al., and the fact that peak distance (in Bregman et al., 1990) may not be exactly comparable to perceptual vowel distance in the current study.

The effect of vowel distance may also explain in part the fact that ACROSS scores do not decrease monotonically. The randomly selected sequences in each $F_{0(2)}$ condition may differ in the ease with which order can be identified. This may contribute to the observed interaction between F_0 and presentation rate in Experiment 1. In an effort to reduce variability across conditions, formant distance was controlled in Experiment 2.

In Experiment 2, it was shown that streaming was reduced when spectral resolution of the stimuli was reduced. As these mechanisms are probably involved in speech-in-speech understanding, it can be argued that they could contribute to the difficulty displayed by HI listeners in multi-talker environments. Given the strength of the F_0 cue for segregation of voices in the current study, it may be assumed that CI users, who have F_0 difference limens roughly one order of magnitude poorer than their NH counterparts (Rogers et al., 2006), will also experience considerable difficulty segregating voices in multi-talker environments. This assumption is supported by studies indicating relatively poor segregation abilities in CI users (e.g. Carlyon et al., 2007; Cooper and Roberts, 2007). It is noteworthy that the current paradigm yields better performance in degraded conditions. This paradigm could therefore prove useful for evaluating primitive segregation in HI listeners, as any increases in performance could not be attributed to cognitive impairment, language impairment or identification impairment.

In the smeared vowels, both harmonics and formants were degraded. However, since harmonics are spaced more closely than formants, smearing may be assumed to have a larger detrimental effect on the perception of harmonics than the perception of formants. This observation is consistent with the ability of subjects to accurately identify the smeared vowels. However, it is also true that component order judgments (ACROSS scores) were more accurate in the Smeared condition (i.e., streaming was reduced) when items all had the same F_0. This result suggests that a broadening of the auditory filters by a factor of three also affects the perception of formants sufficiently to reduce formant-based segregation. This is consistent with the results of the second experiment of Bregman et al. (1990) in which a decrease in segregation accompanied an increase in formant bandwidth by a factor of three. Bregman et al. used peak magnitude to control sharpness, and a three times enlargement is equivalent to a decrease in peak magnitude from 24 to 8 dB. The current results indicate that degradation of spectral cues associated with broadened auditory tuning typical of cochlear hearing loss is sufficient to significantly disrupt streaming.

The real-world ability of HI individuals to understand speech in noisy environments likely involves a number of factors. Although audibility may be the primary concern, the processing of suprathreshold auditory signals is also not normal in these individuals. The loss of outer hair cell function causes a loss of compressive nonlinearity characteristic of NH and a corresponding abnormal growth of loudness (cf. Bacon, 2004). The loss of outer hair cell function is also responsible for broadened auditory tuning. Assuming that signals of interest are generally more spectrally limited than interfering noise, the classic power spectrum model of masking (Patterson and Moore, 1986), predicts that broad tuning will allow larger amounts of noise to enter a given auditory filter, thus reducing the signal to noise ratio at that frequency and disrupting performance.

The current results suggest another influence of broad tuning on auditory performance. It appears that smearing the spectral representation of sequentially-presented speech items reduces the ability to form separate auditory streams. This additional limitation associated with broadened tuning may add to other more well-established limitations to further limit the performance of HI listeners in noisy backgrounds.

Acknowledgements

This study was supported in part by a doctoral grant from the Région Rhône-Alpes (France), and supported by the audioprosthetic group Entendre. Additional support came from NIH/NIDCD Grant DC05795. The authors wish to thank J.P.L. Brokx for providing details on Nooteboom et al. (1978), Frédéric Berthommier, Christophe Micheyl and two anonymous reviewers for their helpful comments, and Mathieu Paquier and Samuel Garcia for technical assistance.

References

Arehart, K.H., King, C.A., McLean-Mudgett, K.S., 1997. Role of fundamental frequency differences in the perceptual separation of competing vowel sounds by listeners with normal hearing and listeners with hearing loss. J. Speech Lang. Hear. Res. 40, 1434–1444.

Bacon, S.P., 2004. Compression: From Cochlea to Cochlear Implants. Springer-Verlag, New York.

Baer, T., Moore, B.C.J., 1993. Effects of spectral smearing on the intelligibility of sentences in noise. J. Acoust. Soc. Am. 94 (3), 1229–1241.

Bregman, A.S., 1978. Auditory streaming is cumulative. J. Exp. Psychol. Hum. Percept. Perform. 4 (3), 380–387.

Bregman, A.S., 1990. Auditory Scene Analysis: The Perceptual Organization of Sound. The MIT Press, Massachusetts.

Bregman, A.S., Campbell, J., 1971. Primary auditory stream segregation and perception of order in rapid sequences of tones. J. Exp. Psychol. 89 (2), 244–249.

Bregman, A.S., Liao, C., Levitan, R., 1990. Auditory grouping based on fundamental frequency and formant peak frequency. Can. J. Psychol. 44 (3), 400–413.

Carlyon, R.P., Long, C.J., Deeks, J.M., McKay, C.M., 2007. Concurrent sound segregation in electric and acoustic hearing. J. Assoc. Res. Otolaryngol. 8 (1), 119–133.

Chatterjee, M., Sarampalis, A., Oba, S.I., 2006. Auditory stream segregation with cochlear implants: A preliminary report. Hear. Res. 222 (1–2), 100–107.

Cooper, H.R., Roberts, B., 2007. Auditory stream segregation of tone sequences in cochlear implant listeners. Hear. Res. 225 (1–2), 11–24.

Darwin, C.J., Bethell-Fox, C.E., 1977. Pitch continuity and speech source attribution. J. Exp. Psychol. Hum. Percept. Perform. 3 (4), 665–672.

de Boer, B., 2000. Self-organization in vowel systems. J. Phonet. 28 (4), 441–465.

de Cheveigné, A., 1999. Waveform interactions and the segregation of concurrent vowels. J. Acoust. Soc. Am. 106 (5), 2959–2972.

Dorman, M.F., Cutting, J.E., Raphael, L.J., 1975. Perception of temporal order in vowel sequences with and without formant transitions. J. Exp. Psychol. Hum. Percept. Perform. 104 (2), 147–153.

Grimault, N., Micheyl, C., Carlyon, R.P., Arthaud, P., Collet, L., 2000. Influence of peripheral resolvability on the perceptual segregation of harmonic tones differing in fundamental frequency. J. Acoust. Soc. Am. 108 (1), 263–271.

Grimault, N., Micheyl, C., Carlyon, R.P., Arthaud, P., Collet, L., 2001. Perceptual auditory stream segregation of sequences of complex sounds in subjects with normal and impaired hearing. Br. J. Audiol. 35 (3), 173–182.

Grose, J.H., Hall, J.W., 1996. Perceptual organization of sequential stimuli in listeners with cochlear hearing loss. J. Speech Hear. Res. 39 (6), 1149–1158.

Hartmann, W.M., Johnson, D., 1991. Stream segregation and peripheral channeling. Music Percept. 9 (2), 115–184.

Hirsh, I.J., 1959. Auditory perception of temporal order. J. Acoust. Soc. Am. 31 (6), 759–767.

Hong, R.S., Turner, C.W., 2006. Pure-tone auditory stream segregation and speech perception in noise in cochlear implant recipients. J. Acoust. Soc. Am. 120 (1), 360–374.

Klatt, D.H., 1980. Software for a cascade/parallel formant synthesizer. J. Acoust. Soc. Am. 67 (3), 971–995.

Lackner, J.R., Goldstein, L.M., 1974. Primary auditory stream segregation of repeated consonant–vowel sequences. J. Acoust. Soc. Am. 56 (5), 1651–1652.

Mackersie, C., Prida, T., Stiles, D., 2001. The role of sequential stream segregation and frequency selectivity in the perception of simultaneous sentences by listeners with sensorineural hearing loss. J. Speech Lang. Hear. Res. 44 (1), 19–28.

Moore, B.C.J., 1998. Cochlear Hearing Loss. Whurr, London.

Moore, B.C.J., Gockel, H., 2002. Factors influencing sequential stream segregation. Acta Acust. (Acustica) 88 (3), 320–332.

Nooteboom, S.G., Brokx, J.P.L., de Rooij, J.J., 1978. Contributions of prosody to speech perception. In: Levelt, W.J.M., d'Arcais, G.B.F. (Eds.), Studies in the Perception of Language. Wiley and Sons, New York, pp. 75–107.

Patterson, R.D., Moore, B.C.J., 1986. Auditory filters and excitation patterns as representations of frequency resolution. In: Moore, B.C. (Ed.), Frequency Selectivity in Hearing. Academic Press, London.

Patterson, R.D., Nimmo-Smith, I., Weber, D.L., Milroy, R., 1982. The deterioration of hearing with age: frequency selectivity, the critical ratio, the audiogram, and speech threshold. J. Acoust. Soc. Am. 72 (6), 1788–1803.

Qin, M.K., Oxenham, A.J., 2005. Effects of envelope-vocoder processing on f_0 discrimination and concurrent-vowel identification. Ear Hear. 26 (5), 451–460.

Remez, R.E., Rubin, P.E., Berns, S.M., Pardo, J.S., Lang, J.M., 1994. On the perceptual organization of speech. Psychol. Rev. 101 (1), 129–156.

Roberts, B., Glasberg, B.R., Moore, B.C.J., 2002. Primitive stream segregation of tone sequences without differences in fundamental frequency or passband. J. Acoust. Soc. Am. 112 (5), 2074–2085.

Rogers, C.F., Healy, E.W., Montgomery, A.A., 2006. Sensitivity to isolated and concurrent intensity and fundamental frequency increments by cochlear implant users under natural listening conditions. J. Acoust. Soc. Am. 119 (4), 2276–2287.

Rose, M.M., Moore, B.C.J., 1997. Perceptual grouping of tone sequences by normally hearing and hearing-impaired listeners. J. Acoust. Soc. Am. 102 (3), 1768–1778.

Rose, M.M., Moore, B.C.J., 2005. The relationship between stream segregation and frequency discrimination in normally hearing and hearing-impaired subjects. Hear. Res. 204 (1–2), 16–28.

Singh, P.G., 1987. Perceptual organization of complex-tone sequences: a tradeoff between pitch and timbre? J. Acoust. Soc. Am. 82 (3), 886–899.

Singh, P.G., Bregman, A.S., 1997. The influence of different timbre attributes on the perceptual segregation of complex-tone sequences. J. Acoust. Soc. Am. 102 (4), 1943–1952.

Stainsby, T.H., Moore, B.C.J., Glasberg, B.R., 2004a. Auditory streaming based on temporal structure in hearing-impaired listeners. Hear. Res. 192 (1–2), 119–130.

Stainsby, T.H., Moore, B.C.J., Medland, P.J., Glasberg, B.R., 2004b. Sequential streaming and effective level differences due to phase-spectrum manipulations. J. Acoust. Soc. Am. 115 (4), 1665–1673.

Studebaker, G.A., 1985. A "rationalized" arcsine transform. J. Speech Hear. Res. 28 (3), 455–462.

Summers, V., Leek, M., 1998. F_0 processing and the separation of competing speech signals by listeners with normal hearing and with hearing loss. J. Speech Lang. Hear. Res. 41, 1294–1306.

Tessier, E., 2001. Étude de la variabilité de l'indice de localisation pour la caractérisation de sources de parole interférentes. Ph.D. thesis, Institut National Polytechnique de Grenoble, France.

Thomas, I.B., Hill, P.B., Carroll, F.S., Garcia, B., 1970. Temporal order in the perception of vowels. J. Acoust. Soc. Am. 48 (4), 1010–1013.

van Noorden, L.P.A.S., 1975. Temporal coherence in the perception of tones sequences. Ph.D. thesis, Eindhoven University of Technology.

Vliegen, J., Moore, B.C.J., Oxenham, A.J., 1999. The role of spectral and periodicity cues in auditory stream segregation, measured using a temporal discrimination task. J. Acoust. Soc. Am. 106 (2), 938–945.

Vliegen, J., Oxenham, A.J., 1999. Sequential stream segregation in the absence of spectral cues. J. Acoust. Soc. Am. 105 (1), 339–346.

Warren, R.M., Obusek, C.J., Farmer, R.M., Warren, R.P., 1969. Auditory sequence: confusion of patterns other than speech or music. Science 164, 586–587.

Chapitre 3

Simulation d'implant cochléaire et indices temporels

Les indices spectraux de hauteur sont fortement affectés par une perte de sélectivité fréquentielle. En revanche, l'enveloppe temporelle est peu affectée par la résolution fréquentielle et sa périodicité est aussi un indice de hauteur permettant la ségrégation. Cette étude avait pour objectifs d'évaluer le rôle des indices temporels de hauteur pour la ségrégation séquentielle de voyelles, et d'estimer si ces indices pourraient être exploités par des implantés cochléaires pour ségréger des flux de voyelles. Nous avons utilisé une simulation d'implant pour réduire les indices spectraux de hauteur. Dans cette simulation, les canaux sont excités avec un bruit filtré. La structure fine temporelle ne contient donc pas d'information de hauteur, et les seuls indices temporels de hauteur sont donc portés par la périodicité de l'enveloppe temporelle.

Nos résultats ont montré que les sujets n'exploitaient pas ces indices temporels dans notre tâche de streaming irrépressible. En revanche, ils semblaient profiter d'un indice spectral lié à la hauteur. Cet indice résulte d'un effet combiné de la variation de la définition des formants avec la hauteur, et de la quantification fréquentielle qui est effectuée dans la simulation. Les implications pour les implantés réels et pour la séparation de voix concurrentes sont discutées.

Cet article a été accepté pour publication sous réserve de corrections mineures par le Journal of the Acoustical Society of America *le 18 janvier 2008.*

Streaming of vowel sequences based on fundamental frequency in a cochlear implant simulation[†]

Etienne Gaudrain and Nicolas Grimault[*]
Neurosciences Sensorielles, Comportement, Cognition, CNRS UMR 5020, Université Lyon 1, 50 avenue Tony Garnier, 69366 Lyon Cedex 07, France

Eric W. Healy
Speech Psychoacoustics Laboratory, Department of Communication Sciences and Disorders, University of South Carolina, Columbia, 29208

Jean-Christophe Béra
Inserm U556, Lyon, France

Abstract

Cochlear-implant users often have difficulties perceiving speech in noisy environments. Although this problem likely involves auditory scene analysis, few studies have examined sequential segregation in cochlear implant (CI) listening situations. The present study aims to assess the possible role of fundamental frequency (F_0) cues for the segregation of vowel sequences, using a cochlear implant simulation. Obligatory streaming was evaluated using an order-naming task in two experiments involving normal-hearing subjects. In the first experiment, it was found that streaming did not occur based on F_0 cues when natural-duration vowels were processed to reduce spectral cues using a noise-band vocoder. In the second experiment, briefer vowels were used to enhance streaming. Under these conditions, F_0-related streaming appeared even when vowels were processed to reduce spectral cues. However, the observed segregation did not appear to result from temporal periodicity cues. Instead, it appears to have resulted from remaining F_0-related spectral cues, despite the fact that the vocoded vowels did not elicit a strong pitch sensation. Thus, streaming under conditions of severely-reduced spec-

[†]Portions of this work were presented in "Segregation of vowel sequences having spectral cues reduced using a noise-band vocoder," Poster presented at the 151st ASA meeting in Providence, RI, USA, June 2006.

[*]Electronic mail : ngrimault@olfac.univ-lyon1.fr

tral cues, such as those associated with CIs may be expected to occur as a result of this cue.

PACS numbers : 43.66.Mk, 43.66.Sr, 43.71.Es, 43.71.Ky

3.1 Introduction

The mechanisms involved in auditory stream segregation have been thoroughly investigated in normal-hearing listeners (*e.g.*, Bregman and Campbell, 1971 ; van Noorden, 1975 ; Bregman, 1990). These studies led to the peripheral channeling theory (Hartmann and Johnson, 1991), which states that two stimuli need to excite different neural populations to produce auditory streaming. This theory and its implementations (Beauvois and Meddis, 1996 ; McCabe and Denham, 1997) assume that the main cues for streaming are spectral. Consequently, frequency selectivity is likely to be critical. Moore and Gockel (2002), in a review of studies involving sequential stream segregation, further concluded that any sufficiently salient perceptual difference may lead to stream segregation, regardless of whether or not it involves peripheral channeling (see also Elhilali and Shamma, 2007). Frequency selectivity can also affect the perceptual salience of cues, and difference limen (DL) measurements can be used to evaluate the salience of stimuli along a given perceptual dimension. Rose and Moore (2005) tested this hypothesis and found that the fission boundary (cf. van Noorden, 1975) was indeed proportional to the frequency DL for pure tones at frequencies between 250 and 2000 Hz. But, when sounds are complex and composed of many interacting features, it is hard to figure out how to define salience. Moreover, when sounds are degraded by the hearing system, such as in hearing-impaired (HI) listeners or in cochlear-implant (CI) users, it is hard to figure out which features of the initial sound still contribute to the salience. The current study aims to clarify the role of fundamental frequency for the segregation of noise-band vocoded vowel sequences.

Experiments involving normal-hearing (NH) listeners have shed light on the mechanisms involved in pitch-based streaming and the influence of reduced frequency resolution. Studies involving the resolvability of harmonic components of complex tones have shown that streaming based on fundamental frequency (F_0) is reduced when resolvability is reduced, but it is still possible to some extent even when harmonics are totally unresolved (Vliegen and Oxenham, 1999 ; Vliegen *et al.*, 1999 ; Grimault *et al.*, 2000). Gaudrain *et al.* (2007) found that F_0-based streaming of vowel sequences was reduced when frequency resolution was reduced by simulated broad auditory

filters (Baer and Moore, 1993). Roberts *et al.* (2002) showed that differences solely in temporal cues (obtained by manipulating the phase relationship between components) can elicit streaming. Grimault *et al.* (2002) also observed streaming based on the modulation rate of sinusoidally amplitude-modulated (SAM) noises, *i.e.* without any spectral cues to pitch. Despite the fact that the pitch elicited by the modulated noise was relatively weak, these authors observed streaming performance with SAM noise similar to that obtained with unresolved complex tones. Thus, streaming is reduced when spectral cues are reduced, but it is apparently possible to some extent when spectral cues are removed and only temporal cues remain.

These results have substantial implications for individuals having sensorineural hearing impairment and those fitted with a cochlear implant (CI). It is well known that these individuals have reduced access to spectral cues (*cf.* Moore, 1998). Moore and Peters (1992) observed F_0 DLs in hearing-impaired (HI) listeners approximately 2.5 times greater than those observed in normal-hearing (NH) listeners. Similarly, Rogers *et al.* (2006) found mean F_0 DLs 7.7 times greater in CI than in NH listeners. These results suggest that pitch differences are far less salient for HI and CI listeners, and that pitch-based streaming might be impaired for these individuals. However, psychoacoustic measures have indicated that temporal resolution is generally intact in the HI ear (for review, see Healy and Bacon, 2002 ; Moore, 1998). Cochlear implant users are also sensitive to the pitch elicited by the temporal repetition rate of electrical pulse trains, up to a limit of approximately 300 Hz (Shannon, 1983 ; Tong and Clark, 1985 ; Townshend *et al.*, 1987). However, their DLs for rate discrimination are larger than in NH (Zeng, 2002). Although these results together provide some insight into the availability of streaming cues in these individuals, relatively little is known.

Auditory streaming has been examined to a limited extent in HI listeners, with mixed results. Grose and Hall (1996) found that listeners with cochlear hearing loss generally required a greater frequency separation for segregation of pure tones. However, Rose and Moore (1997) reported no systematic difference between ears of unilaterally-impaired listeners in the frequency difference required to segregate pure tones. The correlation between auditory filter width and streaming performance with pure tones was also found to be not significant (Mackersie *et al.*, 2001). Grimault *et al.* (2001) found that streaming was hindered for HI listeners relative to NH, but only in conditions where components of complex tones were resolved for NH and unresolved for HI. Finally, Stainsby *et al.* (2004) examined streaming based on phase relationship differences and found results for elderly HI listeners that were similar to those observed in NH listeners.

A few studies have also attempted to examine streaming in CI users (Hong and Turner, 2006 ; Chatterjee *et al.*, 2006 ; Cooper and Roberts, 2007). For these users, different kinds of temporal cues can be related to pitch. Moore and Carlyon (2005) argued that the temporal fine structure of resolved harmonics was the most accurate pitch mechanism. However, when harmonics are unresolved, they interact in auditory filters and can encode pitch by amplitude modulation rate (*i.e.* by the temporal envelope periodicity). Because the spectrum is encoded by a reduced number of bands in CI processors, the harmonics in the range of human voice pitch (F_0 between 100 and 400 Hz) are almost always unresolved, and the fine structure of individual resolved harmonics is not available. On the contrary, the temporal envelope is roughly preserved in each band, so pitch may be coded by periodicity cues. In this paper, the term 'temporal pitch cues' will then refer to temporal envelope periodicity (in the range 100 - 400 Hz) ; in contrast to 'spectral pitch cues' which will refer to the pitch that is evoked by resolved harmonics (*i.e.* issued from the tonotopic analysis of the cochlea, and if relevant, the analysis of temporal fine structure). Because amplitude-modulated broadband noises can produce some impression of pitch (Burns and Viemeister, 1976, 1981) and can induce obligatory streaming (Grimault *et al.*, 2002), it might be possible for these temporal cues to induce streaming in CI users.

Hong and Turner (2006) used the rhythm task described in Roberts *et al.* (2002) to obtain an objective measure of obligatory streaming in NH and CI users. CI had from 16 to 22 electrodes. For a base frequency of 200 Hz, half the CI users performed as poorly as the NH listeners, whereas the other half performed better than normal, demonstrating less stream segregation. The authors showed that this variability correlated moderately but significantly with the ability to perceive speech in noise. Chatterjee *et al.* (2006) used pulse trains in ABA patterns and a subjective evaluation of whether the subjects (fitted with Nucleus-22) heard 1 or 2 streams. These authors observed response patterns that could be explained by streaming for both differences in spatial location (presentation electrode) and amplitude modulation rate. However, they did not observe the characteristic build-up of streaming over time (Bregman, 1978) for simple pulsitile stimuli that differed in location. This observation raises the possibility that the task involved discrimination rather than streaming. On the contrary, they did observe some build-up for the signals that differed in amplitude modulation rate, which suggests that AM-rate based streaming was indeed observed. Cooper and Roberts (2007) conducted an experiment that also involved pulsitile stimuli that differed in electrode location. They observed results that potentially reflected streaming, but a second experiment revealed that the results may have been attributable to pitch (or brightness) discrimination. Altogether, these studies provide only

modest evidence that streaming can occur in CI recipients on the basis of either place-pitch (*i.e.* electrode number) or temporal-pitch.

These previous results together suggest (1) that F_0-based streaming is affected by frequency selectivity, but (2) that streaming can be also induced by temporal pitch cues. It is also clear that (3) frequency selectivity is reduced in HI and CI listeners, but that (4) temporal pitch cues are preserved to some extent in these listeners. The question then becomes to what extent these cues can be utilized to elicit streaming.

Although streaming is often assumed to be a primitive mechanism, some correlation between streaming and higher level processing, such as concurrent speech segregation, has been reported (Mackersie *et al.*, 2001). However, the relation between streaming with pure or complex tones and speech segregation remains difficult to assess. In speech, pitch cues signaling that different talkers are present are mixed with other cues that may not be relevant for concurrent talker segregation. Listeners may then not be able to benefit from these cues in ecological situations. Only a few studies have reported streaming with speech materials (Dorman *et al.*, 1975 ; Nooteboom *et al.*, 1978 ; Tsuzaki *et al.*, 2007 ; Gaudrain *et al.*, 2007), and only the last one examined the effect of frequency selectivity impairment.

The current study is similar to that of Gaudrain *et al.* (2007). Whereas streaming under conditions of broad tuning in accord with sensorineural hearing impairment was examined in that study, streaming under conditions similar to CI stimulation was assessed in the current study. Specifically, the role of reduced spectral pitch cues in vowel sequence streaming was assessed in a first experiment, and the possible role of temporal pitch cues was investigated in a second experiment. Gaudrain *et al.* (2007) used an order task to measure obligatory streaming, *i.e.* streaming that is irrepressible, in opposition to voluntary streaming. This type of streaming does not require an important cognitive load, and is less dependent upon attention and subject strategy. Moreover, when measuring obligatory streaming, the subject actually tries to resist to the segregation, leading to better performances when streaming ability is impaired. Hence, this approach allows studying streaming with degraded stimuli as the performances cannot be attributable to the degradation of the perception of the individual items. In the order task, inspired by Dorman *et al.* (1975), the listener is presented a repeating short sequence of vowels with alternating F_0. The subject is asked to report the whole sequence of vowels. If the sequence splits in to streams, the listener is no longer able to perceive the order of the items between the two streams, and is then not able to give back the whole sequence properly.

3.2 Experiment 1 : Streaming of vowel sequences with quantized spectral cues

3.2.1 Rationale

Although reduced frequency selectivity might be responsible for a portion of the relatively poor performance of CI users in the perception of concurrent voices (Qin and Oxenham, 2005 ; Stickney *et al.*, 2004, 2007), a few studies (*e.g.* Grimault *et al.*, 2002 ; Chatterjee *et al.*, 2006) suggest that temporal pitch cues might be sufficient to elicit streaming in CI users. In the current experiment, an objective paradigm - the order task - was used to assess obligatory streaming induced by vowel sequences having alternating F_0. The current study employed noise-band vocoder models of CIs, to best control the cues available to these listeners. This simulation offers the advantages afforded by NH listeners hearing simulations of cues reduced in controlled manner, without the complications and confounds associated with testing the clinical population.

The reduction of the spectral resolution in the vocoder should reduce the amount of streaming. Consequently, the performances in the order task should be better in the CI simulation than with intact stimuli. But if temporal pitch cues can elicit obligatory streaming, an effect of F_0 separation should be observed.

3.2.2 Material and method

Subjects

Six subjects aged 22 to 30 years (mean 26.2) participated. All were native speakers of French and had pure tone audiometric thresholds below 20 dB HL at octave frequencies between 250 and 4000 Hz (ANSI, 2004). All were paid an hourly wage for participation. These subjects participated in one of the experiments of Gaudrain *et al.* (2007) and were therefore familiar with the paradigm.

Stimuli

Individual vowels were first recorded and processed, then arranged into sequences. The six French vowels /a e i ɔ y u/ were recorded (24 bits, 48000 Hz) using a Røde NT1 microphone, a Behringer Ultragain preamplifier, a Digigram VxPocket 440 soundcard, and a PC. The speaker was instructed to pronounce all 6 vowels at the same pitch and to reduce prosodic variations.

TAB. 3.1 – Cutoff frequencies of the 12- and 20-channel vocoders, from Dorman *et al.* (1998).

channel	Q_{20}	Q_{12}
	166	212
1	198	336
2	362	570
3	414	754
4	600	1056
5	676	1324
6	888	1718
7	992	2098
8	1236	2620
9	1376	3150
10	1658	3848
11	1840	4582
12	2168	5518
13	2400	
14	2786	
15	3078	
16	3534	
17	3900	
18	4438	
19	4894	
20	5532	

The F_0 and duration of each vowel was then manipulated using STRAIGHT (Kawahara *et al.*, 1999). Duration was set to 167 ms, approximating the median duration of syllables in spoken language (Greenberg *et al.*, 1996). Additional versions of each vowel were then prepared in which the average F_0 was 100, 110, 132, 162 and 240 Hz. Fundamental frequency variations related to intonation were constrained to be within 0.7 semitones of the average. Formant positions were held constant across F_0 conditions.

Finally, each vowel was subjected to two conditions of reduced spectral resolution. In Q_{20} the vowels were subjected to 20-bands noise vocoder, and in Q_{12} they were subjected to a 12-band vocoder. Q_∞ refers to the intact vowels. The implementation of the noise-band vocoder followed Dorman *et al.* (1997). The stimulus was first divided into frequency bands using 8th order Butterworth filters. The cutoff frequencies of these bands were the approximately logarithmic values used by Dorman *et al.* (1998) and are listed in Table 3.1. The envelope of each band was extracted using half-wave rectification and 8th order Butterworth lowpass filtering with cutoff frequency of 400 Hz. This lowpass value ensured that temporal pitch cues associated with voicing were preserved. The resulting envelopes were used to modulate white noises using sample point-by-point multiplication, which were then filtered to restrict them to the spectral band of origin. The 12 or 20 bands comprising a condition were then mixed to construct the vocoder. A 10 ms cosine gate was finally applied to each vowel in each condition.

The vowels were then concatenated to form sequences. Fig. 3.1 describes the arrangement of vowels into sequences and the construction of the various conditions. Each sequence contained one presentation of each vowel.

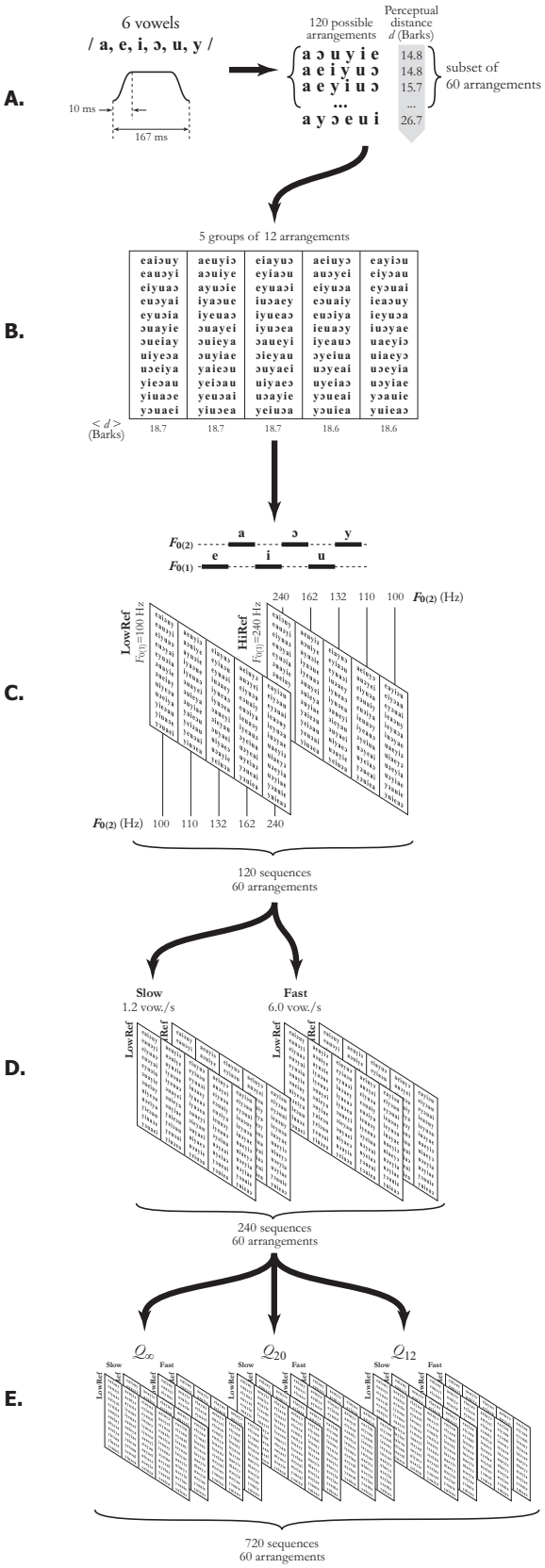

FIG. 3.1 – The arrangement of conditions : *Panel A*. Individual vowels were recorded and modified using STRAIGHT. They were arranged into sequences (120 possible orders), and the 60 arrangements having the lowest perceptual distances d were selected (see text for details). *Panel B*. These 60 arrangements were divided among five groups with similar average perceptual distance (12 in each group). *Panel C*. Each group was attributed a fundamental frequency difference in both LowRef and HiRef conditions (yielding 120 sequences). *Panel D*. These 120 sequences appeared in both Slow and Fast conditions (yielding 240 sequences). *Panel E*. Finally, these 240 sequences appeared in each Q condition (yielding 720 sequences). These 720 sequences were presented across six presentation blocks, such that each condition was equally represented within each block.

Sequences containing all possible arrangements of the six repeating vowels ($[n-1]! = 120$) were first generated, then the 60 arrangements having the smallest differences in formant structure were selected for inclusion (Fig. 3.1, Panel A). The selection of these arrangements having the smallest *perceptual formant differences*[1] was performed to reduce the influence of streaming based on differences in formant structure between successive vowels in a sequence (Gaudrain *et al.*, 2007). These 60 arrangements were then divided into five groups of 12 arrangements each, such that the average perceptual distance of each group was approximately equal across groups (Fig. 3.1, Panel B).

The F_0 of the vowels in a sequence alternated between two values $F_{0(1)}$ and $F_{0(2)}$. In condition LowRef, the value of $F_{0(1)}$ was 100 Hz and $F_{0(2)}$ was one of the five F_0 values (100, 110, 132, 162, 240). In condition HighRef, $F_{0(1)}$ was 240 Hz and $F_{0(2)}$ was one of the five F_0 values. Thus, there were five F_0 differences. Each group of 12 arrangements was then assigned to one of the five F_0 difference. The appearance of the 60 arrangements in both the LowRef and HiRef conditions yielded 120 sequences (Fig. 3.1, Panel C). These 120 sequences appeared in both the Slow and Fast conditions, yielding 240 sequences (Fig. 3.1, Panel D). In addition, each sequence was created with two presentation rates : Slow at 1.2 vowel/s, and Fast at 6 vowel/s. The duration of the vowel remained the same is these two conditions, and silence was added between the vowels in the Slow condition. Slow sequences were used to check vowel identification performance. Fast sequences were used to observe streaming. The Slow sequences were repeated four times and the fast sequences were repeated 20 times, for overall stimulus durations of 20 s. Finally, each of these 120 sequences appeared in each of the three Q conditions, yielding a total of 720 sequences (Fig. 3.1, Panel E).

Stimuli were generated at 16 bits and 44.1 kHz using MATLAB. They were presented using the Digigram VxPocket 440 soundcard, and Sennheiser HD250 Linear II headphones diotically at 85 dB SPL, as measured in an artificial ear (Larson Davis AEC101 and 824 ; ANSI, 1995).

Procedure

Training and selection Two training tasks preceded testing. The first involved simple identification of single vowels. Subjects heard blocks containing

[1]Perceptual distance between vowels was calculated as the weighted Euclidian distance in the F_1–F_2' space, where F_1 is the frequency of first formant and F_2' is frequency of the effective second formant defined as the weighted sum of the second to fourth formants (de Boer, 2003). Perceptual distance for a token is then the sum of the perceptual distances between consecutive vowels.

each vowel at each F_0 twice. They responded using a mouse and computer screen, and visual feedback was provided after each response. This test was repeated, separately for each Q condition, until a score of 98% (59/60) was obtained. On average, proficiency was reached after one block for the Q_∞ vowels, 1.3 blocks for the Q_{20} vowels, and 2.8 blocks for the Q_{12} vowels.

The second training task involved vowel identification using the Slow sequences. In each block, 60 sequences were presented representing all 30 conditions (5 F_0's $\times$ 2 LowRef/HiRef $\times$ 3 Q conditions). The procedure was the same as the test procedure described below, except that visual feedback was provided. To proceed to the test, subjects were required to obtain a score, averaged over two consecutive blocks, greater than 95% in each Q condition. On average, 6.3 blocks were necessary to reach the proficiency criterion. Although intended to be a selection criterion, no subject was eliminated at this step.

Streaming test The subject was presented with a repeating sequence. After an initial 5 s period during which streaming was allowed to stabilize, the subject was asked to report the order of appearance of the constituent vowels. They were allowed to start with any vowel. The response was entered using a computer graphic interface and mouse. The next sequence was presented after the subject confirmed their response or after a maximum of 20 s. The 720 sequences were distributed among six presentation blocks, such that each condition was represented in each block an equal number of times. The average duration of the blocks was approximately 28 min. The experimental procedure was formally approved by a local ethics committee (CCPPRB Léon Bérard).

3.2.3 Results

For each condition, the score is the percentage of responses in which the 6 vowels comprising a sequence were reported in the correct order. Mean scores across subjects are plotted as a function of $F_{0(2)}$ in Fig. 3.2. Chance performance is 0.8%. As in Gaudrain *et al.* (2007), high scores can be interpreted as a tendency toward integration across $F_{0(1)}$ and $F_{0(2)}$ items and a resistance to obligatory streaming. Separate analyses were first conducted on the LowRef and HiRef conditions because the F_0 differences were not the same in the two conditions. The results in the Slow condition (1.2 vowel/s) showed that identification was near perfect in all conditions except in the HiRef, Q_{12} condition, at $F_{0(2)}$=162 Hz. Scores in this condition were therefore not included in the statistical analyses. An analysis of errors in the Slow, HiRef, Q_{12}, $F_{0(2)}$=162 Hz condition showed confusions between /y/ and /e/ in 8/9

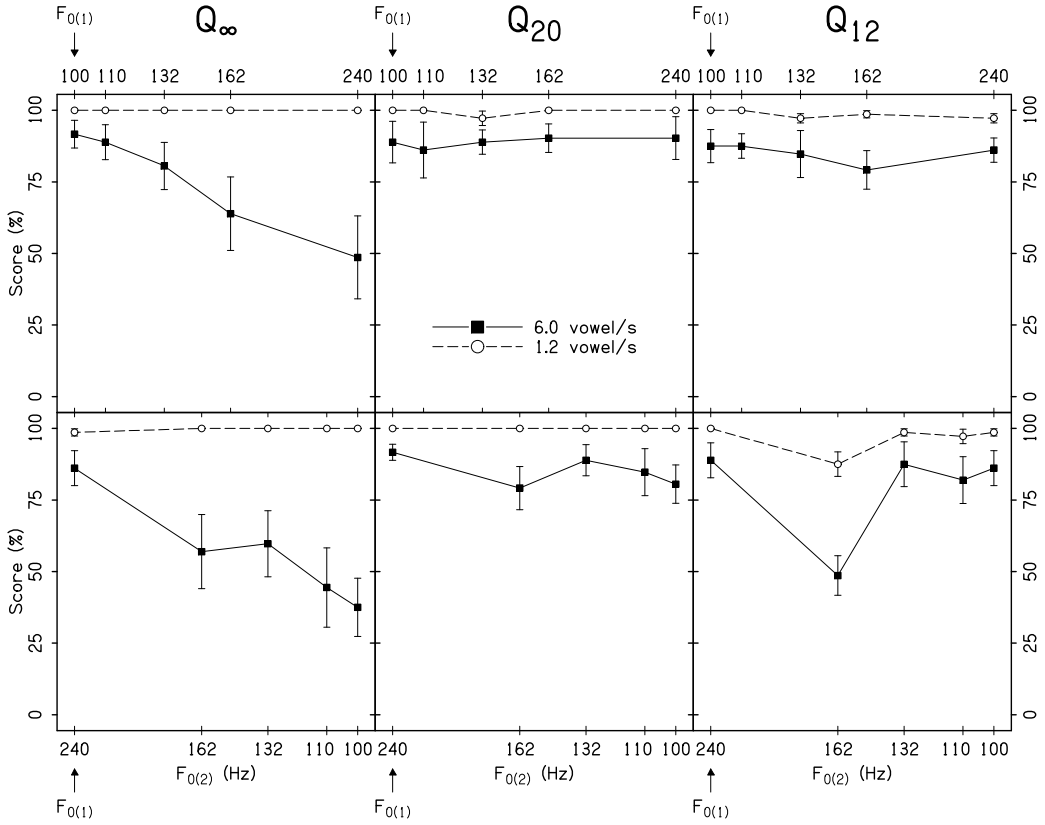

FIG. 3.2 – Shown are group means and standard deviations for a task in which subjects reported the order of six vowels appearing in sequence. Thus, low scores represent a tendency toward segregation. Alternate vowels were at alternate F_0 values ($F_{0(1)}$ and $F_{0(2)}$). In the LowRef conditions, $F_{0(1)}$ was fixed at 100 Hz and in HiRef $F_{0(1)}$ was fixed at 240 Hz. Conditions Q_{20} and Q_{12} involved reduced frequency resolution via 20 and 12-channel noise vocoders. Filled squares represent a Fast condition (6.0 vowel/s) in which streaming can occur, and open circles represent a Slow condition (1.2 vowel/s) that ensures accurate item identification. The abscissa is logarithmic.

false responses. These two vowels therefore seem difficult to discriminate at this particular combination of F_0's and vocoder channel divisions.

A two-way ANOVA on the LowRef data using Q condition and $F_{0(2)}$ as repeated parameters indicated that the effects of both Q condition [$F(2, 10) = 4.14, p < 0.05$] and $F_{0(2)}$ [$F(4, 20) = 12.21, p < 0.001$] were significant, and interacted significantly [$F(8, 40) = 3.93, p < 0.01$]. Separate one-way ANOVAs on each Q condition using $F_{0(2)}$ as a repeated factor showed a significant effect of $F_{0(2)}$ in the Q_∞ condition [$F(4, 20) = 9.28$, $p < 0.001$], but not in Q_{12} [$F(4, 20) = 0.65$, $p = 0.63$] or Q_{20} conditions [$F(4, 20) = 0.21$, $p = 0.93$].

A two-way ANOVA on the HiRef data using Q condition and $F_{0(2)}$ as repeated parameters indicated that both Q condition [$F(2, 10) = 9.02$, $p < 0.01$] and $F_{0(2)}$ [$F(4, 20) = 14.30$, $p < 0.001$] were significant, and interacted significantly [$F(8, 40) = 6.74$, $p < 0.001$]. Separate one-way ANOVAs on each Q condition using $F_{0(2)}$ as a repeated factor showed significant effects of $F_{0(2)}$ in the Q_∞ condition [$F(4, 20) = 10.15$, $p < 0.001$] and in the Q_{12} condition [$F(4, 20) = 14.96$, $p < 0.001$], but not in the Q_{20} condition [$F(4, 20) = 1.50$, $p = 0.24$]. A post-hoc analysis using pairwise t-tests showed that the effect of $F_{0(2)}$ in the Q_{12} condition was due solely to the point $F_{0(2)}$=162 Hz. When the HiRef condition was analyzed with this condition excluded, the pattern of significance was identical to that observed in the LowRef conditions : A significant effect of $F_{0(2)}$ in the Q_∞ condition [$F(3, 15) = 12.24$, $p < 0.001$], but not in the Q_{12} [$F(3, 15) = 0.74$, $p = 0.54$] or Q_{20} conditions [$F(3, 15) = 1.13$, $p = 0.37$].

3.2.4 Discussion

The results in the natural speech condition (Q_∞) are consistent with those observed by Gaudrain *et al.* (2007) in their first experiment. The greater the F_0 difference, the lower the scores, signifying greater streaming. Streaming based on F_0 difference is considered to be obligatory here because the task employed required that streaming be suppressed in order to perform accurately. Although the pattern of results in the current experiment is similar to that obtained by Gaudrain *et al.* (2007), the baseline level of performance differs. Scores in the Q_∞ condition at matched F_0 were over 80% here and approximately 50% in Experiment 1 of Gaudrain *et al.* (2007). One possible reason is that participants in the current experiment were well trained. In addition, there were subtle differences in the stimuli used in the two studies. Gaudrain *et al.* (2007) attributed low scores in the matched F_0 condition to formant-based streaming. Such a phenomenon has been reported by Dorman *et al.* (1975) with synthesized vowels. Formant-based streaming might be reduced with the recorded vowels used in the current experiment, where small F_0 fluctuations were preserved. F_0 fluctuations might serve to strengthen the grouping of components comprising individual vowels and limit the grouping of formants across successive vowels, as suggested by Gestalt theory (Bregman, 1990).

In the conditions having spectral degradation (Q_{20} and Q_{12}), the scores are high and do not depend on F_0. Thus, when spectral cues to pitch were reduced in accord with a CI model, F_0-based streaming was reduced or eliminated. Further, these results indicate that the temporal cues to pitch that remained in the vocoded stimuli were not strong enough, in this case, to elicit

obligatory streaming. It is potentially interesting to note that these results cannot be explained by a loss of intelligibility since degradation of the stimuli yielded an increase in performance. In addition, intelligibility was confirmed during training and in the Slow conditions.

The main finding of this experiment was that no F_0-based streaming appeared when spectral pitch cues were degraded using a model of a 12- or 20-channel CI. This result suggests that obligatory streaming is reduced when spectral cues to pitch are reduced in this manner and may not be possible for vowel stimuli based on remaining temporal pitch cues. This observation is in apparent contrast with studies that observed some streaming in CI recipients (Chatterjee *et al.*, 2006 ; Hong and Turner, 2006 ; Cooper and Roberts, 2007). One possibility is that the noise-band vocoder does not appropriately models some aspects of CI simulation. This point is addressed in the General Discussion. Another possibility is that the paradigm employed here is less sensitive than those employed previously, including the rhythm discrimination task used by Hong and Tuner (2006). It is potentially important that obligatory streaming is strongly influenced by presentation-rate (van Noorden, 1975), and that Hong and Turner (2006), Chatterjee *et al.* (2006) and Grimault *et al.* (2002) all used sequences with higher presentation rates (10 stimuli per second) to observe streaming based primarily or exclusively on temporal pitch cues. It seems then plausible that the natural presentation rate used in Experiment 1 was not sufficiently fast to elicit obligatory streaming under degraded conditions, but that streaming may still be possible. The next experiment assesses this hypothesis.

3.3 Experiment 2 : Assessing the role of temporal cues

3.3.1 Rationale

As shown by van Noorden (1975), the temporal coherence boundary, the threshold corresponding to obligatory streaming, depends on presentation rate. As shown in Experiment 1 of Gaudrain *et al.* (2007), higher presentation rates in the current paradigm do indeed lead to stronger measures of streaming. Thus, increasing the repetition rate of the sequences used in Experiment 1 should strengthen the streaming effect, and reveal if segregation is possible under the current conditions of severely reduced spectral cues, but intact temporal cues to pitch. In addition, two envelope cutoff values were employed to more closely examine the role of temporal cues.

3.3.2 Material and method

Subjects

Nine subjects aged 18 to 27 years (mean 21.9) participated. All were native speakers of French and had normal hearing as defined in Experiment 1. None of these subjects participated in previous similar experiments, and all were paid an hourly wage for participation.

Stimuli

The same six recorded vowels employed in Experiment 1 were used. The durations of vowels were reduced to 133 ms using STRAIGHT. Again, 10 ms ramps were employed. The average F_0 of each vowel was set to 100, 155 and 240 Hz using the same method used in Experiment 1. Five conditions of spectral degradation were then created. The intact vowels were used for a Q_∞ condition. The same noise-band vocoder used in Experiment 1 was again used to process vowels for Q_{20} and Q_{12} conditions. However, unlike Experiment 1, two cutoff frequencies (f_c) were used for envelope extraction. A value of f_c=400 Hz, was employed to preserve temporal pitch cues, and a value of f_c=50 Hz, was employed to eliminate temporal pitch cues. As in Experiment 1, envelope extraction involved half-wave rectification and 8th order Butterworth lowpass filtering.

The processed vowels were then concatenated to form sequences in the five processing conditions (Q_∞, Q_{20} f_c=400 Hz, Q_{20} f_c=50 Hz, Q_{12} f_c=400 Hz, and Q_{12} f_c=50 Hz). The 36 arrangements having the lowest perceptual distance were selected and divided into three groups having approximately equal mean perceptual distance values. As in Experiment 1, these three groups were used for the three F_0 separation, conditions. Thus, as in Experiment 1, the particular arrangements of vowels were distributed across the F_0 separation conditions, but repeated across the other conditions to ensure that effects associated with the particular order of items were constant.

In this experiment, only the LowRef condition was used, so that $F_{0(1)}$ was always 100 Hz. As in the first experiment, Slow (1.2 vowel/s) and Fast (7.5 vowel/s) sequences were employed. Slow sequences were repeated 4 times, and Fast sequences were repeated 25 times so that the overall duration in both conditions was 20 s. Stimuli were generated with MATLAB as 16 bit, 44.1 kHz sound files, and were presented as in Experiment 1.

Procedure

Training and selection Training again began with simple identification task of single vowels. Five blocks of 72 vowels were presented. Each block contained each vowel at each $F_{0(2)}$ in a single degradation condition, in random order (4 repetitions $\times$ 6 vowels $\times$ 3 F_0s = 72 items). The blocks were presented from the least degraded (Q_∞) to the most degraded (Q_{12} f_c=50 Hz). Visual feedback was provided. Each block was repeated until a score of 96% (69/72) was obtained or a maximum of three repetitions was reached. On average the blocks were repeated 1.7 times for Q_∞, 2.1 times for Q_{20} f_c=400 Hz, 2 times for Q_{20} f_c=50 Hz, 2.3 times for Q_{12} f_c=400 Hz, and 2 times for Q_{12} f_c=50 Hz.

As in Experiment 1, training involving the Slow condition sequences followed. The test consisted of seven blocks. Each block was composed of 36 sequences and all the conditions were represented at least twice in random order. Subjects were required to score greater than 95% correct over three successive blocks in each condition to advance to the next stage. One subject was unable to reach the criterion and was dismissed. For seven of the remaining participants, five blocks were sufficient to reach the criterion. The last subject reached the criterion after seven blocks.

Streaming test The test consisted of 5 blocks of 72 sequences each. All conditions (3 $F_{0(2)}$s, 5 Qs, 2 Slow/Fast) were as much as possible equally represented in all blocks. For each $F_{0(2)}$, streaming was measured over 12 different arrangements of vowels. Slow and Fast sequences were used, respectively, to check identification and to measure obligatory streaming. Other aspects of the experiment, including the initial 5 s response lockout and the manner of responding were identical to those Experiment 1. The experimental procedure was formally approved by a local ethics committee (CPP Sud Est II).

3.3.3 Results

Results averaged across the 8 subjects are plotted in Fig. 3.3. As can be seen, scores were uniformly high in the Slow conditions (mean : 98.6% correct), reflecting accurate identification of the constituent items. In these Fast conditions, scores were considerably lower than in the corresponding conditions of Experiment 1.

A two-way ANOVA using processing condition (Q_∞, Q_{20} f_c=400 Hz, Q_{20} f_c=50 Hz, Q_{12} f_c=400 Hz, and Q_{12} f_c=50 Hz) and F_0 as repeated parameters, showed a significant effect of processing condition [$F(4, 28) = 23.57$,

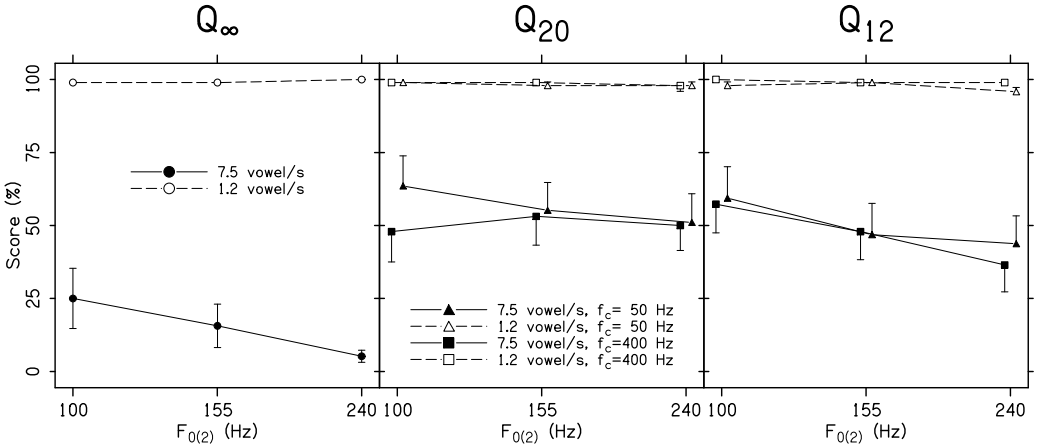

FIG. 3.3 – Shown are group means and standard deviations for the vowel order identification task in Experiment 2. Again, low scores represent a tendency toward segregation. $F_{0(1)}$ was at 100 Hz and $F_{0(2)}$ is shown. In each Q condition, scores for the Slow conditions (1.2 vowel/s) are plotted as open symbols, and score in the Fast conditions (7.5 vowel/s) are plotted as filled symbols. In the conditions Q_{20} and Q_{12} (second and third panels), scores for conditions in which temporal smoothing (f_c) was 50 Hz are plotted with triangles, and scores for f_c=400 Hz are plotted with squares The abscissa is logarithmic.

$p < 0.001$] as well as F_0 [$F(2, 14) = 8.25$, $p < 0.01$]. These factors did not interact [$F(8, 56) = 1.31$, $p = 0.26$], so the effect of F_0 was not significantly different between processing conditions. Multiple pairwise comparisons with Bonferroni correction showed that the processing condition effect originated from condition Q_∞ differing from the other conditions [$p < 0.001$]. Planned contrasts did not revealed any significant influence of f_c [for Q_{20}, $F(1, 7) = 3.86$, $p = 0.09$; for Q_{12}, $F(1, 7) = 0.90$, $p = 0.37$].

3.3.4 Discussion

The scores in the matched F_0 conditions were lower in the current experiment than in Experiment 1. This result can likely be attributed to at least two sources. The first is that naïve listeners in the current experiment produced lower scores than the trained listeners in Experiment 1. The second is that increasing the presentation rate enhances segregation based on formant structure. Gaudrain *et al.* (2007) argued that vowels having matched F_0 can elicit streaming based on formant structure, as found by Dorman *et al.* (1975). The higher scores in the Q_{20} and Q_{12} conditions support this hypothesis suggesting, as in Gaudrain *et al.* (2007), that formant based streaming is hindered by loss of frequency resolution.

Despite the fact that scores were reduced in the $F_{0(2)} = 100$ Hz conditions, a significant effect of $F_{0(2)}$ was nevertheless observed. This further suggests that streaming was observed for these sequences, and that this streaming interfered with the ability to accurately identify the order of constituent items in the sequences. The absence of an interaction with processing condition indicates that with this faster presentation rate, the vowel sequences elicited pitch-based streaming even in the degraded conditions.

However, the presence (f_c=400 Hz) or absence (f_c=50 Hz) of temporal pitch cues did not influence performance. The use of an 8th order smoothing filter during envelope extraction ensured that modulation frequencies above the temporal cutoff were not present at meaningful levels (Healy and Steinbach, 2007). This result suggests that streaming was not based on temporal pitch cues. Thus, other reasons for pitch-based streaming must be sought.

The main effect of F_0 suggests that streaming was based on a pitch-related cue, and the lack of an interaction suggests that this cue was present in all Q conditions. The first F_0 related cue that can be considered is the modulation product. The amplitude modulation in the channels of the vocoder may be highly periodic. Because of basilar membrane nonlinearity, this modulation could evoke a component at this modulation frequency and harmonic multiples. If the envelope modulation periodicity represents the F_0, the modulation product can recreate the original first harmonic. At $F_{0(2)}$=100 Hz, this modulation product falls below the lower frequency bound of the first band of the vocoder and would not be masked by that lowest noise band. It could then be heard and provide a pitch-related cue for streaming. However, when f_c=50 Hz, pitch cues associated with F_0 are suppressed from the envelope. Thus if amplitude modulation had elicited streaming, a difference should have been observe between the two f_c conditions. Thus, modulation products may be evoked with the present stimuli, but they are likely not responsible for the observed pitch-based streaming.

It must be then considered that some remaining spectral cues could be related to the F_0. The first such cue involves the first harmonic. The first harmonic can fall either in the first band of the vocoder or below it. It could then possibly provide a pitch related cue since it can influence the level of the first band. To test this hypothesis, the level in the first channel was measured in the different conditions. A three-way ANOVA on this measure using F_0, Q and f_c as repeated parameters, across the 6 vowels, showed no effect of F_0 [$F(2, 10) = 0.03$, $p = 0.97$] nor f_c [$F(1, 5) = 0.84$, $p = 0.40$], but a significant effect of Q [$F(1, 5) = 449.3$, $p < 0.001$]. Thus, the first channel does not appear to contain a consistent pitch-related cue. However, the distribution of harmonics across channels could have an effect when considering all the channels. A Fisher Discriminant Analysis (FDA) was used to find a linear

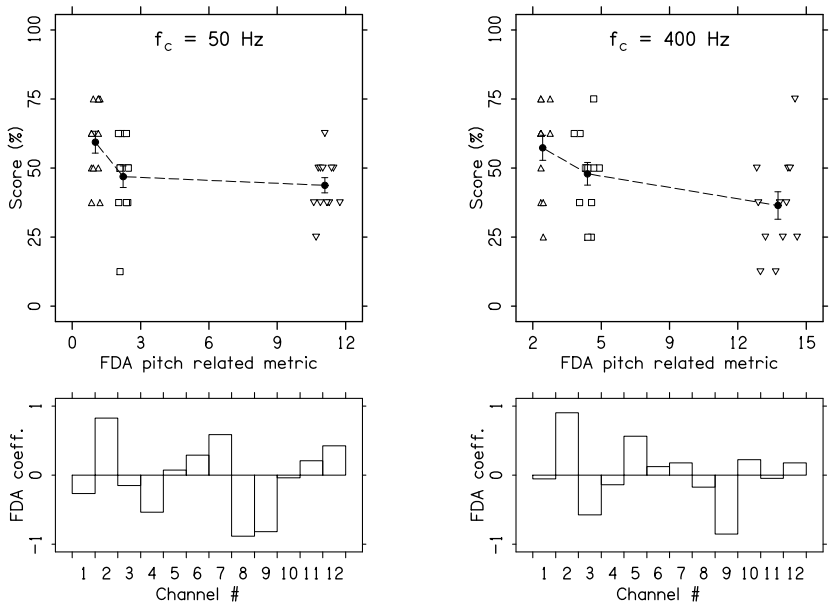

FIG. 3.4 – *Left panels* : Scores as a function of the pitch related metric found by FDA on vocoder channel mean levels for vowels processed in the Q_{12} condition, f_c=50 Hz. The upward triangles represents sequences with $F_{0(2)}$=100 Hz, squares for $F_{0(2)}$=155 Hz, and downward triangles for $F_{0(2)}$=240 Hz. The coefficients of the linear combination used as a metric are represented in the lower subpanel. *Right panels* : Same representation for the Q_{12} condition, f_c=400 Hz.

combination of channel levels that provides the best representation of the F_0 in the Q_{12} condition[2] (using Python MDP, Berkes and Zito, 2007). As shown in Fig. 3.4, for both f_c=400 Hz and 50 Hz, a linear combination of channel levels can be found to represent the F_0. This linear combination can be used as a metric that represents F_0. In Fig. 3, scores were plotted against this metric to show the relation between this metric and segregation. In conclusion, it is possible that some spectral cues related to the F_0, but that do not elicit a strong pitch sensation, could persist even in absence of harmonics. The coefficients found with the FDA suggest that channels #2, #8 and #9 were most likely carry these F_0 related cues.

Although the evidence for remaining F_0 related spectral cues is relatively clear, it remains unclear whether these cues would be present in a real CI. It is then important to determine the origin of these cues. In this experiment,

[2]It was not possible to analyze the Q_{20} condition, because FDA requires the number of items (6 vowels × 3 F_0s = 18) to be greater than the number of parameters (20 vocoder bands).

two successive manipulations were applied to the vowels. They were first transformed using the STRAIGHT algorithm, and then processed with a noise-band vocoder. In this sequence of processes, the F_0 cues are originated in the first stage, namely in STRAIGHT. Interestingly, it means that they are preserved or emphasized by the noise-band vocoder.

The vocoder discards the harmonic structure of the vowels. Hence, if a spectral cue is preserved by the vocoder, it should be part of the spectral envelope which is partially preserved. To determine if the F_0 manipulation by STRAIGHT induces F_0 related spectral cues that could be preserved in the vocoder, the spectral envelopes of the natural vowels (Q_∞) were reanalyzed using STRAIGHT again, for the 3 different F_0s. For each F_0, the average spectral envelope across vowels was computed, as well as the standard error. It was found that the greatest difference between averaged spectral envelopes at different F_0 was greater than the smallest standard error (across vowels) in only 3 frequency regions : below 470 Hz, around 2500 Hz, around 3250 Hz. These frequency regions correspond to channels #1, #2, #8 and #10 in the Q_{12} condition. It is then probable that the channels found by the FDA as describing the F_0 are actually reflecting these cues produced by the F_0 manipulation. The vowel /i/ seems to produce most of the observable difference between F_0 conditions. Further study of the spectrum of this vowel, and its spectral envelope as extracted by STRAIGHT, revealed that these cues are due to the position of formants relative to the position of harmonics. In the /i/ used in the current experiment, the frequency of the second formant was 2273 Hz. The closest harmonics for F_0=100 Hz are at 2198, 2295 and 2393 Hz. For F_0=240 Hz, the closest harmonics are at 2153 and 2393 Hz, missing the position of the peak. The peak corresponding to this formant is not shifted in frequency, but its magnitude is slightly lower at F_0=240 Hz than at F_0=100 Hz.

This analysis of the stimuli strongly suggests that the F_0 related spectral cues are due to modification of formant definition associated with changes in F_0. This phenomenon happens when perceiving natural vowels uttered at different F_0s. The frequency quantization in the noise-band vocoder emphasized this effect, as well as it makes it relatively stochastic across vowels. Though both these phenomena can occur with real speech and real CIs, it is difficult to determine whether this cue can occur in ecological situations. However, in the current experiment, the formant positions were held constant while changing F_0. In real speech, the formant position is related to F_0. This F_0-related cue may then appear along with some timbre cues such as vocal tract length (VTL) which can also elicit streaming (Tsuzaki *et al.*, 2007). Although streaming in NH listeners seems to be less influenced by changes in VTL than in F_0 (Tsuzaki *et al.*, 2007), this kind of cue could be preserved in

CI listening. Further investigation is then required to determine the potential role of these cues in multi-talker environments.

3.4 General discussion

The main result of this study is that obligatory streaming was reduced, but still present with spectral cues reduced in accord with a CI model, but that this streaming was not attributable to temporal pitch cues. This is consistent with Gaudrain *et al.* (2007) who observed that impoverishing the spectral representation of vowels via simulated broadened auditory tuning hindered pitch-based streaming. Reducing the spectral cues and preserving most of the temporal cues in Experiment 1 reduced sufficiently the salience of the two different streams to eliminate obligatory streaming. In the second experiment, emphasizing streaming through a higher presentation rate did not lead to the observation of streaming based on temporal cues, but instead led to the observation of streaming based on a stochastic spectral cue.

The current results are also consistent with previous observations of concurrent speech perception in NH listeners hearing CI simulations. Normal-hearing listeners are able to take advantage of a pitch difference between a target and masker to enhance speech recognition in noise (Brokx and Nooteboom, 1982). In contrast, Stickney *et al.* (2004) found that NH listeners were not able to take advantage of pitch differences between speech target and masker when the stimuli were processed with a noise-band vocoder. The absence of pitch-based streaming observed in the first experiment suggests that impairment of this mechanism could be partially responsible for the low speech perception performance in noise under CI simulation.

However, this result contrasts with those previously obtained in CI recipients using simple stimuli (Hong and Turner, 2006 ; Chatterjee *et al.*, 2006). One explanation is that the complexity of spectral and temporal cues associated with even simple speech items such as vowels diminishes the availability of F_0 cues following CI processing. Another possible reason is that the duration of the stimuli employed here was different from that used in previous studies. Hong and Turner (2006) and Chatterjee *et al.* (2006) used 60 and 50 ms stimuli, while the briefest vowels in the current study were 133 ms. Presentation rate strongly influences streaming, and a slower presentation rate can prevent streaming. However, to evaluate if obligatory streaming can possibly be involved in concurrent speech segregation in ecological situations, it is important to examine natural speech rates. Because streaming was not observed with vowels matching the median duration of English syllables (167 ms) in Experiment 1 of the current study, it can be concluded that CI users

may have difficulty taking advantage of streaming cues in cocktail party situations.

The comparison with CI user performances remains difficult because the noise-band vocoder simulation with NH listeners does not exactly mimic the perception of sounds by CI users. There is a long list of patient variables that are associated with CI users, but absent from consideration when using simulations. For instance, although the number of physical channels in modern cochlear implants is 16-22, the number of actual auditory channels depends on a variety of patient variables and can be far lower. Similarly, a typical processor bandwidth is approximately 100 to approximately 10,000 Hz, but varies across patients. Also, CI recipients have generally experienced their CI for months prior to experimentation while NH participants in the current experiment were trained on noise-band vocoders for a few hours. Notably, the noise-band vocoder does not exactly mimic stimulation by the CI. The main difference is that the output of the vocoder is acoustic and then subjected to peripheral processing by the ear, while the output of the CI involves electric pulses. Laneau *et al.* (2006) compared pitch discrimination in CI users (fitted with Nucleus-24) and in NH subjects using noise-band vocoders. They observed that NH subjects hearing vocoders were less sensitive to temporal pitch differences than CI users. Laneau *et al.* (2006) suggested that the main factor for this poor sensitivity in NH was the absence of compression/expansion of the temporal envelope in the vocoder. Acoustic stimuli, such as that produced by the vocoder, are encoded via a compressive function in the cochlea. In contrast, electric stimuli, such as those produced by the CI, are encoded via an expansive function (Zeng and Shannon, 1994 ; Laneau *et al.*, 2006). Laneau *et al.* (2006) suggest that the amplitude modulation depth might then be greater for CI users than for NH subjects hearing a vocoder.

Another effect that potentially reduces the strength of temporal pitch in the vocoder has been suggested by Hanna (1992). In the noise-band vocoder, the analysis filter bank and the resynthesis filter bank are typically the same. Thus, the noise carrier after modulation with the temporal envelope in each band is then filtered again to suppress the sidebands, *i.e.* the modulation products that fall outside the band. Modulation depth is reduced by this resynthesis filtering for the narrowest bands, and temporal pitch cues are then weakened. To fully preserve the amplitude modulation, the bandwidth must be greater than twice the F_0. Using this metric, temporal pitch cues are then intact in the current vocoder channels beyond #6 in the Q_{12} condition, and beyond #14 in the Q_{20} condition. In addition, as described by Hanna (1992), the peripheral filters of the normal ear play the same role as the resynthesis filters of the vocoder, and weaken again the modulation depth. Thus,

although subjects were probably able to perceive temporal pitch cues, their depth was reduced in the lower bands of the vocoder. In CI processors, such peripheral filtering does not occur and temporal pitch cues are not degraded in this way. Indeed, Laneau *et al.* (2006) observed that, using Butterworth filters, CI users had better F0 discrimination abilities than normal-hearing listeners with noise-band vocoder.

In the current study, it has been observed that temporal pitch cues were not sufficient to produce obligatory streaming of vowel sequences. However it cannot be entirely ruled out that pitch-based obligatory streaming can occur in CI users. In particular, it was found that an F_0 related spectral cue induced obligatory streaming. The current findings then suggest that these cues could potentially be available to CI listeners. Many attempts have been made in the last few years to provide a better encoding of pitch for CI recipients. Increasing the number of bands in the low frequencies could allow capturing the first harmonic, and has been shown to improve the perception of pitch (Carroll and Zeng, 2007). Unfortunately, with a fixed number of bands, increasing the number of bands in the low frequency region leads to decrease the number of bands in the higher regions, which appeared to be detrimental for speech intelligibility (Carroll and Zeng, 2007). Hence, while the number of bands is limited, there seems to be a tradeoff between pitch perception and speech intelligibility. Moreover, despite the increasing number of channels in cochlear-implants (up to 22), it seems that most CI recipients do not benefit from more than seven bands for speech recognition (Friesen *et al.*, 2001). These results suggest that pitch cues should be enhanced in existing bands to avoid degradation of spectral cues required for speech perception. Instead of increasing the spectral pitch cues, Green *et al.* (2005) have enhanced the temporal pitch cues by adding to standard processing a 100% amplitude-modulation at the F0 frequency in all the channels This manipulation did improve the perception of prosody. However, again, it appeared that the modified processing had a detrimental effect on vowel recognition. These two strategies to enhance pitch perception in CI do not account for the F_0 related spectral cues found in the current study. Further investigation is then required to evaluate the ability of CI to take advantage of these cues for segregation of speech in ecological situations.

3.5 Conclusions

1. Temporal pitch cues available from a noise-band vocoder are not sufficient to induce obligatory streaming at realistic speech rates.

2. The quantization of the spectrum in the vocoder enhances a F_0 related spectral cue that is able to induce streaming. This cue might be available to CI users.
3. It cannot be ruled out that envelope periodicity cues could induce obligatory streaming in CI users since this cues are stronger in real CI than in CI simulations.

Aknowledgments

The authors whish to thank Andrew J. Oxenham and Christophe Micheyl for their helpful comments on this study. This work was supported in part by NIH grants DC05795 and DC08594, by a doctoral grant from the Région Rhône-Alpes (France), and by grant JC6007 from the Ministère de l'Enseignement supérieur et de la Recherche (France).

References

American National Standard Institute (1995). ANSI S3.7-1995 (R2003), Methods for coupler calibration of earphones (New-York).

American National Standard Institute (2004). ANSI S3.6-2004, Specifications for audiometers (New-York).

Baer, T. and Moore, B.C.J. (1993). "Effects of spectral smearing on the intelligibility of sentences in noise," J. Acoust. Soc. Am. 94, 1229-1241.

Beauvois, M.W. and Meddis, R. (1996). "Computer simulation of auditory stream segregation in alternating-tone sequences," J. Acoust. Soc. Am. 99, 2270-2280.

Berkes, P. and Zito, T. (2007). Modular Toolkit for Data Processing (version 2.1), `http://mdp-toolkit.sourceforge.net`.

de Boer, B. (2000). "Self-organization in vowel systems," J. Phonetics 28(4), 441-465.

Bregman, A.S. (1990). Auditory Scene Analysis : The Perceptual Organization of sound (The MIT Press, Massachusetts, USA).

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

Bregman, A.S. and Campbell, J. (1971). "Primary auditory stream segregation and perception of order in rapid sequences of tones," J. Exp. Psychol. 89, 244-249.

Brokx, J.P.L. and Nooteboom, S.G. (1982). "Intonation and the perceptual separation of simultaneous voices," J. Phonetics 10, 23.

Burns, E.M. and Viemeister, N.F. (1981). "Played-again SAM : Further observations on the pitch of amplitude-modulated noise," J. Acoust. Soc. Am. 70, 1655-1660.

Burns, E.M. and Viemeister, N.F. (1976). "Nonspectral pitch," J. Acoust. Soc. Am. 60, 863-869.

Carroll, J. and Zeng, F. (2007). "Fundamental frequency discrimination and speech perception in noise in cochlear implant simulations," Hear. Res. 231, 42-53.

Chatterjee, M., Sarampalis, A., and Oba, S.I. (2006). "Auditory stream segregation with cochlear implants : A preliminary report," Hear. Res. 222, 100-107.

Cooper, H.R. and Roberts, B. (2007). "Auditory stream segregation of tone sequences in cochlear implant listeners," Hear. Res. 225, 11-24.

Dorman, M.F., Cutting, J.E., and Raphael, L.J. (1975). "Perception of temporal order in vowel sequences with and without formant transitions," J. Exp. Psychol. Hum. Percept. Perform. 104, 147-153.

Dorman, M.F., Loizou, P.C., and Rainey, D. (1997). "Speech intelligibility as a function of number of channels of stimulation for signal processors using sine-wave and noise-band outputs," J. Acoust. Soc. Am. 102, 2403-2410.

Dorman, M.F., Loizou, P.C., Fitzke, J., and Tu, Z. (1998). "The recognition of sentences in noise by normal-hearing listeners using simulations of cochlear-implant signal processors with 6-20 channels," J. Acoust. Soc. Am. 104, 3583-3585.

Elhilali, M. and Shamma, S. (2007). "The correlative brain : A stream segregation model," in Hearing - From Sensory Processing to Perception, edited by Kollmeier, B., Klump, G., Hohmann, V., Langemann, U., Mauermann, M., Uppenkamp, S., and Verhey, J. (Springer).

Friesen, L.M., Shannon, R.V., Baskent, D., and Wang, X. (2001). "Speech recognition in noise as a function of the number of spectral channels : comparison of acoustic hearing and cochlear implants," J. Acoust. Soc. Am. 110, 1150-1163.

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

Green, T., Faulkner, A., Rosen, S., and Macherey, O. (2005). "Enhancement of temporal periodicity cues in cochlear implants : effects on prosodic perception and vowel identification," J. Acoust. Soc. Am. 118, 375-385.

Greenberg, S., Hollenback, J., and Ellis, D. (1996). "Insights into spoken language gleaned from phonetic transcription of the switchboard corpus," Proceedings of the International Conference on Spoken Language Processing, pp. S24-27.

Grimault, N., Bacon, S.P., and Micheyl, C. (2002). "Auditory stream segregation on the basis of amplitude modulation rate," J. Acoust. Soc. Am. 111, 1340-1348.

Grimault, N., Micheyl, C., Carlyon, R.P., Arthaud, P., and Collet, L. (2001). "Perceptual auditory stream segregation of sequences of complex sounds in subjects with normal and impaired hearing," Br. J. Audiol. 35(3), 173-182.

Grimault, N., Micheyl, C., Carlyon, R.P., Arthaud, P., and Collet, L. (2000). "Influence of peripheral resolvability on the perceptual segregation of harmonic tones differing in fundamental frequency," J. Acoust. Soc. Am. 108, 263-271.

Grose, J.H., and Hall, J.W. (1996). "Perceptual organization of sequential stimuli in listeners with cochlear hearing loss," J. Speech Hear. Res. 39, 1149-1158.

Hanna, T.E. (1992). "Discrimination and identification of modulation rate using a noise carrier," J. Acoust. Soc. Am. 91, 2122-2128.

Hartmann, W.M. and Johnson, D. (1991). "Stream segregation and Peripheral channeling," Music Percept. 9, 115-184.

Healy, E.W. and Bacon, S.P. (2002). "Across-frequency comparison of temporal speech information by listeners with normal and impaired hearing," J. Speech Lang. Hear. Res. 45, 1262-75.

Healy, E.W. and Steinbach, H.M. (2007). "The effect of smoothing filter slope and spectral frequency on temporal speech information," J. Acoust. Soc. Am. 121, 1177-1181.

Hong, R.S., and Turner, C.W. (2006). "Pure-tone auditory stream segregation and speech perception in noise in cochlear implant recipients," J. Acoust. Soc. Am. 120, 360-374.

Kawahara, H., Masuda-Katsuse, I. and de Cheveigné, A. (1999). "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction : Possible role of a repetitive structure in sounds," Speech Comm. 27, 187-207.

Laneau, J., Moonen, M., and Wouters, J. (2006). "Factors affecting the use of noise-band vocoders as acoustic models for pitch perception in cochlear implants," J. Acoust. Soc. Am. 119, 491-506.

Mackersie, C., Prida, T., and Stiles, D. (2001). "The role of sequential stream segregation and frequency selectivity in the perception of simultaneous sentences by listeners with sensorineural hearing loss," J. Speech Lang. Hear. Res. 44, 19-28.

McCabe, S.L. and Denham, M.J. (1997). "A model of auditory streaming," J. Acoust. Soc. Am. 101, 1611.

Moore, B.C.J. (1998). Cochlear hearing loss (Whurr, London).

Moore, B.C.J. and Carlyon, R.P. (2005). "Perception of Pitch by People with Cochlear Hearing Loss and Cochlear Implant Users," in Pitch : Neural coding and perception, edited by Plack, C.J., Oxenham, A.J., Fay, R.R., and Popper, A.N. (Springer).

Moore, B.C.J. and Gockel, H. (2002). "Factors influencing sequential stream segregation," Acta Acustica united with Acustica 88, 320-332.

Moore, B.C.J. and Peters, R.W. (1992). "Pitch discrimination and phase sensitivity in young and elderly subjects and its relationship to frequency selectivity," J. Acoust. Soc. Am. 91, 2881-2893.

van Noorden, L.P.A.S. (1975). "Temporal coherence in the perception of tones sequences," PhD dissertation, Eindhoven University of Technology, The Netherlands.

Nooteboom, S.G., Brokx, J.P.L. and de Rooij, J.J. (1978). "Contributions of prosody to speech perception," in Level, W.J.M., d'Arcais, G.B.F. (Eds), Studies in the Perception of Language. Wiley and Sons, New-York, pp. 75-107.

Qin, M.K. and Oxenham, A.J. (2005). "Effects of envelope-vocoder processing on F_0 discrimination and concurrent-vowel identification," Ear Hear. 26, 451-460.

Roberts, B., Glasberg, B.R., and Moore, B.C.J. (2002). "Primitive stream segregation of tone sequences without differences in fundamental frequency or passband," J. Acoust. Soc. Am. 112, 2074-2085.

Rogers, C.F., Healy, E.W., and Montgomery, A.A. (2006). "Sensitivity to isolated and concurrent intensity and fundamental frequency increments by cochlear implant users under natural listening conditions," J. Acoust. Soc. Am. 119, 2276-2287.

Rose, M.M. and Moore, B.C.J. (2005). "The relationship between stream segregation and frequency discrimination in normally hearing and hearing-impaired subjects," Hear. Res. 204, 16-28.

Rose, M.M. and Moore, B.C.J. (1997). "Perceptual grouping of tone sequences by normally hearing and hearing-impaired listeners," J. Acoust. Soc. Am. 102, 1768-1778.

Shannon, R.V. (1983). "Multichannel electrical stimulation of the auditory nerve in man. I. Basic psychophysics," Hear. Res. 11, 157-189.

Shannon, R.V., Zeng, F., Kamath, V., Wygonski, J., and Ekelid, M. (1995). "Speech recognition with primarily temporal cues," Science 270, 303-304.

Stainsby, T.H., Moore, B.C.J., and Glasberg, B.R. (2004). "Auditory streaming based on temporal structure in hearing-impaired listeners," Hear. Res. 192, 119-130.

Stickney, G.S., Assmann, P.F., Chang, J. and Zeng, F. (2007). "Effects of cochlear implant processing and fundamental frequency on the intelligibility of competing sentences," J. Acoust. Soc. Am. 122, 1069-1078.

Stickney, G.S., Zeng, F., Litovsky, R. and Assmann, P. (2004). "Cochlear implant speech recognition with speech maskers," J. Acoust. Soc. Am. 116, 1081-1091.

Tong, Y.C. and Clark, G.M. (1985). "Absolute identification of electric pulse rates and electrode positions by cochlear implant patients," J. Acoust. Soc. Am. 77, 1881-1888.

Townshend, B., Cotter, N., Compernolle, D.V., and White, R.L. (1987). "Pitch perception by cochlear implant subjects," J. Acoust. Soc. Am. 82, 106-115.

Tsuzaki, M., Takeshima, C., Irino, T., and Patterson, R.D. (2007). "Auditory stream segregation based on speaker size, and identification of size-modulated vowel sequences," in Hearing - From Sensory Processing to Perception, edited by Kollmeier, B., Klump, G., Hohmann, V., Langemann, U., Mauermann, M., Uppenkamp, S. and Verhey, J. (Springer).

Vliegen, J., Moore, B.C.J., and Oxenham, A.J. (1999). "The role of spectral and periodicity cues in auditory stream segregation, measured using a temporal discrimination task," J. Acoust. Soc. Am. 106, 938-945.

Vliegen, J. and Oxenham, A.J. (1999). "Sequential stream segregation in the absence of spectral cues," J. Acoust. Soc. Am. 105, 339-346.

Zeng, F.G. (2002). "Temporal pitch in electric hearing," Hear. Res. 174, 101-106.

Zeng, F.G. and Shannon, R.V. (1994). "Loudness-coding mechanisms inferred from electric stimulation of the human auditory system," Science 264, 564-566.

Chapitre 4

Streaming, perception de voix concurrentes et sélectivité fréquentielle

Dans cette étude, le streaming, la perception de voix concurrentes et la sélectivité fréquentielle de 25 sujets ont été évalués et comparés. Le streaming était mesuré en utilisant notre paradigme des 6 voyelles. La perception de voix concurrentes était mesurée en présentant des mots dans un flux de parole retourné temporellement, et en faisant varier le rapport entre les niveaux des mots cible et du masque ainsi que la différence de hauteur fondamentale entre la cible et le masque. Enfin la sélectivité fréquentielle était mesurée en présentant un son pur dans un bruit à échancrure permettant ainsi d'obtenir la valeur de l'ERB.

Comme suggéré par de précédentes études, une relation a été trouvée entre les performances de streaming avec des séquences de voyelles et les performances de perception de la parole dans un masque. Une relation a aussi été trouvée entre la sélectivité fréquentielle et la perception de voix concurrentes. En revanche cette relation ne concerne que le niveau moyen d'intelligibilité, indépendamment du F_0 de la cible. L'apport de la différence de F_0 ne s'est pas révélé dépendant de l'ERB. De même, l'effet de la différence d'ERB sur les performances de streaming ne s'est pas révélé significatif.

La contradiction de ces résultats avec ceux du chapitre 2 est discutée. Des hypothèses et pistes de recherche sont proposées afin d'éclaircir les mécanis-

mes permettant de profiter d'une différence de F_0 pour séparer deux voix concurrentes.

Article en préparation.

Pitch-based streaming of vowel sequences and speech-in-speech segregation in relation to frequency selectivity

Etienne Gaudrain and Nicolas Grimault*
Neurosciences Sensorielles, Comportement, Cognition, CNRS UMR 5020, Université Lyon 1, 50 avenue Tony Garnier, 69366 Lyon Cedex 07, France

Eric W. Healy
Speech Psychoacoustics Laboratory, Department of Communication Sciences and Disorders, University of South Carolina, Columbia, 29208

Jean-Christophe Béra
Inserm U556, Lyon, France

Abstract

Simultaneous and sequential segregation are described as the base mechanisms for the auditory scene analysis and are likely to be involved in concurrent speech segregation. However, simultaneous segregation has been found to be uncorrelated to speech-in-noise perception whereas it appeared to be related to pure-tones fusion threshold. This study aimed to clarify the relationship between pitch-based speech-in-speech segregation, pitch-based streaming and frequency selectivity. Twenty-five listeners with close to normal hearing were involved in this study. Auditory filter widths were derived from a notch-noise method. Speech-in-speech perception was measured using words presented in a single talker background time reversed, with various pitch differences between target and masker. The streaming performance was measured using an objective name-order task on vowel sequences. The results showed a correlation between frequency selectivity and performance in speech-in-speech perception suggesting that intelligibility mainly relies on simultaneous masking. A correlation was also found between the effect of pitch on the speech-in-speech perception and the effect of pitch on the streaming performance. However, no correlation was found between streaming and frequency selectivity. These latter results suggest that pitch-based segregation probably relies on pitch discrimination which is only poorly correlated to frequency selectivity.

PACS numbers : 43.66.Mk, 43.66.Sr, 43.71.Es, 43.71.Ky

*Electronic mail : ngrimault@olfac.univ-lyon1.fr

4.1 Introduction

The simultaneous and the sequential segregation are commonly considered as the base mechanisms of Auditory Scene Analysis (ASA, Bregman, 1990). This hypothesis implies that these mechanisms are involved in the resolution of Cocktail Party situations (Cherry, 1953), or in speech-in-speech perception tasks. Correlations have then been searched between simultaneous or sequential segregation and speech perception performances. Perception of speech within an acoustic masker and ASA base mechanisms seem to be submitted to many common factors, among which pitch is probably one of the most important.

Brokx and Nooteboom (1982) reported that introducing a pitch difference between two sentences uttered by the same speaker increased the intelligibility of the target sentence. A pitch difference as small as 3 semitones was enough to increase the percentage of correct responses of 20%. Similarly, Summers and Leek (1998) found an improvement of more than 10% in normal-hearing (NH) listeners when adding a pitch difference of 4 semitones between two simultaneous synthetic sentences. In these two reports, the percentage of correct responses grown roughly linearly with the pitch difference in semitones. Bird and Darwin (1998) used speech masker that was almost entirely voiced to enhance the effect of F_0. They observed that the percentage of words correctly reported increased by 40% between 0 and 2 semitones, and kept growing of 20% between 2 and 8 semitones.

Contrastively, the benefit of F_0 difference (ΔF_0) for simultaneous concurrent vowel identification is saturated over two semitones (*e.g.* Assmann and Summerfield, 1990; Meddis and Hewitt, 1992). The difference in the range of ΔF_0 where concurrent sentences and double vowel identification are improved suggests that these two phenomena may rely on different underlying mechanisms or at least that simultaneous segregation is not the only pitch-based segregation mechanism involved in concurrent sentences perception. Summers and Leek (1998) highlighted this difference by comparing the performances in a concurrent-vowel task and in a concurrent-sentence task in NH and hearing-impaired (HI) listeners. These authors found that the F_0 benefit in the concurrent-sentence task was not clearly associated with the F_0 benefit in the concurrent-vowel task, especially in HI listeners.

Early reports on sequential segregation involving pure/complex tones have shown that streaming can be induced by a wider range of $\Delta F/\Delta F_0$ than simultaneous segregation does (van Noorden, 1975). The effect of ΔF_0 on the streaming of synthesized vowel sequences was found to grow continuously from 0 to 12 semitones (Gaudrain *et al.*, 2007). Mackersie *et al.* (2001) studied the relationship between streaming and performances in concurrent-

sentence task in NH and HI listeners. Streaming was evaluated measuring the fission threshold (Rose and Moore, 1997) that ranged from 2 to 6 semitones across listeners. Performance in concurrent-speech task was the percentage of correct words in first repeated sentence. Each sentence pair consisted of one sentence spoken by a female talker (mean F_0 =240 Hz) and one sentence spoken by a male talker (mean F_0 =115 Hz). The results revealed that sequential segregation and concurrent speech perception were strongly correlated. Hong and Turner (2006) also observed such a correlation in cochlear implant users while measuring obligatory streaming in pure tones sequences and speech perception in steady-state noise and multi-talker babble. Although both simultaneous and sequential segregation are involved in speech-in-speech perception, the later seems to predict more accurately the performances of speech-in-speech perception, especially for ΔF_0 greater than 2 semitones.

Pitch perception is impaired in HI listeners (Moore and Peters, 1992) and impaired frequency selectivity induces a deficit in F_0-based streaming (Grimault *et al.*, 2001 ; Gaudrain *et al.*, 2007). The effect of frequency selectivity on speech-in-noise has been clearly established when the masker is steady-state noise or amplitude-modulated noise (Festen and Plomp, 1983 ; Glasberg and Moore, 1989 ; Baer and Moore, 1993 ; 1994). Similarly, HI listeners benefit less from ΔF_0 in concurrent-sentences task than NH listeners (Summers and Leek, 1998). However no relationship was found between this benefit of ΔF_0 and the frequency selectivity (Mackersie *et al.*, 2001). It is then possible that the frequency selectivity determines the performances of speech perception in noise but does not accurately determines the speech segregation based on a ΔF_0. Rose and Moore (1997) have investigated the sequential segregation of pure tone sequences in unilaterally-impaired listeners, but found no clear relationship between auditory filter width and fission boundary. Gaudrain *et al.* (2007) used the smearing algorithm developed by Baer and Moore (1993) to simulate auditory filter broadening in NH in a streaming task involving vowel sequences. They observed a significant deficit of streaming due to spectral smearing. However the frequency selectivity was artificially altered in NH listeners, and did not account therefore for ecological variations of frequency selectivity and of other factors that could be related to segregation performances.

The purpose of the current study was to clarify the relations between frequency selectivity, F_0-based streaming with vowels and F_0 benefit in speech-in-speech perception. The variation in frequency selectivity was realized selecting listeners with perfect hearing to slight hearing-loss in low frequencies. Each participant was submitted to three measurements : auditory filter width was evaluated with a probe tone in a notched-noise, obligatory streaming

TAB. 4.1 – Audiometric thresholds in dB-HL of the 25 listeners involved in the study. Tested ear is indicated in the second column.

	Test ear	Frequency (Hz) 250	500	1000	2000	4000	6000	8000
S01	R	20	20	10	0	0	15	5
S02	R	20	15	0	10	15	15	10
S03	L	20	10	5	10	30	40	5
S04	L	20	25	15	20	5	35	30
S05	R	65	70	60	80	85	90	65
S06	L	25	25	25	5	10	10	5
S07	L	20	20	5	10	5	15	5
S08	L	15	10	10	5	10	10	10
S09	L	0	0	0	10	5	20	25
S10	R	10	15	10	15	10	10	0
S11	R	15	10	10	5	15	25	20
S12	L	10	10	5	0	5	5	5
S13	L	20	15	10	0	5	10	10
S14	R	20	15	5	15	10	10	15
S15	L	30	15	10	10	30	10	0
S16	L	25	15	10	5	0	25	20
S17	L	15	10	0	10	25	25	10
S18	L	20	10	15	0	5	20	20
S19	R	5	5	0	5	0	15	0
S20	L	5	5	5	5	20	10	15
S21	R	5	5	0	0	5	15	10
S22	L	30	15	15	30	10	10	15
S23	R	15	15	10	20	-10	15	0
S24	L	15	5	5	0	0	20	10
S25	R	10	5	5	10	5	10	0

performances were evaluated using an objective name-order task on vowel sequences, and speech-in-speech reception was measured using lists of words in reverse speech.

4.2 Method

4.2.1 Listeners

Twenty five listeners with normal hearing to moderate hearing loss participated in the experiment. They ranged in age from 18 to 27 years, with a mean age of 20 years. Listeners were selected on the basis of their audiometric thresholds. To ease the selection of listeners they were tested in only one ear. Their audiometric thresholds (ANSI, 2004) on tested ear are listed in Table 4.1. Listeners were paid an hourly wage for their participation, and signed an inform consent.

4.2.2 Frequency selectivity

Auditory filters were derived using symmetric notched-noise masker and sinusoidal probe tone (Patterson, 1976 ; Glasberg and Moore, 1990), with fixed probe level (*e.g.* Bernstein and Oxenham, 2006). Auditory filter width was measured for two frequencies : 370 Hz and 1394 Hz. These correspond to the mean frequencies of the first and second formant of the vowels used in the streaming test described in the next section. The masker was built using a white noise in which the notch was created using a 16th order Butterworth stop-band filter. Cutoff frequencies for the notch are expressed as a proportion r of the center frequency f_c as follow : $(1-r)\ f_c$ and $(1+r)\ f_c$. In addition, the signal was bandpass filtered between $0.2f_c$ and $1.8f_c$ with a 4th order Butterworth filter. Finally a lowpass noise was added below $0.2f_c$ (4th order filter), 20 dB lower than the notched noise. An additional 16th order lowpass filter with a cutoff frequency of $(1-r)f_c$ was added to avoid this lowpass noise to appear in the notch. The noise duration was 700 ms while the probe tone duration was 500 ms.

Detection threshold of the probe tone were obtained with a two down, one up, two-interval, two-alternative forced-choice (2I-2AFC) paradigm to estimate the 79% point on the psychometric function (Levitt, 1971). The probe tone level was held constant at 63 dB-SPL for 370 Hz, and at 44 dB-SPL for 1394 Hz. These levels were chosen as the mean spectral levels of the two first formants of the six vowels used in the streaming test. At the beginning of the procedure, the probe tone and the masker had the same spectral level. The masker level was then increased according to the responses of the participant. The initial step size was 8 dB before the two first turnarounds, then 4 dB for 2 turnarounds, and finally 2 dB for 8 turnarounds. These 8 turnarounds were averaged to compute the threshold. The thresholds were measured for at least 3 notch ratios by participants, always including 0.0. The other values were determined for each subject in order to avoid saturation of the procedure, and were always smaller than 0.2. For subjects S01 to S07, three measures by f_c were performed, whereas at least six measures were done for subjects S08 to S25.

A fitting procedure was performed to derive auditory filter shapes from the data, using a roex(p) model (Glasberg and Moore, 1990). The average spectrum of 4000 repetitions of the masker noise was used for integration under the filter-shape. The fitting procedure took into account the Sennheiser HD250 Linear II transfer function, the middle-ear transfer function, and variations in filter bandwidth with center frequency. The ERB was then com-

puted from the fitting, as well as the ratio to the ERB_N[1]. The ERB centered on 370 Hz and 1394 Hz are noted ERB_{370} and ERB_{1394} respectively.

4.2.3 Streaming with vowels

The method used to evaluate streaming is based on a naming order-task (Dorman *et al.*, 1975 ; Gaudrain *et al.*, 2007). Sequences of vowels with alternating pitches are presented and the subject is asked to give back the vowels in the correct order. In case of segregation, the perception of the order is lost between the auditory streams. This paradigm provides an objective estimation of obligatory streaming abilities (Gaudrain *et al.*, 2007). The experimental procedure was formally approved by a local ethics committee (CPP Lyon Sud).

Material

The material was similar to the one used in Gaudrain *et al.* (2007) except that in the current study, the vowels were recorded vowels instead of synthesized vowels. The six French vowels /a e i ɔ y u/ were recorded 24 bit, 48 kHz, using a Røde NT1 microphone, a Behringer Ultragain preamplifier, a Digigram VxPocket 440 soundcard and a PC. Speaker was instructed to pronounce all 6 vowels at the same pitch, and to reduce prosodic variations.

The fundamental frequency and duration of each vowel was then manipulated using STRAIGHT (Kawahara *et al.*, 1999). Duration was set to 165 ms, which is approximately the median duration of syllables in English (Greenberg *et al.*, 1996). Average F_0 was adjusted to 100, 134, 179 and 240 Hz. F_0 variations related to intonation were constrained to be less than 0.7 semitone apart from average fundamental frequency. Formant positions were held constant across F_0s.

The vowels were then concatenated to form sequences. Each sequence contained the six different vowels. The F_0 of the vowels alternated between two values $F_{0(1)}$ and $F_{0(2)}$. For all sequences, $F_{0(1)}$ was 100 Hz, and $F_{0(2)}$ was one of the following values : 100, 134, 179 and 240 Hz. Each sequence was created with two presentation rates : Slow at 1.2 vowel/s, and Fast at 6 vowel/s. Slow sequences were made by inserting some silence between the vowels, and were used to check vowel identification performance. Fast sequences were used to observe streaming. Finally, each sequence was repeated to form the final stimuli. Slow sequences were repeated 4 times, and Fast sequences were repeated 20 times, for overall durations of 20 s. The perceptual

[1]The standard ERB was defined by Glasberg and Moore (1990) as : $\mathrm{ERB}_N(f_c) = 24.7\ (4.37 \cdot 10^{-3}\ f_c + 1)$.

distance related to the formants was set minimal and held constant across F_0 conditions (see Gaudrain *et al.*, 2007 for details). Stimuli were generated 16 bits, 44.1 kHz using MATLAB, and played at 85 dB-SPL.

Procedure

Training The training begun with a simple identification task on single vowels. Each vowel, at each F_0 (100, 134, 179 and 240 Hz), was presented twice in random order. Visual feedback was provided after each response. All subjects achieved more than 93% correct, except S05 who reached 69% correct at this stage.

The second step of training involved another form of vowel identification. In this step, vowels were presented in Slow sequences. In each of two blocks, 20 sequences were presented : 5 in each of the F_0s. The procedure was the same as the test procedure described in the next paragraph, except that visual feedback was provided.

Streaming test The streaming test was composed of 2 blocks of 40 sequences each. Half the sequences were in the Slow condition to check identification within test, and half the sequences were in the Fast condition to test streaming. Each sequence was presented to the subject during 20 s, but he/she was locked out from responding during the first 5 s. The subject had then to select 6 vowels to "Write the sequence in the correct order" using a mouse and computer graphic interface. The next sequence was presented when subject clicked the "Submit" button or after the 20 s expired. The average duration of each block was about 12.5 min. For each subject, this procedure provides score as a function of $F_{0(2)}$. The score is the percentage of responses where the subject successfully gave back the six vowels in the correct order. Thus high scores correspond to perception of an integrated stream, while low scores correspond to segregation of the stimuli in two streams.

4.2.4 Speech-in-speech reception

Material

Concurrent speech test consisted in target words presented simultaneously with a constant masker. The masker was time reversed speech so it did not content any semantic information. A male talker was digitally recorded (with the same apparatus as for the vowels in section 4.2.3) reading a newspaper article during 5 min. Silences were deleted, and the RMS level was adjusted to be constant over 12 s time windows with Hann shape and 50%

overlap. The resulting signal was then segmented in pieces of 45 s, and down-sampled to 24 kHz. Finally, each piece was processed with STRAIGHT to set the average F_0 to 100 Hz (min : 86 Hz, max : 116 Hz). The targets were lists of monosyllabic French words uttered by another male speaker. The lists were extracted from the Vocales audio-CD as used in Hoen *et al.* (2007). The words were all balanced in occurrence frequency, phonological neighborhood, number of letters, duration, number of consonant-vowel alternations and gender. The RMS level of all words was adjusted to 85 dB-SPL, and F_0 was set using STRAIGHT. Words were gathered in 24 lists of 10 words. Six lists were processed to reach each of the four F_0s : 100, 134, 179, 240 Hz. As in the streaming test, formant positions were held constant. The level of the target words was fixed while the masker level varied. In each F_0 condition, each of the six lists was combined with a masker with a different target-to-masker ratio (SNR) : -9, -6, -3, 0, 3 and 6 dB.

Procedure

The participants were asked to listen to the word lists and to repeat each word they heard. Each word counted true or false yielding a score for each F_0 and SNR. Fitting of a logistic function was used to compute the SNR that yielded 50% correct, used as the speech reception threshold (SRT). Either the scores or the SRT were used in the analysis presented in the next sections to ease the comparison with previous studies.

4.2.5 Common apparatus

All stimuli were presented using a PC, a Digigram VxPocket 440 soundcard, a Behringer Ultragain amplifier and a Sennheiser HD250 Linear II headphone. Sound levels were calibrated in an artificial ear (Larson Davis AEC101 and 824, ANSI, 1995). All the experiments took place in a sound attenuated booth.

4.3 Results and discussion

4.3.1 Frequency selectivity

The fitting procedure failed at f_c =370 Hz for subjects S01, S07 and S09, and at f_c =1394 Hz for S22. These participants were excluded from any further analysis involving ERB measures. The ERB of the auditory filters of the subjects as a function of audiometric thresholds at center frequency are displayed in Fig. 4.1. The mean ERB_{370} was 80.7 Hz (1.2 times the normal

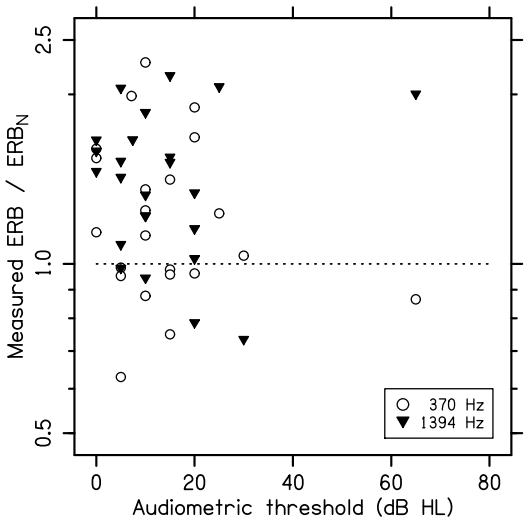

FIG. 4.1 – Values of the ERB of the auditory filter for the subjects plotted as a function of absolute threshold (dB HL) interpolated at the test frequency. The ERB values are plotted relative to the ERBN (Glasberg and Moore, 1990). For test frequency 370 Hz (empty circles), the fitting procedure succeeded for 22 participants. For test frequency 1394 Hz (filled downward triangles), the fitting procedure succeeded for 24 participants.

ERB), and the mean ERB_{1394} was 249 Hz (1.4 times the normal ERB). Most of the points are over ERB_N as reported by Moore (1998, Fig. 3.20).

4.3.2 Streaming with vowels

The results of the streaming test are plotted in the Fig. 4.2. The average identification score (Slow condition) is over 95% correct for all values of $F_{0(2)}$, and only three subjects had scores lower than 90%. As described in Gaudrain *et al.* (2007), the score in the Fast condition reflects the segregation that occurs. The scores decreased from about 50% for $F_{0(2)}$=100 Hz, to about 10% for $F_{0(2)}$=240 Hz. A one-way repeated measure ANOVA with $F_{0(2)}$ as repeated parameter revealed a significant effect of $F_{0(2)}$ [$F(3, 72) = 26.4,\ p < .001$]. This effect reflects the effect of the F_0 difference on streaming : the greater the F_0 difference, the more segregation occurs. The scores are consistent with those observed in Gaudrain *et al.* (2007). In particular, the score at matched F_0 is the same as the one observed in naïve NH listeners with synthetic vowels of 175 ms. It has been argued that the score in this particular

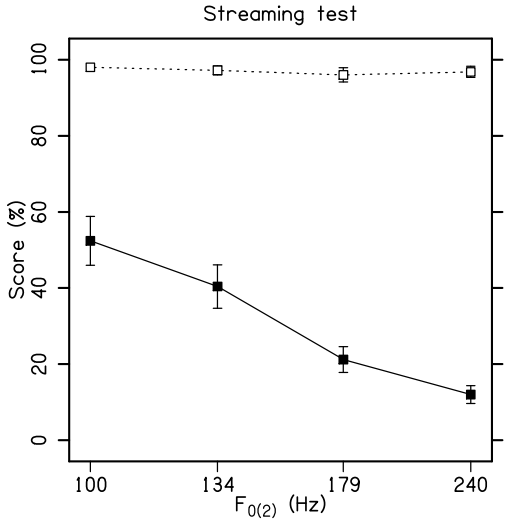

FIG. 4.2 – Average score across participants for the streaming task as a function of $F_{0(2)}$. The score is the percentage of sequences where the subject answered the six vowels in the correct order. The scores for the Fast condition (streaming) are plotted with the solid line and filled squares. The scores for the Slow condition (identification) are plotted with the dashed line and empty squares. The errorbars represent the standard error across participants.

condition reflected the streaming induced by formant structure, as described by Dorman *et al.* (1975).

In the following sections, two measures of streaming are used : the score in the streaming task, and the F_0 benefit in the streaming task. The former is the mean score for a listener, and for a given $F_{0(2)}$. The latter is defined for conditions where $F_{0(2)}$ >100 Hz only, as the difference between the score for $F_{0(1)} = F_{0(2)}$ =100 Hz and the score in another $F_{0(2)}$ condition. So the F_0 benefit in streaming is the decrease of the score using $\Delta F_0 = 0$ condition as reference score.

4.3.3 Speech-in-speech reception

The results of the speech reception test are displayed in the Fig. 4.3. The scores are the percentage of words correctly repeated averaged for each subject, SNR and F_0. Without any F_0 difference between the target and the masker, the mean score is about 50% correct. For F_0 differences greater than 5 semitones, the mean score raises up to more than 60%. A two-way repeated measure ANOVA using SNR and F_0 as repeated parameters indicated that

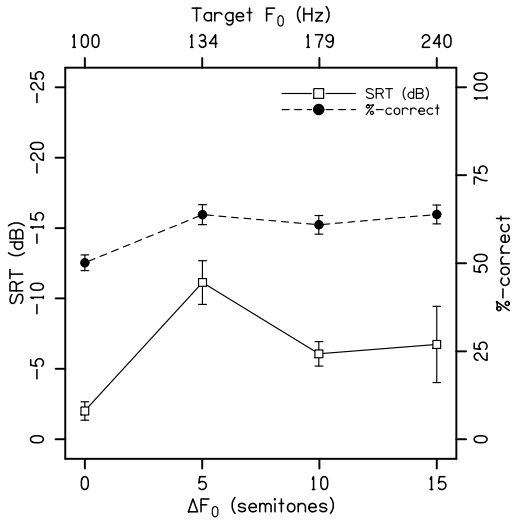

FIG. 4.3 – Speech-in-speech perception performances. *Left axis, open symbols* : Mean SRT across subjects as a function of the target F_0. SRT were computed by fitting a psychometric function on scores against SNR, and choosing the SNR that yields 50% correct. *Right axis, filled symbols* : Mean scores of correct identification of the words in percent correct as a function of the F_0 of the target. The lower axis represents the F_0 of the target as the difference in semitones with the masker (F_0 =100 Hz). The upper axis is the absolute F_0 of the target in Hertz. The errorbars show the inter-individual standard deviation.

the effect of the F_0 difference between target and masker was significant $[F(3, 72) = 31.48,\ p < .001]$, as well as the SNR $[F(5, 120) = 90.17,\ p < .001]$, and their interaction $[F(15, 360) = 11.10,\ p < .001]$.

Because the scores were mostly greater than 50%, the evaluation of the SRT failed at least for one F_0 condition in 11 participants. A one-way repeated measure ANOVA on SRT values in the 14 remaining listeners, using F_0 as repeated parameter revealed a significant effect of the F_0 of the target $[F(3, 39) = 9.12,\ p < .001]$. The mean SRTs, displayed in Fig. 4.3, were calculated on 25, 21, 25 and 17 participants at 100, 134, 179 and 240 Hz respectively.

These results are very similar to those obtained by Brokx and Nooteboom (1982). These authors found an increase of the identification scores from about 40% to 60% for 0 to 3 semitones. The scores observed in the current experiment are slightly greater. This is probably due to the fact that the SNR ranged from -9 to 6 dB in the current experiment while it ranged from -15 to 0 dB in Brokx and Nooteboom (1982). The benefit of F_0 difference for the identification of the target words depends on the SNR as revealed by

the significant interaction. The tendency is that the benefit becomes smaller when the SNR increases.

In the following sections, the speech-in-speech perception score differs from the F_0 benefit in the speech-in-speech perception task like for the streaming task. The former is the mean score for a listener, and for a given target F_0. The latter is defined for conditions where the target F_0 >100 Hz only, as the difference between the score with no F_0 difference (both masker and target F_0 =100 Hz) — considered as the baseline — and the score with a F_0 difference. The same distinction is made between SRT and F_0 benefit in SRT.

4.4 General discussion

4.4.1 Speech-in-speech perception and streaming

A F_0 difference of 5 to 15 semitones yielded an improvement of the speech-in-speech perception scores, and was detrimental for streaming scores, *i.e.* increased the amount of streaming. If streaming is an underlying mechanism involved in speech-in-speech segregation, a relationship between the scores in these two tasks can be expected. In particular a correlation might be found between the F_0 benefit in SRT that yields smaller SRT for greater F_0 difference, and the F_0 benefit in streaming that yields smaller identification scores for the vowel sequences for greater F_0 differences. For both SRT and streaming score, the maximal F_0 benefit has been used to observe the level of correlation. The maximal F_0 benefit in SRT is plotted against the maximal F_0 benefit in streaming for each listener in Fig. 4.4. A simple regression analysis revealed a significant correlation $[r(25) = -0.43,\ F(1, 23) = 5.06,\ p < 0.05]$. This correlation illustrates the fact that listeners who benefit the most of F_0 in speech-in-speech perception are also the ones who have the more difficulties to resist to segregation in the streaming task. Mackersie *et al.* (2001) found a relationship between fusion threshold for pure tones and concurrent sentences perception. Similarly, Hong and Turner (2006) found a relationship in cochlear implant users between obligatory streaming of pure tones and speech-in-noise perception. The current results confirm that a similar relationship can be observed in close to normal hearing listeners with complex tones. Furthermore, since the relationship concerns the F_0 benefit in scores, the current results suggest more clearly that the F_0-based streaming mechanism is involved in speech-in-speech perception.

Contrastively, the same analysis on the raw scores instead of F_0 benefit showed no correlation between speech-in-speech perception and streaming

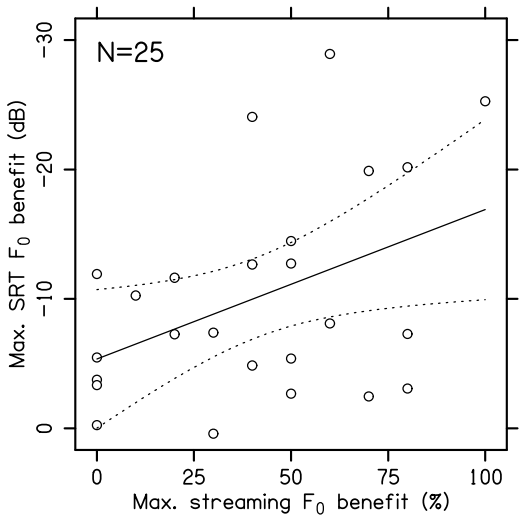

FIG. 4.4 – Maximal F_0 benefit on the SRT as a function of maximal effect of $F_{0(2)}$ on the streaming scores, for 25 listeners. The maximal SRT benefit due to F_0 difference is computed for each listener as the difference between the SRT for target and masker at same F_0 and the best SRT for this listener. Similarly, the maximal benefit of F_0 on streaming is computed as the difference between the score for $F_{0(1)} = F_{0(2)} = 100$ Hz and the lowest score. Each open circle represents a participant. The solid line shows a linear regression function fit to the data for the 25 listeners. The dotted lines define the 95% confidence interval of the regression function.

$[r(24) < .001,\ F(1, 22) < .001,\ p = .998]$. Although the effect of F_0 was comparable in these two tasks, the overall performances were also modulated independently of the F_0. Hence, this absence of correlation reveals the existence of at least one factor that differs between concurrent speech segregation and obligatory streaming.

4.4.2 Frequency selectivity and speech-in-speech perception

The effects of F_0 and frequency selectivity on speech-in-masker perception have been separately studied in various paradigms that have yielded various results. Festen and Plomp (1983) have found a correlation between the speech-in-noise perception and the logarithm of auditory bandwidth estimated using a comb-filter noise masker and a probe tone. Similarly, Glasberg and Moore (1989) measured SRT in quiet and in speech-shaped noise along with many psychoacoustic tests. They found a correlation between SRT in

noise and some measures related to the perception of frequency : tonal frequency difference limens, fundamental frequency difference limens, and ERB. The authors of these two studies argued that speech-in-quiet perception relies on the audiometric threshold, while speech-in-noise perception depends on supra-threshold abilities such as the spectral resolution. Mackersie *et al.* (2001) used simultaneous sentences which F_0s were 115 Hz and 240 Hz. Contrastively to the previous literature (Festen and Plomp, 1983 ; Glasberg and Moore, 1989), these authors did not found any significant correlation between the slopes of the notched-noise masking function (as a representation of frequency selectivity) and the percentage of words correct in sentence repeated. Mackersie *et al.* (2001) argued that simultaneous sentences contain more contextual evidence and acoustic variability than the steady-state noise used in the previous studies in which peripheral masking would have then been enhanced.

In the literature, to carefully observe the effect of frequency selectivity, many researchers have first partialled out the effect of audiometric threshold. In the current experiment, the hearing-losses are moderate and should not have a significant effect. Indeed, no correlation was found between mean audiometric threshold below 2000 Hz and mean speech perception scores [$r = .15,\ p = .49$]. Hence, the variation in audiometric threshold has not to be taken into account for the ERB effect analysis.

In the current study the score of speech-in-speech identification was observed as a function of the ERB value at 370 Hz and 1394 Hz, as plotted in Fig. 4.5. Two separate two-way ANOVAs on identification scores, using ERB as first-order linear predictor and target F_0 as repeated parameter, were performed for each ERB center frequency. These two ANOVAs could not be merged in a single model that would have included both ERB_{370} and ERB_{1394} because the number of listeners for whom the data was available was not the same. These ANOVAs revealed no significant correlation[2] between the average speech-in-speech perception score and the ERB_{370} [$r = 0.26,\ F(1,19) = 1.43,\ p = 0.25$], but a significant correlation with the ERB_{1394} [$r = 0.50,\ F(1,21) = 6.86,\ p < 0.05$]. The partial correlations in this analysis showed that this correlation with ERB_{1394} was only significant for target F_0=179 Hz [$r = 0.54,\ p < 0.01$] and very close from significance for F_0=240 Hz [$r = 0.41,\ p = 0.051$]. Contrastively, a similar analysis of the F_0 benefit in word identification scores revealed no significant effect of the

[2]Subject S05 was excluded from this analysis because his very low average speech-in-speech perception score led to consider him as an outlier. However, the results of the analysis when including this participant are only slightly different : no significant correlation with the ERB at 370 Hz [$r = 0.10,\ F(1,20) = 0.22,\ p = 0.64$], but a significant correlation with the ERB at 1394 Hz [$r = -0.44,\ F(1,22) = 5.25,\ p < 0.05$].

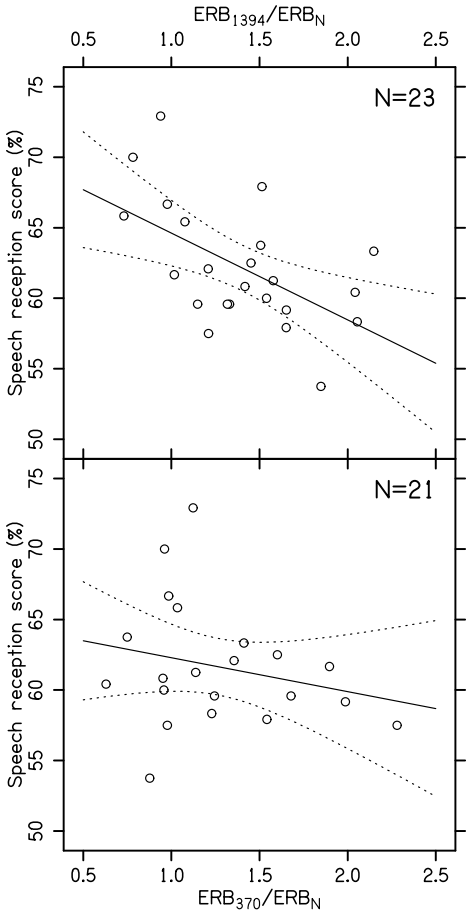

FIG. 4.5 – Average speech-in-speech perception score as function of ERB ratio. The average speech-in-speech perception score is the mean score across target F_0 and SNR for each listener. The ERB ratio is displayed as the ratio relative to the ERB_N. The *upper panel* represents the data for 23 listeners for the ERB centered on 1394 Hz. The *lower panel* represents the data for 21 listeners for the ERB centered on 370 Hz. The solid line shows a linear regression function fit to the data. The dotted lines define the 95% confidence interval of the regression function.

ERB value, neither at 370 Hz [$r = -0.08,\ F(1,20) = 0.13,\ p = 0.72$] nor at 1394 Hz [$r = 0.02,\ F(1,22) = 0.01,\ p = 0.92$]. These results indicate that the ERB_{1394} is related to the overall intelligibility but not to the effect of F_0 on segregation. This effect of the frequency selectivity on the overall performance is consistent with the observations of speech-in-noise perception (Festen and Plomp, 1983 ; Glasberg and Moore, 1990), but contrasts with the results found by Mackersie *et al.* (2001).

4.4.3 Frequency selectivity and streaming with vowels

The peripheral channeling theory (Hartmann and Johnson, 1991) hypothesized a relationship between the frequency resolution and the streaming performances. However, a few studies have attempted to observe the effect of frequency selectivity on streaming. Rose and Moore (1997) have measured the fission boundary for pure tones in listeners with normal hearing

and in listeners with unilateral cochlear hearing-loss. For the normal hearing listeners, they found that the frequency difference at the fission boundary was constant when expressed in ERBs. Contrastively, they observed no clear difference in the fission boundary for normal and impaired ears of the unilaterally impaired listeners. More recently, these authors have compared fission boundary to frequency difference limens in NH and HI listeners (Rose and Moore, 2005). They observed, for the NH listeners, that the fission boundary was fairly constantly 8 times larger that the frequency difference limen in the frequency region 250-2000 Hz. They did not found such a clear relationship in HI listeners, however they evidenced that enlarged frequency difference limens may contribute to enlarged fission boundary, though some other factors are likely to be involved. Grimault *et al.* (2001) also observed streaming performances in NH and HI listeners, but using resolved and unresolved complex tones. They found that complex tones that were unresolved for both NH and HI yielded to similar streaming performances, while complex tones resolved for NH but not HI induced more streaming in NH than in HI. Gaudrain *et al.* (2007) used a spectral smearing algorithm to simulate some auditory filter broadening (Baer and Moore, 1993) in a streaming task that involved synthetic vowels. They observed that the spectral smearing hindered the F_0-based obligatory streaming.

In the current study, the performances of the participants were compared to the measured ERBs. Two separate two-way ANOVAs on F_0 benefit in streaming scores, using ERB as first-order linear predictor and $F_{0(2)}$ as repeated parameter, were performed for each ERB center frequency. They revealed no significant effect of ERB_{370} [$r = .28,\ F(1, 20) = 1.68,\ p = .21$] neither of ERB_{1394} [$r = .02,\ F(1, 22) = .01,\ p = .93$] on the streaming score improvement due to the F_0 difference. Average F_0 benefit in streaming score is plotted as a function of ERB in the Fig. 4.6.

Two separate two-way ANOVAs on streaming scores, using ERB as first-order linear predictor and $F_{0(2)}$ as repeated parameter, were performed for each ERB center frequency. They revealed no significant effect of ERB_{370} [$r = .05,\ F(1, 20) = .06,\ p = .82$] nor ERB_{1394} [$r = .02,\ F(1, 22) = .01,\ p = .94$].

Neither the average scores of streaming, nor the F_0 benefit, were correlated to any of the two ERBs measured. This result is relatively consistent with those obtained with pure tones by Rose and Moore (1997) and Mackersie *et al.* (2001). However, they are in contradiction with Gaudrain *et al.* (2007) where simulated broadened auditory filters were found to significantly improve the scores in a similar streaming task. More specifically, they reported that the spectral smearing improved both the mean scores and the F_0 benefit in streaming. Gaudrain *et al.* (2007) argued that the spectral smearing

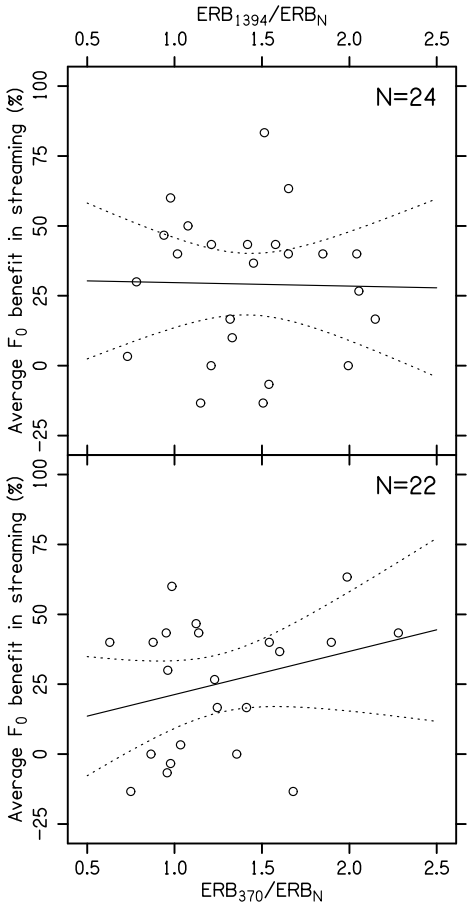

FIG. 4.6 – Average F_0 benefit in streaming as a function of ERB ratio. Each data point indicates the individual data of a listener. The average F_0 benefit is the difference between the score for $F_{0(1)} = F_{0(2)}$ =100 Hz and the mean score for $F_{0(2)}$ >100 Hz. The ERB ratio is displayed as the ratio relative to the ERB_N. The *upper panel* represents the 24 points for the ERB centered on 1394 Hz. The *lower panel* represents the 22 points for the ERB centered on 370 Hz. The solid lines show the linear regression function fits to the data. The dotted lines define the 95% confidence interval of the regression function.

hindered the streaming based on the F_0 difference as well as the streaming based on the formant structure. The overall performance in the streaming task depends on F_0-based and formant structure-based streaming but also reflects the ability of the listeners to perform the name-order task beside any segregation effect. Contrastively, the F_0 benefit reflects solely the effect of F_0-based streaming.

The fact that the F_0 benefit was not correlated to the frequency selectivity in the current experiment suggests that F_0-based streaming relies on another psychoacoustic factor. As found by Rose and Moore (2005) with pure tones, F_0 based streaming probably depends on F_0 discrimination performances, which has been found to be poorly correlated to frequency selectivity (Moore and Peters, 1992). The discrimination of the F_0 in complex tones involves the perception of temporal cues : temporal envelope periodicity and fine structure (Moore and Moore, 2003). The ability of each subject to benefit from these cues is probably not directly correlated to its frequency

selectivity, especially when its frequency selectivity is close to normal. The spectral smearing algorithm of Baer and Moore (1993) used by Gaudrain *et al.* (2007) to simulate broadened auditory filters actually mimics the spectral aspect of frequency selectivity impairment but does not account for the time coding of the waveform. Indeed, the temporal fine structure is markedly altered in the smearing algorithm, at least by the time windowing. Hence the effect of spectral smearing observed in Gaudrain *et al.* (2007) could have been caused by both the degradation of the frequency selectivity and the degradation of the temporal fine structure. In the current study, the ability to handle temporal fine structure probably varies across participants, relatively independently of frequency selectivity. Further investigations are required to assess this hypothesis.

4.4.4 Conclusions

(1) F_0 benefit in streaming and F_0 benefit in speech-in-speech perception were correlated, suggesting that they are driven by the same underlying factor. Since none of them was correlated to ERB, F_0 discrimination ability would be a good candidate as common underlying factor.

(2) Streaming scores did not depend on the ERB while speech-in-speech performance did. Whatever are the segregation abilities, the identification performances are determined by the amount of available information. Energetic masking is probably the main factor that determines the intelligibility of the words, and it depends on frequency selectivity.

Aknowledgments

The authors whish to thank Claire Grataloup and François Pellegrino for providing word list material and support in building the speech-in-speech experiment, and Joshua G.W. Bernstein for providing a MATLAB implementation of the auditory filter function fitting procedure. This work was supported in part by NIH grants DC05795 and DC08594, by a doctoral grant from the Région Rhône-Alpes (France), and by grant JC6007 from the Ministère de l'Enseignement supérieur et de la Recherche (France).

References

American National Standard Institute (2004). ANSI S3.6-2004, Specifications for audiometers (New-York).

American National Standard Institute (1995). ANSI S3.7-1995 (R2003), Methods for coupler calibration of earphones (New-York).

Assmann, P.F. and Summerfield, Q. (1990). "Modeling the perception of concurrent vowels : vowels with different fundamental frequencies," J. Acoust. Soc. Am. 88, 680-697.

Baer, T. and Moore, B.C.J. (1993). "Effects of spectral smearing on the intelligibility of sentences in noise," J. Acoust. Soc. Am. 94, 1229-1241.

Baer, T. and Moore, B.C.J. (1994). "Effects of spectral smearing on the intelligibility of sentences in the presence of interfering speech," J. Acoust. Soc. Am. 95, 2277-2280.

Bernstein, J.G.W. and Oxenham, A.J. (2006). "The relationship between frequency selectivity and pitch discrimination : sensorineural hearing loss," J. Acoust. Soc. Am. 120, 3929-3945.

Bird, J. and Darwin, C.J. (1998). "Effects of a difference in fundamental frequency in separating two sentences," in Psychophysical and physiological advances in hearing, edited by Palmer, A. R., Rees, A., Summerfield, A. Q. and Meddis, R. (Whurr, London).

Bregman, A.S. (1990). Auditory Scene Analysis : The Perceptual Organization of sound (The MIT Press, Massachusetts, USA).

Brokx, J.P.L. and Nooteboom, S.G. (1982). "Intonation and the perceptual separation of simultaneous voices," J. Phonetics 10, 23.

Cherry, E.C. (1953). "Some Experiments on the Recognition of Speech, with One and with Two Ears," J. Acoust. Soc. Am. 25, 975-979.

Dorman, M.F., Cutting, J.E., and Raphael, L.J. (1975). "Perception of temporal order in vowel sequences with and without formant transitions," J. Exp. Psychol. Hum. Percept. Perform. 104, 147-153.

Festen, J.M. and Plomp, R. (1983). "Relations between auditory functions in impaired hearing," J. Acoust. Soc. Am. 73, 652-662.

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

Glasberg, B.R. and Moore, B.C.J. (1989). "Psychoacoustic abilities of subjects with unilateral and bilateral cochlear hearing impairments and their relationship to the ability to understand speech," Scand. Audiol. Suppl. 32, 1-25.

Glasberg, B.R. and Moore, B.C.J. (1990). "Derivation of auditory filter shapes from notched-noise data," Hear. Res. 47, 103-138.

Greenberg, S., Hollenback, J., and Ellis, D. (1996). "Insights into spoken language gleaned from phonetic transcription of the switchboard corpus," Proceedings of the International Conference on Spoken Language Processing, pp. S24-27.

Grimault, N., Micheyl, C., Carlyon, R.P., Arthaud, P., and Collet, L. (2001). "Perceptual auditory stream segregation of sequences of complex sounds in subjects with normal and impaired hearing," Br. J. Audiol. 35, 173-182.

Hartmann, W.M. and Johnson, D. (1991). "Stream segregation and Peripheral channeling," Music Percept. 9, 115-184.

Hoen, M., Meunier, F., Grataloup, C., Pellegrino, F., Grimault, N., Perrin, F., Perrot, X., and Collet, L. (2007). "Phonetic and lexical interferences in informational masking during speech-in-speech comprehension," Speech Comm. 49, 905-916.

Hong, R.S. and Turner, C.W. (2006). "Pure-tone auditory stream segregation and speech perception in noise in cochlear implant recipients," J. Acoust. Soc. Am. 120, 360-374.

Kawahara, H., Masuda-Katsuse, I., and de Cheveigné, A. (1999). "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F_0 extraction : Possible role of a repetitive structure in sounds," Speech Comm. 27, 187-207.

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

Mackersie, C., Prida, T., and Stiles, D. (2001). "The role of sequential stream segregation and frequency selectivity in the perception of simultaneous sentences by listeners with sensorineural hearing loss," J. Speech Lang. Hear. Res. 44, 19-28.

Meddis, R. and Hewitt, M.J. (1992). "Modeling the identification of concurrent vowels with different fundamental frequencies," J. Acoust. Soc. Am. 91, 233-245.

Moore, B.C.J. (1998). Cochlear hearing loss (Whurr, London).

Moore, B.C.J. and Moore, G.A. (2003). "Discrimination of the fundamental frequency of complex tones with fixed and shifting spectral envelopes by normally hearing and hearing-impaired subjects," Hear. Res. 182, 153-163.

Moore, B.C.J. and Peters, R.W. (1992). "Pitch discrimination and phase sensitivity in young and elderly subjects and its relationship to frequency selectivity," J. Acoust. Soc. Am. 91, 2881-2893.

van Noorden, L.P.A.S. (1975). "Temporal coherence in the perception of tones sequences," PhD dissertation, Eindhoven University of Technology, The Netherlands.

Patterson, R.D. (1976). "Auditory filter shapes derived with noise stimuli," J. Acoust. Soc. Am. 59, 640-654.

Rose, M.M. and Moore, B.C.J. (2005). "The relationship between stream segregation and frequency discrimination in normally hearing and hearing-impaired subjects," Hear. Res. 204, 16-28.

Rose, M.M. and Moore, B.C.J. (1997). "Perceptual grouping of tone sequences by normally hearing and hearing-impaired listeners," J. Acoust. Soc. Am. 102, 1768-1778.

Summers, V. and Leek, M. (1998). "F_0 processing and the separation of competing speech signals by listeners with normal hearing and with hearing loss," J. Speech Lang. Hear. Res. 41, 1294-1306.

Chapitre 5

Streaming de séquences de voyelles chez les malentendants

Les résultats présentés au chapitre 2 montrent qu'en simulant un élargissement des filtres auditifs, les sujets voyaient leurs scores s'améliorer dans notre tâche de streaming. Cette tâche est donc particulièrement séduisante pour mettre en évidence des déficits sensoriels. Il est en effet difficile, lorsque l'on travaille avec des malentendants, de distinguer les déficits sensoriels de possibles déficits cognitifs. Avec le paradigme des séquences de voyelles, un déficit de ségrégation devrait conduire à une amélioration des scores, tandis qu'il est difficile d'imputer ce résultat à un déficit cognitif.

Deux expériences ont été réalisées : l'une avec des normo-entendants et malentendants âgés atteints de presbyacousie, l'autre avec des malentendants jeunes.

5.1 Expérience 1 : Malentendants âgés

Cette première expérience s'est déroulée au Laboratoire Entendre de Montluçon, en collaboration avec Patrick Arthaud (Audioprothésiste D.E., responsable scientifique du groupe Entendre) et Christelle Chabrier (étudiante en Audioprothèse, 3ème année). Elle se déroulait en champ libre de façon à ce que les malentendants puissent être testés avec et sans prothèse.

5.1.1 Sujets

Deux groupes de sujets ont été constitués : un groupe de 8 sujets normo-entendants âgés de 21 à 61 ans, et un groupe de 6 sujets malentendants âgés de 62 à 77 ans. Les normo-entendants présentaient tous une audition normale. Leurs audiogrammes sont présentés figure 5.1. Les malentendants souffraient de pertes d'origine neurosensorielle, généralement de nature presbyacousique. Les pertes observées étaient des pertes en pente, en moyenne de -20 dB-HL à 250 Hz, à -70 dB-HL à 8000 Hz. Les audiogrammes individuels sont tracés figure 5.2. Les sujets ont été répartis en trois groupes, justifiés par l'analyse des résultats : un groupe de normo-entendants de moins de 30 ans (NE-30, 4 sujets), un groupe de normo-entendants de plus de 30 ans (NE+30, 4 sujets), et un groupe de malentendants de plus de 30 ans (ME+30, 6 sujets). Les sujets du groupe ME+30 ayant des seuils auditifs entre 20 et 80 dB HL, on peut estimer qu'ils devaient avoir des filtres auditifs allant de 1 à 5 ERB_N (Moore, 1998, voir la figure 1.14).

Tous les sujets ont été recrutés par l'intermédiaire du même cabinet d'audioprothèse, où ils ont ensuite passé les tests. Les sujets malentendants sont des patients de ce cabinet qui ont exprimé leur intérêt pour cette étude. Enfin, tous les sujets étaient de langue maternelle française.

5.1.2 Matériel et méthode

La méthode employée est identique à celle utilisée dans l'Experiment 1 du chapitre 2. Comme dans cette précédente expérience, les sujets devaient essayer de donner les 6 voyelles dans l'ordre (score ACROSS) ; et si malgré leurs efforts ils n'y parvenaient pas, ils devaient donner deux groupes de 3 voyelles (score WITHIN). Il a été montré que cette tâche permettait d'estimer le profil du seuil de cohérence temporelle (van Noorden, 1975). Une tâche supplémentaire, subjective, a ici été ajoutée : après avoir donné les 6 voyelles dans l'ordre, les sujets devaient indiquer s'ils avaient entendu 1 voix ou 2.

Suite à quelques tests préliminaires, la durée des voyelles a été fixée à 175 ms. Cette valeur est cohérente avec les travaux de Nooteboom *et al.* (1978) et correspond à la durée moyenne d'une voyelle dans les signaux de parole courants (Richie, Kewley-Port, et Coughlin, 2003). Chaque voyelle était suivie d'un silence de 5 ms, ce qui devait avoir pour effet d'en faciliter

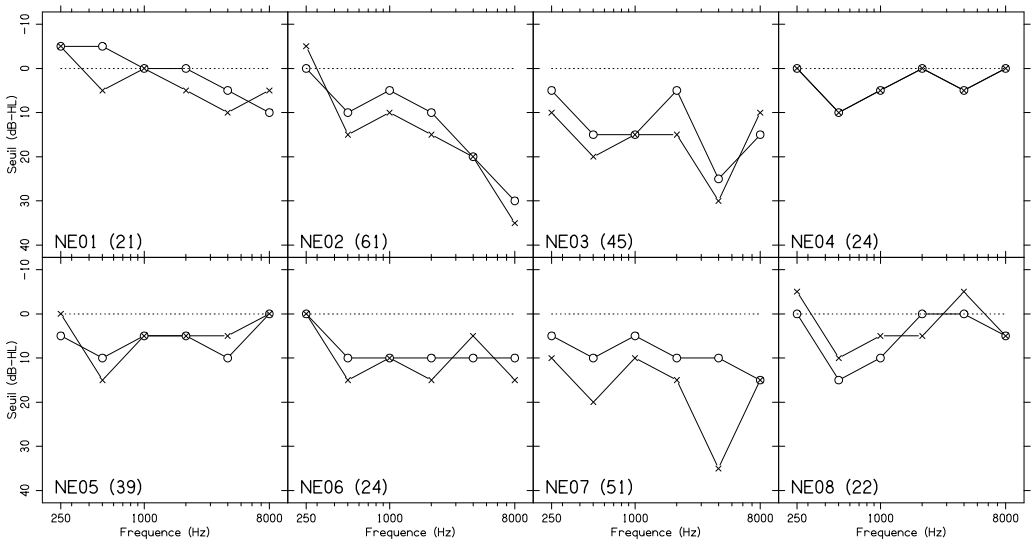

FIG. 5.1 – Audiogrammes des normo-entendants. Les cercles ∘ représentent l'oreille droite, et les croix × l'oreille gauche. L'âge des sujets est indiqué entre parenthèses.

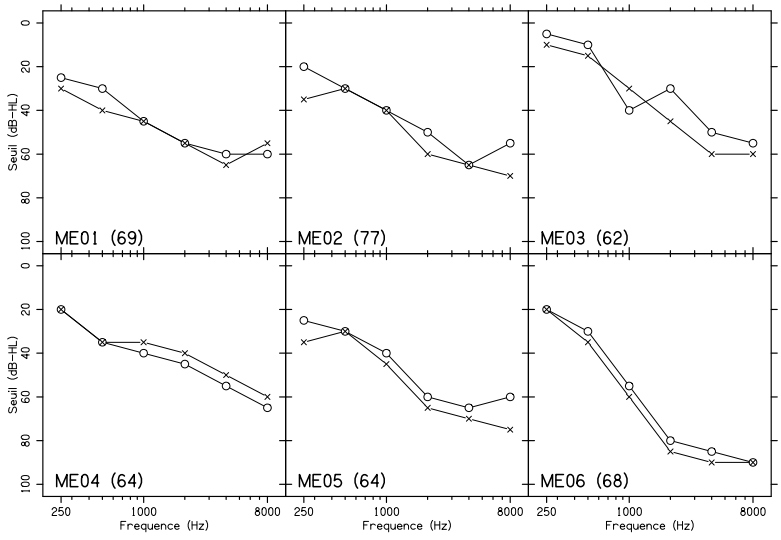

FIG. 5.2 – Comme la figure 5.1 pour les sujets malentendants âgés.

l'intelligibilité. Le tempo de ces séquences était donc caractérisé par un intervalle inter-stimuli d'une durée de 180 ms, soit une cadence syllabique de 5,6 voy./s.

Les stimuli étaient présentés en champ libre *via* une carte son externe Roland UA-30, un amplificateur Yamaha AX-V401 et une enceinte Elipson Audiopro 80 W. Ce matériel est celui employé habituellement au sein du laboratoire d'audioprothèse pour les présentations en champ libre. La pièce dans laquelle étaient testés les sujets est traitée de façon à réduire la réverbération et atténuer les sons provenant de l'extérieur.

Les sujets se trouvaient face à l'enceinte, de façon à ce que leurs oreilles se situent dans une zone de l'espace dans laquelle le niveau des voyelles a été mesuré à 85 dB SPL. Un écran était disposé à droite de l'enceinte. Les sujets devaient donner oralement leurs réponses à l'expérimentateur, et contrôler sur l'écran que les réponses inscrites par ce dernier étaient correctes. Tous les sujets ont rapporté bien entendre les voyelles.

Les sujets normo-entendants (NE-30 et NE+30) ont participé à 2 séances de 2 blocs. Les malentendants (ME+30) ont participé à 2 séances d'un bloc chacune. L'une des séances était réalisée avec prothèse, l'autre sans prothèse.

5.1.3 Résultats et discussion

Les résultats des normo-entendants sont représentés figure 5.3. Les scores ont été analysés à l'aide d'une ANOVA avec $F_{0(2)}$ et le numéro du bloc comme facteurs répétés, et l'âge comme facteur non répété. Cette analyse a mis en évidence un effet significatif de $F_{0(2)}$ $[F(9,54) = 56,78,\ p < 0,0001]$, du numéro du bloc $[F(3,18) = 3,21,\ p < 0,05]$ ainsi que de l'âge $[F(1,6) = 192,38,\ p < 0,0001]$. Toutes les interactions se sont révélées significatives. La même ANOVA sur les réponses subjectives a révélé un fort effet de $F_{0(2)}$ $[F(9,54) = 31,24,\ p < 0,0001]$, mais ni de l'âge $[F(1,6) = 0,09,\ p = 0,77]$, ni du numéro de bloc $[F(3,18) = 0,51,\ p = 0,68]$.

Les résultats pour les malentendants sont présentés figure 5.4. Dans le groupe des malentendants (ME+30), une ANOVA similaire montre un effet significatif du $F_{0(2)}$ $[F(9,45) = 3,60, p < 0,005]$, mais un effet non significatif de la présence de la prothèse $[F(1,5) = 0,007,\ p = 0,30]$.

Le très fort effet de $F_{0(2)}$ sur les scores dans les deux groupes est cohérent avec les résultats présentés au chapitre 2. De plus, le profil moins accidenté

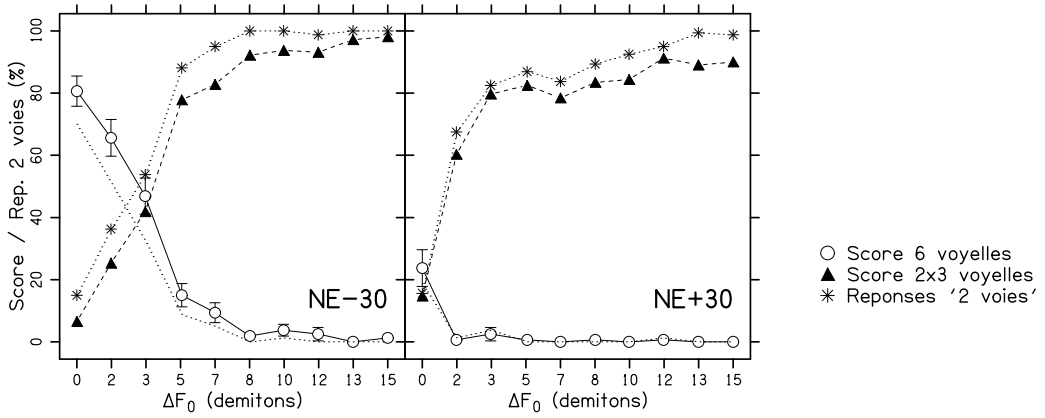

FIG. 5.3 – Scores en fonction de la différence de F_0 entre les deux voies. Les cercles vides et le trait continu représentent les scores dans la tâche principale (ACROSS, donner les 6 voyelles dans l'ordre) en moyenne sur les 4 blocs. Les pointillés sans symbole représentent la moyenne sur les deux premiers blocs seulement. Les triangles noirs et le trait tireté représentent la proportion de réponses justes pour les deux groupes de 3 voyelles (WITHIN). Les astérisques représentent la proportion de réponses "2 voies" dans la tâche supplémentaire. Le *cadre de gauche* représente les résultats pour le groupe de normo-entendants de moins de 30 ans (NE-30). Le *cadre de droite* représente les résultats pour le groupe de normo-entendants de plus de 30 ans (NE+30).

des résultats et le score de 80% dans la condition $\Delta F_0 = 0$ montrent que l'ajout des 5 ms de silence a réduit l'effet de la ségrégation sur la base de la structure formantique. Cependant, la grande variabilité intersujets, et le fort effet de l'âge, nous ont conduit à séparer les sujets normo-entendants en 2 classes d'âge : plus et moins de 30 ans. Les deux groupes ont des résultats très clairement différents, comme le montre la figure 5.3. Les sujets du groupe NE+30 ont obtenu des scores significativement plus faibles que les sujets du groupe NE-30.

L'effet du numéro de bloc démontre un apprentissage de la part des sujets. Les sujets ont eu plus de réponses justes lors du quatrième test (16,9% de réponses justes en moyenne sur les 8 sujets) que lors du premier (7,6% de réponses justes). L'absence d'effet du numéro de bloc sur la tâche subjective suggère qu'il s'agit d'un apprentissage procédural, en particulier pour la tâche d'ordre, mais que les performances de ségrégation restent identiques. Pour pouvoir comparer le groupe ME+30 avec le groupe NE+30, nous avons utilisé

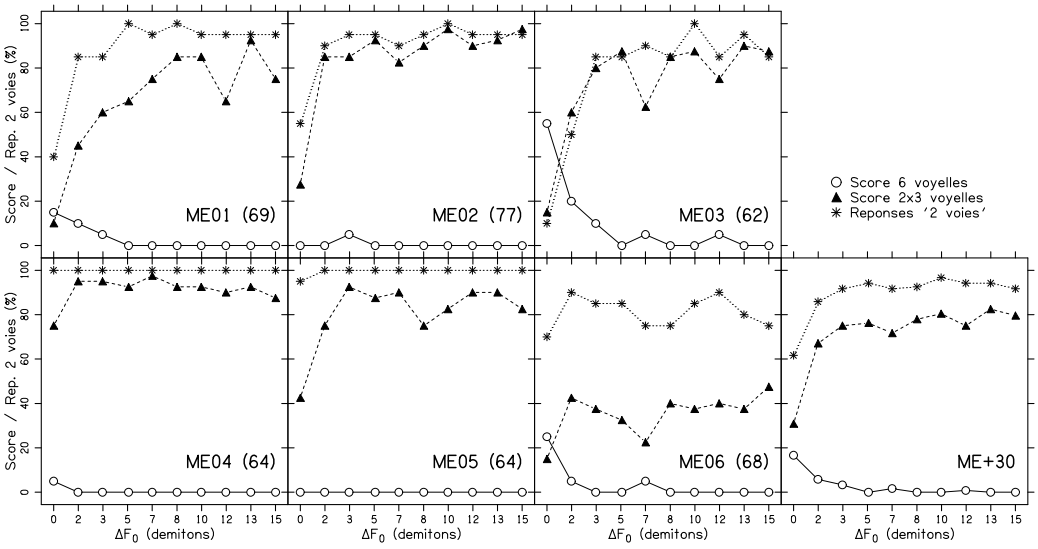

FIG. 5.4 – Comme la figure 5.3 sauf que seule la moyenne sur les deux blocs (avec et sans prothèse) est donnée. Les résultats individuels pour chaque malentendant sont donnés, et le dernier cadre représente les résultats moyens pour ce groupe (ME+30).

les scores moyens des NE+30 sur les deux premières séances. L'entraînement pour ces groupes est alors équivalent.

Lorsque l'on compare les différents groupes, il apparaît que l'effet de la perte auditive n'est pas significatif lorsque l'on compare des classes d'âge équivalentes. Ainsi, aucune différence significative n'apparaît entre le groupe NE+30 et le groupe ME+30 [$F(1,8) = 0,007$, $p = 0,94$]. Dans notre expérience, au-delà de 30 ans, les performances chutent et les normo-entendants ont des performances tout à fait similaires à celles des malentendants. Les malentendants étant tous âgés de plus de 30 ans, on peut estimer que l'effet de la surdité est masqué par l'effet de l'âge. La réduction des performances liée à l'âge, pour ce type de tâche a été décrite par Trainor et Trehub (1989) et Andrews, Dowling, Bartlett, et Halpern (1998) dans le cas de sons purs. En revanche, les performances de ségrégation sont généralement conservées avec l'âge (par exemple, Alain, Ogawa, et Woods, 1996). Ne parvenant pas à percevoir les 6 voyelles dans l'ordre pour effectuer la tâche principale, les sujets ont effectué la tâche subsidiaire et donné deux groupes de 3 voyelles formant des séquences plus lentes. Étant donné leur déficit de perception de

l'ordre, pour pouvoir donner cette réponse, les sujets âgés devaient ségréger les séquences, et le score WITHIN permet alors probablement d'estimer le seuil de fission. Cependant, le seuil de fission ne doit se situer que vers quelques demitons et les ΔF_0 mis en œuvre dans cette expérience ne permettent pas de mettre en évidence une différence entre les groupes NE+30 et ME+30. La puissance de cet effet de l'âge dans notre tâche rend donc notre tâche impossible, et nous empêche donc de mesurer l'effet d'une perte presbyacousique sur la ségrégation de séquences de voyelles. Néanmoins, ce résultat suggère que les problèmes cognitifs liés à l'âge pourraient potentiellement avoir un effet plus dégradant que la perte auditive sur l'intelligibilité de la parole dans le bruit. Une nouvelle expérience impliquant des malentendants jeunes a été réalisée.

5.2 Expérience 2 : Malentendants jeunes

Cette seconde expérience est rigoureusement identique à l'Experiment 1 du chapitre 2 mais implique des sujets malentendants de moins de 30 ans. Elle s'est déroulée au Laboratoire ENTENDRE Xavier Debruille (Audioprothésiste D.E., Reims).

5.2.1 Sujets

Cette seconde expérience a impliqué 5 sujets malentendants âgés de 19 à 24 ans (ME-30) de langue maternelle française. Les pertes étaient neurosensorielles, postlinguales, d'origine génétique. Les audiogrammes des sujets sont donnés figure 5.5 et montrent des pertes plus sévères que celles constatées dans le groupe ME+30 (sauf pour MEj04). Ils étaient tous appareillés, cependant, un des sujets (MEj04) ne portait pas quotidiennement ses prothèses et était malvoyant. L'écran de test a donc été agrandi afin qu'il puisse donner ses réponses.

5.2.2 Stimuli et matériel

Les stimuli et le matériel utilisés étaient les mêmes que dans l'Experiment 1 du chapitre 2. La tâche subjective (1 voix/2 voix) a cependant

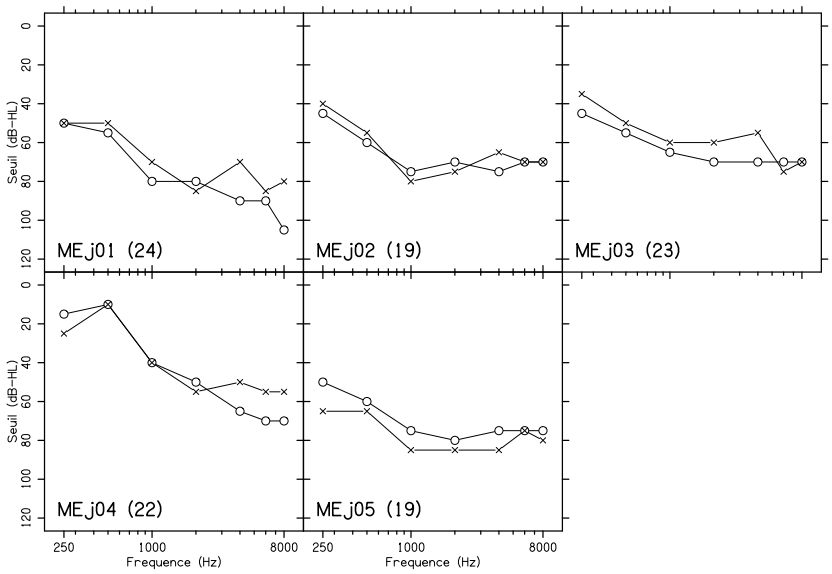

FIG. 5.5 – Audiogrammes des sujets. Les cercles (∘) représentent l'oreille droite, et les croix (×) l'oreille gauche. L'âge des sujets est indiqué entre parenthèses.

été ajoutée. Les stimuli étaient présentés *via* une carte son Digigram Vx-Pocket 440 et casque Sennheiser HD-250 Linear II. La chaîne de stimulation était calibrée de façon à ce que les stimuli aient un niveau de 85 dB-SPL sur un sonomètre Larson Davis 824 (ANSI Type 1), et à travers une oreille artificielle Larson Davis AEC101 munie des coupleurs adéquats (IEC 318, ANSI S3.7–1995). Tous les sujets ont rapporté bien entendre les voyelles.

5.2.3 Résultats et discussion

Les résultats individuels et moyens sont tracés figure 5.6. Les résultats ne s'étant pas révélés dépendants du tempo, ils ont été moyennés sur les deux conditions. Contrairement à notre hypothèse, les sujets du groupe ME-30 ne parviennent pas à donner les 6 voyelles dans l'ordre et obtiennent des scores très faibles par rapport à ceux des normo-entendants (voir chapitre 2). La forme des courbes de réponses est aussi très différente de celle du groupe ME+30 (sauf pour MEj04).

Étant donné les faibles scores WITHIN (2×3 voyelles) pour 4 des 5 sujets, et bien que les sujets aient rapporté bien percevoir les voyelles, nous avons

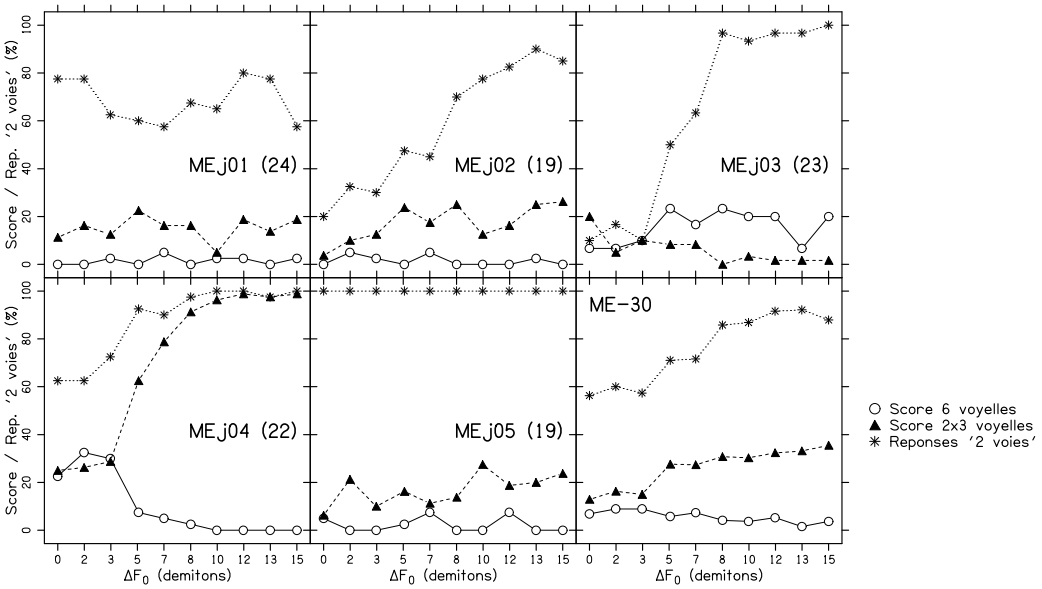

FIG. 5.6 – Scores en fonction de la différence de F_0 entre les deux voies, individuels et moyens (dernier cadre) pour le groupe ME-30. Les cercles vides et le trait continu représentent les scores dans la tâche principale (ACROSS, donner les 6 voyelles dans l'ordre) en moyenne sur les 4 blocs. Les triangles noirs et le trait tireté représente la proportion de réponses justes pour les deux groupes de 3 voyelles (WITHIN). Les astérisques représentent la proportion de réponses "2 voies" dans la tâche supplémentaire.

voulu vérifier *a posteriori* leurs performances d'identification des voyelles isolées. Nous n'avons pu obtenir ces résultats que pour trois des sujets (table 5.1). Le sujet MEj03 a des scores d'identification quasiment parfaits et malgré tout, il présente de grandes difficultés à donner les 6 voyelles dans l'ordre et ne parvient pas à redonner les groupes de 3 voyelles lorsqu'il rapporte percevoir deux flux. Les résultats de ce sujet montrent néanmoins des scores ACROSS relativement indépendants de ΔF_0 et supérieurs à la chance ($< 1\%$). Le sujet MEj04 a lui aussi de bonnes performances d'identification des voyelles isolées, et ses bons scores WITHIN montrent qu'il parvient à identifier les voyelles dans les séquences. La forme de sa courbe de réponses rappelle celle des sujets des groupes NE+30 et ME+30 de l'expérience 1. Il faut noter que ses seuils auditifs sont du même niveau que ceux constatés dans le groupe ME+30. Le niveau de ses performances laisse à penser que la tâche était beaucoup plus dure pour lui que pour les normo-entendants

TAB. 5.1 – Pourcentages d'identification correcte de chacune des 6 voyelles pour trois sujets du groupe ME-30. La dernière colonne contient le score moyenné sur les voyelles.

Sujet	/a/	/e/	/i/	/ɔ/	/u/	/y/	*M*
MEj03	100,0	97,5	100,0	97,5	100,0	100,0	99,2
MEj04	100,0	95,0	87,5	92,5	100,0	95,0	95,0
MEj05	85,0	72,5	52,5	80,0	60,0	55,0	67,5

jeunes. Enfin, le sujet MEj05 a des scores d'identification des voyelles isolées assez bas. Étant donné la difficulté de notre tâche, une identification même faiblement réduite des voyelles a un effet dramatique sur les scores.

Ces résultats montrent de très grandes différences individuelles qui peuvent s'expliquer en partie par la sévérité de la perte auditive et la capacité des sujets à identifier les voyelles isolées. Cependant, même lorsque l'identification est bonne (MEj03 et MEj04), les performances des sujets sont très différentes. Ces résultats suggèrent donc que la perte auditive, si elle semble parfois réduire l'effet de ΔF_0 sur le streaming (MEj03), peut aussi avoir un effet délétère dans notre tâche.

5.3 Discussion

Les résultats des normo-entendants de moins de 30 ans, en champ libre (expérience 1) ou au casque (chapitre 2), sont cohérents avec les résultats obtenus dans les précédentes expériences décrites dans la littérature (notamment, Nooteboom *et al.*, 1978).

L'expérience 1 a montré un fort effet de l'âge dans la tâche imposée aux sujets, dominant sur l'effet de la perte auditive. Il semble que cet effet de l'âge ne porte pas directement sur les performances de ségrégation des sujets âgés, mais sur leur capacité à percevoir l'ordre temporel d'éléments sonores de courte durée. Trainor et Trehub (1989) et Andrews *et al.* (1998) ont en effet montré que les performances de perception de l'ordre temporel décroissaient avec l'âge. Les performances de streaming ne semblent en revanche pas affectées par l'âge (Alain *et al.*, 1996). Il s'avère que l'âge limite qui ressort de nos résultats était 30 ans. Étant donné le nombre de sujets, cette valeur

ne peut avoir de caractère normatif. Cependant, ces résultats indiquent que cette tâche impose d'être très rigoureux sur l'âge des sujets. De nouvelles investigations seraient nécessaires pour déterminer l'impact de cet effet sur l'analyse des scènes auditives. Il est en effet possible que les performances d'analyse soient altérées avec l'âge de façon indépendante d'une perte auditive, en raison de phénomènes plus cognitifs que sensoriels. On a peiné, jusqu'à maintenant, à mettre en évidence ces effets cognitifs (Richie *et al.*, 2003), et cette tâche, combinée à d'autres mesures, pourrait permettre d'éclaircir ce point.

Les résultats obtenus pour les malentendants jeunes dans l'expérience 2 sont déroutants vis-à-vis de notre hypothèse. Les données récoltées ne permettent que de spéculer sur les causes possibles de la dégradation des performances observées. La tâche de perception de l'ordre, contrairement à la perception des voyelles isolées, induit plusieurs types de masquage : du masquage informationnel et du *forward-masking*. Le processus d'identification d'une voyelle nécessite un certain temps. On peut imaginer que pendant ce temps, l'identification de la voyelle suivante est rendue plus difficile, occasionnant ainsi du masquage informationnel. Si le sujet donne une réponse WITHIN, il devient moins sensible au masquage informationnel puisque le tempo de séquence est deux fois plus lent. Il dispose donc de plus de temps pour identifier la voyelle. Cependant, les scores WITHIN du sujet MEj03 restent faibles, même lorsqu'il rapporte avoir perçu deux voix. Le forward-masking est un masquage énergétique. En raison des effets de seuil et de recrutement de sonie, il est possible que les malentendants souffrent plus du forward-masking que les normo-entendants, à niveau sonore égal (Glasberg, Moore, et Bacon, 1987). L'effet de masquage d'un bruit filtré sur un son pur ne décroît que d'environ 25 dB en 80 ms chez les malentendants, contre 60 dB chez les normo-entendants. Le forward-masking pourrait donc réduire l'intelligibilité de voyelles, même lorsqu'elles sont dans des flux différents. Cependant, il est probable que la durée des voyelles aurait alors un effet important sur ce masquage, ce qui n'a pas été constaté. La faiblesse des scores observés n'autorise néanmoins pas une grande variabilité, et l'effet du tempo est peut-être simplement écrasé dans l'effet de saturation des scores. D'autres hypothèses sont discutées dans le chapitre 6.

Chapitre 6

Discussion, perspectives et conclusions

Dans cette dernière partie, le paradigme de l'ordre des séquences de voyelles utilisé dans les quatre études présentées est d'abord discuté. Puis une vue d'ensemble des résultats est proposée. Des hypothèses sont avancées pour apporter une explication aux résultats parfois non conformes à nos prédictions. Enfin, des perspectives à donner à ces travaux sont énoncées.

6.1 Paradigme de l'ordre des voyelles

Streaming ou discrimination

Dans certaines études, le streaming a été confondu avec de la discrimination (l'exemple le plus récent est probablement Cooper et Roberts, 2007). Lorsqu'une tâche subjective est employée et qu'il est simplement demandé aux sujets s'ils ont perçu 1 flux ou 2, il est possible que les sujets cherchent simplement à savoir s'ils arrivent à distinguer un son A d'un son B. C'est pour cette raison que le motif ABA–ABA– est généralement préféré au motif ABAB dans la littérature. Le fait de demander au sujet s'il perçoit ou non un rythme de galop (ce qui correspond à un seul flux) rend cette tâche un peu plus objective. Cependant, ces tâches subjectives n'offrent pas de garantie de permettre d'observer effectivement un phénomène de streaming. Il convient donc de vérifier que le phénomène observé présente bien les caractéristiques du streaming : une dépendance au tempo, une durée de construction des

flux de plusieurs secondes, et surtout que les flux créés contiennent bien les éléments attendus.

Dans le cas de la parole, demander au sujet s'il a entendu un ou deux locuteurs dans une séquence de voyelles est déjà une tâche différente que de demander le nombre de flux dans une séquence de sons complexes harmoniques ABAB. Dans un flux de parole écologique, l'intonation (ou la prosodie) produisent des fluctuations de hauteur au cours du temps. Les alternances de hauteur présentes dans les séquences de 6 voyelles utilisées dans notre paradigme peuvent donc éventuellement être interprétées comme une ligne prosodique, et donc attribuées à un seul locuteur. Cependant, si la différence de hauteur devient trop grande, il deviendra improbable pour l'auditeur qu'un seul locuteur ait pu produire naturellement de tels changements. La tâche consistant simplement à donner le nombre de locuteurs implique donc soit le streaming, soit la détection de la probabilité qu'un seul locuteur puisse produire les changements de hauteur perçus. Au cours de nos investigations, cette tâche n'a été utilisée que de façon subsidiaire pour compléter la tâche objective. Dans ce cas, les résultats ont montré que le jugement subjectif calquait très bien le résultat objectif (voir figure 5.3, p. 129). Néanmoins, la tâche subjective, demandée après la tâche objective, a certainement été fortement influencée par celle-ci.

Pour s'assurer que la tâche objective permet bien d'observer un phénomène de streaming, nous avons, dans l'Experiment 1 du chapitre 2, fait varier le tempo. Lorsque le tempo passait de 5,7 voy./s à 7,4 voy./s, les scores baissaient de plus 10%. Il pourrait être argumenté que le tempo augmentant, la tâche était rendue plus difficile pour d'autres raisons que le streaming, ce qui expliquerait bien la baisse des scores. Par exemple, si le temps nécessaire à l'identification d'une voyelle est constant, la charge cognitive dédiée à la tâche augmente vraisemblablement de façon significative quand les séquences sont accélérées. Cependant, lorsque l'on observe les réponses WITHIN, c'est-à-dire les flux formés par la ségrégation, il apparaît qu'augmenter le tempo conduit bien à une augmentation du nombre de réponses justes de ce type. Ceci signifie bien qu'augmenter le tempo rend plus difficile la résistance à la ségrégation.

Le troisième point à vérifier pour s'assurer que nous avons bien affaire, dans ce paradigme, à un phénomène de streaming, est la durée de la con-

struction des flux. Dans notre paradigme, nous avons réservé 5 s à cette construction en empêchant le sujet de répondre pendant ce laps de temps au début de chaque séquence. Nous n'avons donc pas réalisé d'expérience visant à attester de cet effet de construction. Cependant, il apparaît très clairement qu'à la première présentation de la séquence de 6 voyelles, cette dernière est perçue comme un seul flux. Les sujets ont souvent rapporté leur frustration de ne pouvoir répondre durant les 5 premières secondes alors qu'à cet instant de la présentation, ils avaient tendance à toujours percevoir les 6 voyelles dans l'ordre. Tout concorde donc pour indiquer que la méthode employée permet d'observer le phénomène de streaming. De plus, la forte dépendance au tempo indique que c'est bien le profil du seuil de cohérence qui est observé.

Comparaison avec d'autres méthodes

Le rôle de l'attention dans la formation de flux auditifs est encore flou malgré les récentes études qui y sont consacrées (Sussman, Horváth, Winkler, et Orr, 2007; Snyder, Alain, et Picton, 2006; Carlyon, Cusack, Foxton, et Robertson, 2001; Macken, Tremblay, Houghton, Nicholls, et Jones, 2003; Cusack, Deeks, Aikman, et Carlyon, 2004). Pour réduire l'influence de l'attention dans les tâches de streaming, il faut donner au sujet une tâche qui requiert son attention durant toute la durée du stimulus (notamment pendant la phase de construction), et orienter son effort vers la fusion ou la fission. Les tâches subjectives de jugement du nombre de flux sont donc à éviter.

Il existe d'autres méthodes objectives pour observer le streaming. Il est possible d'intercaler deux mélodies et de demander au sujet s'il reconnaît l'une ou l'autre des mélodies (par exemple Cusack et Roberts, 2000). Cependant, cette méthode impose des écarts de hauteur importants, au moins de l'ordre de grandeur de l'amplitude des variations de hauteur dans chacune des mélodies. En outre, une telle tâche conduit à mesurer le seuil de fission et ne reflète donc pas un processus automatique.

Une autre tâche très souvent employée consiste à détecter un changement de rythme dans la séquence (par exemple Roberts *et al.*, 2002). Au début d'une séquence A B A B A, les sons B sont placés à égale distance avant et après les sons A. Puis les sons B se décalent progressivement pour se rapprocher du son A suivant : A B A B A. Le tempo étant deux fois plus

lent dans chacun des flux créés, il est beaucoup plus difficile d'y détecter un changement de rythme que dans la séquence fusionnée. En mesurant le seuil de détection d'un changement de rythme, il est donc possible d'obtenir une mesure objective très fiable du profil du seuil de cohérence temporelle. Comme notre méthode, cette mesure a la particularité de conduire à de meilleures performances pour les sujets présentant un déficit de ségrégation.

Cette méthode présente néanmoins le défaut d'être très lourde puisqu'il faut effectuer une mesure de seuil par condition à tester, ce qui peut être rédhibitoire pour tester des personnes malentendantes. Par ailleurs, dans cette méthode, l'identification des éléments n'est pas nécessaire. Une telle expérience réalisée avec des séquences de voyelles, en plus de présenter des difficultés méthodologiques (les attaques des voyelles différant d'une voyelle à l'autre), ne permettrait pas de s'assurer que les voyelles sont identifiées et donc traitées comme de la parole. Le sujet pourrait potentiellement ignorer la structure formantique et se focaliser uniquement sur la hauteur ou le rythme des événements sonores. Dans notre tâche d'ordre, pour donner leur réponse, les sujets sont obligés de traiter intégralement la voyelle jusqu'à la comparer avec une représentation graphique à l'écran. Ceci implique que toutes les caractéristiques acoustiques de la voyelle nécessaires à son identification sont traitées, en plus de la hauteur.

Limites de la méthode

La principale limitation dont a souffert notre tâche d'ordre a été mise en évidence dans la dernière étude (chapitre 5). Cette tâche ne mesure pas seulement les performances de streaming, mais rend aussi compte de la capacité des sujets à percevoir l'ordre des éléments constituant une séquence. Cette performance peut être altérée, notamment chez les personnes âgées. Néanmoins il est difficile d'estimer comment les performances de perception de l'ordre de séquences de voyelles sont reliées à la perception de voix concurrentes. L'âge a un effet important sur les performances de perception de la parole dans le bruit (Glasberg et Moore, 1989; Mackersie *et al.*, 2001). Il semble que les difficultés des sujets presbyacousiques dans les situations de cocktail party ne soient pas dues qu'à leur perte auditive, mais que des effets plus cognitifs (comme, par exemple, au niveau de la mémoire de travail)

soient impliqués (George, Zekveld, Kramer, Goverts, Festen, et Houtgast, 2007).

L'implication de la mémoire dans notre tâche est aussi une autre source de variabilité possible dans nos résultats. Les séquences sont en effet présentées en boucle pendant que le sujet répond, de façon à réduire l'effet des capacités de mémorisation des sujets. Cependant, les flux auditifs ne se forment pas instantanément dès la première occurrence de la séquence (Bregman, 1978; Anstis et Saida, 1985). Le streaming est initialement biaisé vers la fusion et la ségrégation met quelques secondes à se mettre en place Bregman (1990). Les six voyelles de la première occurrence de la séquence sont donc quasiment systématiquement perçues dans l'ordre. Une stratégie du sujet pourrait donc être de mémoriser cette première occurrence jusqu'à l'apparition de l'interface de réponse. Un moyen de réduire partiellement cet effet serait de faire apparaître graduellement les séquences. Ainsi, l'accumulation d'indices pour la ségrégation pourrait débuter avant que les voyelles ne deviennent intelligibles et ainsi réduire la durée du biais vers la fusion. Une autre possibilité serait de présenter une autre séquence de voyelles au début du stimuli. Cependant, comme le montrent les résultats, la tâche permet effectivement de mesurer les performances de streaming des sujets. Ceci illustre le fait qu'il est extrêmement difficile de garder en mémoire une séquence de six voyelles quand ces mêmes voyelles sont présentées par la suite selon une autre organisation. Plus que des capacités mnésiques, c'est donc la capacité du sujet à faire abstraction de ce qu'il entend qui influence ses performances. Il semble en effet que les sujets dotés d'une grande capacité de concentration obtiennent de meilleures performances. Il pourrait être intéressant d'évaluer cette résistance à un flux de voyelles distracteur conjointement avec les performances de perception de phrases concurrentes d'une part, et avec les performances de résistance à l'effet de son déviant (*irrelevant sound effect*, voir par exemple Ellermeier et Zimmer, 1997).

Enfin, l'emploi de voyelles dans une tâche de streaming était destiné à rendre les observations plus facilement comparables aux performances de séparation de voix concurrentes. Les séquences de voyelles sont cependant encore très différentes de la parole réelle. En particulier, nous avons pu observer de la ségrégation sur la base de la structure formantique. Il est peu probable que ce phénomène intervienne dans la parole réelle où les voyelles

sont connectées par des consonnes qui assurent la continuité des formants. Les trajectoires formantiques sont effectivement issues des trajectoires articulaires qui ne peuvent être discontinues. Dans notre méthode, les transitions formantiques ont été supprimées et les distances formantiques contrôlées de façon à limiter l'implication de la structure formantique dans le streaming. L'absence de consonnes permettait aussi de ne pas créer de mots et ainsi de ne pas avoir à contrôler un éventuel contenu sémantique. Cependant, dans la langue, en supprimant les consonnes, toutes les séquences de voyelles ne sont pas équiprobables. Bien que le streaming observé soit irrépressible et réputé d'origine périphérique (Bregman, 1990), il est possible que la probabilité de co-occurrence des voyelles ou le contenu sémantique puissent jouer un rôle dans cette tâche.

6.2 Interprétation spéculative des résultats

Les résultats des quatre études présentées dans cette thèse montrent qu'une différence de hauteur (F_0) a, sur le streaming de flux de voyelles, un effet continu, relativement linéaire, entre 0 et 15 demitons. Comme cela a été discuté au chapitre 4, cette plage de fréquences est similaire à celle sur laquelle l'effet d'une différence de hauteur est profitable pour la séparation de voix concurrentes (Bird et Darwin, 1998). Ceci signifie que le streaming peut effectivement être impliqué dans la séparation de voix concurrentes. La corrélation mise en évidence au chapitre 4 suggère en outre que les performances de streaming rendent assez bien compte du profit que peuvent tirer les sujets d'une différence de hauteur dans une tâche de séparation de voix concurrentes. Le niveau moyen de performance de séparation de voix concurrentes serait donc piloté par la sélectivité fréquentielle (comme le proposent Glasberg et Moore, 1989), tandis que le bénéfice d'une différence de hauteur serait relié aux performances de streaming sur la base de la hauteur.

De récentes investigations ont montré que des pertes modérées (seuils auditifs entre 20 et 60 dB-HL) pouvaient conduire à une réduction de la capacité à exploiter la structure fine temporelle de sons complexes harmoniques pour en déduire leur hauteur (Hopkins et Moore, 2007). La perception de la structure fine semble également impliquée dans la capacité des sujets à tirer profit de trous temporels dans un bruit masquant (Lorenzi *et al.*, 2006). Ces

résultats récents pourraient éclairer différemment les résultats obtenus dans notre tâche de streaming. Les sujets quasi normo-entendants du chapitre 4 et les malentendants du chapitre 5 ont montré des performances très variables dans notre tâche. Dans le chapitre 4, nous avons montré que cette variabilité ne pouvait s'expliquer ni par la sélectivité fréquentielle, ni par les seuils auditifs. Il apparaît donc qu'une autre fonction auditive est impliquée dans la ségrégation séquentielle de voyelles, et donc probablement dans la perception de voix concurrentes. Il semble tentant de spéculer que les capacités de perception de la structure fine pourraient fortement moduler les performances des sujets dans notre paradigme. Les investigations menées jusqu'à maintenant concernent le rôle de la structure fine pour la perception de la hauteur ou pour l'exploitation de trous temporels dans un masque. En revanche, trop peu de connaissances ont été rassemblées sur l'implication de la structure fine dans les tâches de streaming.

Au-delà de la perception de la hauteur, il semble que la manipulation du F_0 puisse induire du streaming même lorsqu'aucune sensation de hauteur n'est perçue. Les résultats présentés au chapitre 3 suggèrent que les implantés cochléaires ne pourront aisément bénéficier d'une différence de F_0 pour séparer deux locuteurs. La périodicité de l'enveloppe temporelle ne semble pas un indice de hauteur facilement exploitable dans la parole. Cependant, ces résultats ont aussi mis en évidence qu'une information spectrale relativement fine et de nature relativement stochastique pouvait induire du streaming.

Contrairement aux expériences habituelles de streaming où seuls deux sons différents (A et B) sont généralement utilisés dans une séquence, les sons composants les séquences de voyelles varient suivant plus d'une dimension acoustique (les positions des différents formants varient suivant les voyelles). Seules deux valeurs de F_0 sont utilisées dans chaque séquence, mais les signaux acoustiques qui forment chaque voyelle sont très différents aussi bien temporellement que spectralement. Les premières expériences (chapitre 2) ont montré que les sujets arrivaient à former des flux auditifs à partir de ces séquences en utilisant la différence de F_0 malgré les variations suivant les autres dimensions, comme lors de la formation de flux de parole. Ceci indique que des objets sonores différents mais ayant le même F_0 peuvent être groupés dans un même flux. L'assertion de Moore et Gockel (2002) selon laquelle tout indice acoustique perceptible peut permettre de séparer un sig-

nal en flux auditifs est donc incomplète. Il semble que malgré l'existence de différences acoustiques contradictoires, le système auditif parvienne à extraire des indices des régularités présentes dans le signal pour en créer les flux auditifs. Ceci expliquerait pourquoi la durée de construction des flux auditifs est si longue (comme observé par Bregman, 1978; Anstis et Saida, 1985). Dans cette hypothèse, les indices spectraux reliés à la hauteur, et mis en évidence au chapitre 3, seraient donc extraits par analyse de la régularité — de la consistance — de cet indice à travers les différentes voyelles, de façon similaire à l'analyse statistique utilisée pour les mettre en évidence. Dans cette hypothèse, la saillance de cet indice, et donc sa propension à induire du streaming, serait définie par la consistance de ces régularités. Ces phénomènes seraient à rapprocher des phénomènes d'amorçage (*priming*) mis en évidence notamment pour le langage et la musique.

Ce phénomène pourrait être pris en compte différement dans les modèles de streaming. Le modèle développé par Elhilali et Shamma (2007) cherche les corrélations à différentes échelles temporelles et spectrales dans un espace multidimensionnel. Ces auteurs ont suggéré que le modèle pourrait être complété par une représentation de la hauteur fondamentale. Avec un tel complément, ce modèle pourrait vraisemblablement reproduire les résultats que nous avons obtenus dans les différentes expériences avec les voyelles intactes. Cependant, il est difficile d'évaluer si ce modèle pourrait permettre d'exploiter des régularités plus complexes. Les modèles implémentés pour exploiter ce genre de régularité sont généralement basés sur des réseaux de neurones comme ceux qui constituent le modèle de Grossberg *et al.* (2004).

6.3 Perspectives

Structure fine temporelle

Les considérations sur la méthode et l'interprétation spéculative des résultats amènent de nombreuses questions et esquissent de nouvelles pistes de recherche. Tous les résultats obtenus au cours des investigations présentées dans cette thèse suggèrent que les représentations de la hauteur fondamentale par des indices spectraux liés à la tonotopie et par la périodicité de l'enveloppe temporelle ne permettent pas d'interpréter tous les résultats obtenus. Il semble qu'un autre mécanisme détermine les performances de ségrégation sur la

base de la hauteur. Les avancées récentes sur la compréhension du rôle de la structure fine temporelle dans la perception de la parole dans le bruit et de la hauteur, font de cet indice un bon candidat qui mérite d'être étudié plus en profondeur. En particulier, l'implication de la structure fine dans le streaming doit être évaluée, et la variabilité de la capacité à utiliser cette structure fine dans la population doit être estimée. Pour y parvenir il faudra se munir d'une définition précise de cette structure fine, ainsi que d'un test psychoacoustique permettant d'évaluer la capacité des sujets à l'exploiter.

Vers la parole

L'objectif des différentes études présentées ici était d'amorcer un rapprochement entre les mécanismes fondateurs de l'analyse des scènes auditives et la perception de la parole dans le bruit. La ségrégation séquentielle avec des signaux de parole a donc été étudiée afin de pouvoir la comparer à la ségrégation simultanée. Cependant le chemin à parcourir pour intégrer ces deux mécanismes dans une compréhension plus globale du phénomène de la perception dans le bruit est encore long. Cette section propose quelques pistes pour prolonger nos travaux dans cette direction.

Les stimuli utilisés ici étaient des voyelles. Dans la littérature, les signaux de parole les plus complexes qui ont été utilisés sont des syllabes (par exemple Darwin et Bethell-Fox, 1977). Afin de se rapprocher de la parole réelle, des syllabes devraient être utilisées dans notre tâche. Il pourrait même être envisageable de réaliser des phrases imbriquées dans lesquelles la perception d'un seul flux conduit à une phrase, et la perception de deux flux conduit à deux autres phrases. Par exemple :

le		pin		vit		du		vent		frais	
	la		fu-		sée		o-		sci-		le

La perception d'un seuil flux conduit à "Le lapin fut visé du auvent si frèle", et la perception de deux flux conduit à "Le pin vit du vent frais" et "La fusée oscille". Bien qu'il soit probablement difficile de constituer un corpus important de phrases de ce type, celles-ci pourraient permettre de mieux étudier l'effet de la hauteur sur le streaming dans un contexte plus écologique.

Le fait d'utiliser des stimuli différents de la plupart des études qui peuplent la littérature fait apparaître de nouveaux indices de ségrégation pertinents. Ainsi Tsuzaki *et al.* (2007) ont montré qu'une différence de longueur

du tractus vocal pouvait induire du streaming. Cette observation n'est possible qu'avec des sons complexes harmoniques qui contiennent une structure formantique élaborée. La classification des indices conduisant à la ségrégation proposée par Hartmann et Johnson (1991) serait donc à revoir dans le cas de la parole. Les résultats du chapitre 3 ont mis en évidence l'existence d'indices spectraux liés à la hauteur. Deux aspects sont particulièrement intéressants dans ces indices. D'abord, ils agissent comme un filtre fréquentiel, c'est-à-dire qu'ils modifient le spectre d'un son existant. Pour percevoir cette modification, il faut qu'il y ait de l'énergie à la fréquence où cette modification intervient. Dans les situations réelles, d'autres indices agissent de façon similaire : les indices spectraux de localisation (*head-related transfer function*, HRTF). Les indices de localisation sont réputés plutôt faibles pour la ségrégation séquentielle. Ceci provient vraisemblablement du fait que les multiples réverbérations qui interviennent quand un son se propage rendent sa localisation peu fiable. Néanmoins, l'implication de ces indices pour le streaming de signaux de parole, et en particulier chez les implantés cochléaires, pourrait être ré-évaluée.

Une autre dimension importante de la parole est qu'elle nous est familière. L'oreille humaine est spécialisée dans le traitement du langage. Il est connu qu'une voix familière est plus facile à séparer d'un bruit de fond qu'une voix inconnue. Dans ce cas, la ségrégation repose sur une connaissance *a priori* des caractéristiques de la voix. Il serait intéressant de tester si la familiarité d'une voix peut être un indice de streaming irrépressible. De la même façon, il est possible de réaliser des voix dont l'existence dans la nature est improbable, par exemple en utilisant certaines combinaisons de longueur de tractus vocal et de hauteur fondamentale. La crédibilité d'une voix pourrait aussi être un indice important pour le système auditif.

Enfin, en plus d'étudier l'importance de tous ces indices spécifiques à la parole, le rapprochement entre ségrégation simultanée et ségrégation séquentielle devrait être opéré. Il est possible de faire évoluer une séquence ABABAB purement séquentielle vers une séquence dans laquelle les A et les B sont simultanés. La principale difficulté méthodologique réside dans le fait que la tâche usuelle de streaming consiste à fusionner la séquence tandis que la tâche couramment employée pour observer la ségrégation simultanée consiste à séparer les deux voyelles simultanées. Il serait néanmoins possible de

contourner cet écueil en utilisant des tâches subjectives. Une autre méthode pourrait consister à placer des distracteurs simultanés dans une tâche objective de streaming. Enfin, les séquences utilisées pour le streaming sont souvent isochrones, ce qui est loin d'être écologique. L'anisochronie des évènements sonores d'une séquence pourrait aussi avoir un rôle dans cette interaction. Une autre approche pourrait consister à mesurer les performances de perception de deux phrases concurrentes, et de les comparer aux performances de ségrégations simultanée et séquentielle en utilisant des stimuli extraits des deux phrases de départ. Les signaux pour la ségrégation simultanée ne seraient composés que des segments du signal où les deux voix présentent des niveaux comparables. Les signaux pour la ségrégation séquentielle seraient composés alternativement de segments des deux phrases. Cette approche pourrait permettre de comparer l'implication de chacun de ces mécanismes dans la perception de voix concurrentes.

Références

Alain, C., Ogawa, K. H., et Woods, D. L. (**1996**). "Aging and the segregation of auditory stimulus sequences", J. Gerontol. B. Psychol. Sci. Soc. Sci. **51**, 91–3.

American National Standard Institute (**1995**). *ANSI S3.7-1995 (R2003), Methods for coupler calibration of earphones*, New-York.

American National Standard Institute (**2004**). *ANSI S3.6-2004, Specifications for audiometers*, New-York.

Andrews, M. W., Dowling, W. J., Bartlett, J. C., et Halpern, A. R. (**1998**). "Identification of speeded and slowed familiar melodies by younger, middle-aged, and older musicians and nonmusicians", Psychol. Aging **13**, 462–71.

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

Arehart, K. H., King, C. A., et McLean-Mudgett, K. S. (**1997**). "Role of fundamental frequency differences in the perceptual separation of competing vowel sounds by listeners with normal hearing and listeners with hearing loss", J. Speech Lang. Hear. Res. **40**, 1434–1444.

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

Bacon, S. P. (**2004**). *Compression : From Cochlea to Cochlear Implants* (Springer-Verlag, New York).

Baer, T. et Moore, B. C. J. (**1993**). "Effects of spectral smearing on the intelligibility of sentences in noise", J. Acoust. Soc. Am. **94**, 1229–1241.

Baer, T. et Moore, B. C. J. (**1994**). "Effects of spectral smearing on the intelligibility of sentences in the presence of interfering speech", J. Acoust. Soc. Am. **95**, 2277–2280.

Beauvois, M. W. et Meddis, R. (**1996**). "Computer simulation of auditory stream segregation in alternating-tone sequences", J. Acoust. Soc. Am. **99**, 2270–2280.

Berkes, P. et Zito, T. (**2007**). "Modular toolkit for data processing (version 2.1)", `http://mdp-toolkit.sourceforge.net`.

Bernstein, J. G. W. et Oxenham, A. J. (**2006**). “The relationship between frequency selectivity and pitch discrimination : sensorineural hearing loss”, J. Acoust. Soc. Am. **120**, 3929–3945.

Bird, J. et Darwin, C. J. (**1998**). “Effects of a difference in fundamental frequency in separating two sentences”, *in Psychophysical and physiological advances in hearing*, edité par A. R. Palmer, A. Rees, A. Q. Summerfield, et R. Meddis (Whurr, London).

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

Bregman, A. S. (**1990**). *Auditory Scene Analysis : The Perceptual Organization of sound* (The MIT Press, Massachusetts, USA).

Bregman, A. S. et Campbell, J. (**1971**). “Primary auditory stream segregation and perception of order in rapid sequences of tones”, J. Exp. Psychol. **89**, 244–249.

Bregman, A. S., Liao, C., et Levitan, R. (**1990**). “Auditory grouping based on fundamental frequency and formant peak frequency”, Can. J. Psychol. **44**, 400–413.

Broadbent, D. E. et Ladefoged, P. (**1957**). “On the fusion of sounds reaching different sense organs”, J. Acoust. Soc. Am. **29**, 708–710.

Brokx, J. P. L. et Nooteboom, S. G. (**1982**). “Intonation and the perceptual separation of simultaneous voices”, J. Phonetics **10**, 23.

Bronkhorst, A. W. (**2000**). “The cocktail party phenomenon : A review of research on speech intelligibility in multiple-talker conditions”, Acustica **86**, 117–128.

Burns, E. M. et Viemeister, N. F. (**1976**). “Nonspectral pitch”, J. Acoust. Soc. Am. **60**, 863–869.

Burns, E. M. et Viemeister, N. F. (**1981**). “Played-again sam : Further observations on the pitch of amplitude-modulated noise”, J. Acoust. Soc. Am. **70**, 1655–1660.

Byrne, D., Dillon, H., Tran, K., Arlinger, S., Wilbraham, K., Cox, R., Hagerman, B., Hetu, R., Kei, J., Lui, C., Kiessling, J., Kotby, M. N., Nasser, N. H. A., Kholy, W. A. H. E., Nakanishi, Y., Oyer, H., Powell, R., Stephens, D., Meredith, R., Sirimanna, T., Tavartkiladze, G., Frolenkov, G. I., Westerman, S., et Ludvigsen, C. (**1994**). “An international comparison of long-term average speech spectra”, J. Acoust. Soc. Am. **96**, 2108–2120.

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

Carlyon, R. P., Long, C. J., Deeks, J. M., et McKay, C. M. (**2007**). “Concurrent sound segregation in electric and acoustic hearing”, J. Assoc. Res. Otolaryngol. **8**, 119–133.

Carroll, J. et Zeng, F.-G. (**2007**). “Fundamental frequency discrimination and speech perception in noise in cochlear implant simulations”, Hear. Res. **231**, 42–53.

Chatterjee, M., Sarampalis, A., et Oba, S. I. (**2006**). "Auditory stream segregation with cochlear implants : A preliminary report", Hear. Res. **222**, 100–107.

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

Chi, T., Ru, P., et Shamma, S. A. (**2005**). "Multiresolution spectrotemporal analysis of complex sounds", J. Acoust. Soc. Am. **118**, 887–906.

Cohen, M. A., Grossberg, S., et Wyse, L. L. (**1995**). "A spectral network model of pitch perception", J. Acoust. Soc. Am. **98**, 862–879.

Cooper, H. R. et Roberts, B. (**2007**). "Auditory stream segregation of tone sequences in cochlear implant listeners", Hear. Res. **225**, 11–24.

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

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

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

Darwin, C. J. et Bethell-Fox, C. E. (**1977**). "Pitch continuity and speech source attribution", J. Exp. Psychol. Hum. Percept. Perform. **3**, 665–672.

de Boer, B. (**2000**). "Self-organization in vowel systems", J. Phonetics **28**, 441–465.

de Cheveigné, A. (**1999**). "Waveform interactions and the segregation of concurrent vowels", J. Acoust. Soc. Am. **106**, 2959–72.

de Cheveigné, A. (**2005**). "Pitch perception models", *in Pitch : Neural Coding and Perception*, edité par C. J. Plack, A. J. Oxenham, R. R. Fay, et A. N. Popper (Springer).

Dorman, M. F., Cutting, J. E., et Raphael, L. J. (**1975**). "Perception of temporal order in vowel sequences with and without formant transitions", J. Exp. Psychol. Hum. Percept. Perform. **104**, 147–153.

Dorman, M. F., Loizou, P. C., Fitzke, J., et Tu, Z. (**1998**). "The recognition of sentences in noise by normal-hearing listeners using simulations of cochlear-implant signal processors with 6-20 channels", J. Acoust. Soc. Am. **104**, 3583–3585.

Dorman, M. F., Loizou, P. C., et Rainey, D. (**1997**). "Speech intelligibility as a function of number of channels of stimulation for signal processors using sine-wave and noise-band outputs", J. Acoust. Soc. Am. **102**, 2403–2410.

Elhilali, M. et Shamma, S. (**2007**). "The correlative brain : A stream segregation model", *in Hearing - From Sensory Processing to Perception*, edité par B. Kollmeier, G. Klump, V. Hohmann, U. Langemann, M. Mauermann, S. Uppenkamp, et J. Verhey, chapitre 27, 247 (Springer).

Ellermeier, W. et Zimmer, K. (**1997**). “Individual differences in susceptibility to the "irrelevant speech effect"”, J. Acoust. Soc. Am. **102**, 2191–2199.

Festen, J. M. et Plomp, R. (**1983**). “Relations between auditory functions in impaired hearing”, J. Acoust. Soc. Am. **73**, 652–662.

Friesen, L. M., Shannon, R. V., Baskent, D., et Wang, X. (**2001**). “Speech recognition in noise as a function of the number of spectral channels : comparison of acoustic hearing and cochlear implants”, J. Acoust. Soc. Am. **110**, 1150–1163.

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

George, E. L. J., Zekveld, A. A., Kramer, S. E., Goverts, S. T., Festen, J. M., et Houtgast, T. (**2007**). “Auditory and nonauditory factors affecting speech reception in noise by older listeners”, J. Acoust. Soc. Am. **121**, 2362–2375.

Glasberg, B. R. et Moore, B. C. J. (**1986**). “Auditory filter shapes in subjects with unilateral and bilateral cochlear impairments”, J. Acoust. Soc. Am. **79**, 1020–1033.

Glasberg, B. R. et Moore, B. C. J. (**1989**). “Psychoacoustic abilities of subjects with unilateral and bilateral cochlear hearing impairments and their relationship to the ability to understand speech”, Scand. Audiol. Suppl. **32**, 1–25.

Glasberg, B. R. et Moore, B. C. J. (**1990**). “Derivation of auditory filter shapes from notched-noise data”, Hear. Res. **47**, 103–138.

Glasberg, B. R., Moore, B. C. J., et Bacon, S. P. (**1987**). “Gap detection and masking in hearing-impaired and normal-hearing subjects”, J. Acoust. Soc. Am. **81**, 1546–1556.

Green, T., Faulkner, A., Rosen, S., et Macherey, O. (**2005**). “Enhancement of temporal periodicity cues in cochlear implants : effects on prosodic perception and vowel identification”, J. Acoust. Soc. Am. **118**, 375–385.

Greenberg, S., Hollenback, J., et Ellis, D. (**1996**). “Insights into spoken language gleaned from phonetic transcription of the switchboard corpus”, *in Proceedings of the International Conference on Spoken Language Processing*, S24–27.

Grimault, N., Bacon, S. P., et Micheyl, C. (**2002**). “Auditory stream segregation on the basis of amplitude modulation rate”, J. Acoust. Soc. Am. **111**, 1340–1348.

Grimault, N., Micheyl, C., Carlyon, R. P., Arthaud, P., et Collet, L. (**2000**). “Influence of peripheral resolvability on the perceptual segregation of harmonic tones differing in fundamental frequency”, J. Acoust. Soc. Am. **108**, 263–271.

Grimault, N., Micheyl, C., Carlyon, R. P., Arthaud, P., et Collet, L. (**2001**). “Perceptual auditory stream segregation of sequences of complex sounds in subjects with normal and impaired hearing”, Br. J. Audiol. **35**, 173–182.

Grose, J. H. et Hall, J. W. (**1996**). “Perceptual organization of sequential stimuli in listeners with cochlear hearing loss”, J. Speech Hear. Res. **39**, 1149–1158.

Grossberg, S., Govindarajan, K. K., Wyse, L. L., et Cohen, M. A. (**2004**). "ARTSTREAM : a neural network model of auditory scene analysis and source segregation", Neural Netw. **17**, 511–536.

Hanna, T. E. (**1992**). "Discrimination and identification of modulation rate using a noise carrier", J. Acoust. Soc. Am. **91**, 2122–2128.

Hartmann, W. M. et Johnson, D. (**1991**). "Stream segregation and peripheral channeling", Music Percept. **9**, 115–184.

Hawkins, Jr., J. E. et Stevens, S. S. (**1950**). "The masking of pure tones and of speech by white noise", J. Acoust. Soc. Am. **22**, 6–13.

Healy, E. W. et Bacon, S. P. (**2002**). "Across-frequency comparison of temporal speech information by listeners with normal and impaired hearing", J. Speech Lang. Hear. Res. **45**, 1262–75.

Healy, E. W. et Steinbach, H. M. (**2007**). "The effect of smoothing filter slope and spectral frequency on temporal speech information", J. Acoust. Soc. Am. **121**, 1177–1181.

Hirsh, I. J. (**1959**). "Auditory perception of temporal order", J. Acoust. Soc. Am. **31**, 759–767.

Hoen, M., Meunier, F., Grataloup, C.-L., Pellegrino, F., Grimault, N., Perrin, F., Perrot, X., et Collet, L. (**2007**). "Phonetic and lexical interferences in informational masking during speech-in-speech comprehension", Speech Comm. **49**, 905–916.

Hong, R. S. et Turner, C. W. (**2006**). "Pure-tone auditory stream segregation and speech perception in noise in cochlear implant recipients", J. Acoust. Soc. Am. **120**, 360–374.

Hopkins, K. et Moore, B. C. J. (**2007**). "Moderate cochlear hearing loss leads to a reduced ability to use temporal fine structure information", J. Acoust. Soc. Am. **122**, 1055–1068.

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

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

Lackner, J. R. et Goldstein, L. M. (**1974**). "Primary auditory stream segregation of repeated consonant–vowel sequences", J. Acoust. Soc. Am. **56**, 1651–1652.

Laneau, J., Moonen, M., et Wouters, J. (**2006**). "Factors affecting the use of noise-band vocoders as acoustic models for pitch perception in cochlear implants", J. Acoust. Soc. Am. **119**, 491–506.

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

Lorenzi, C., Gilbert, G., Carn, H., Garnier, S., et Moore, B. C. J. (**2006**). "Speech perception problems of the hearing impaired reflect inability to use temporal fine structure", Proc. Natl. Acad. Sci. U.S.A. **103**, 18866–18869.

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

Mackersie, C., Prida, T., et Stiles, D. (**2001**). "The role of sequential stream segregation and frequency selectivity in the perception of simultaneous sentences by listeners with sensorineural hearing loss", J. Speech Lang. Hear. Res. **44**, 19–28.

Mantakas, M., Schwartz, J. L., et Escudier, P. (**1986**). "Modèle de prédiction du 'deuxième formant effectif' F_2' – application à l'étude de la labialité des voyelles avant du français", *in Actes des 15èmes journées d'étude sur la parole*, edité par la Société Française d'Acoustique, 157–161.

McCabe, S. L. et Denham, M. J. (**1997**). "A model of auditory streaming", J. Acoust. Soc. Am. **101**, 1611.

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

Miller, G. A. (**1947**). "The masking of speech", Psychol. Bull. **44**, 105–129.

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

Moore, B. C. J. (**1998**). *Cochlear hearing loss* (Whurr, London).

Moore, B. C. J. et Carlyon, R. P. (**2005**). "Perception of pitch by people with cochlear hearing loss and cochlear implant users", *in Pitch : Neural coding and perception*, edité par C. J. Plack, A. J. Oxenham, R. R. Fay, et A. N. Popper (Springer).

Moore, B. C. J. et Glasberg, B. R. (**1993**). "Simulation of the effects of loudness recruitment and threshold elevation on the intelligibility of speech in quiet and in a background of speech", J. Acoust. Soc. Am. **94**, 2050–2062.

Moore, B. C. J., Glasberg, B. R., et Vickers, D. A. (**1995**). "Simulation of the effects of loudness recruitment on the intelligibility of speech in noise", Br. J. Audiol. **29**, 131–143.

Moore, B. C. J. et Gockel, H. (**2002**). "Factors influencing sequential stream segregation", Acta Acustica united with Acustica **88**, 320–332.

Moore, B. C. J., Johnson, J. S., Clark, T. M., et Pluvinage, V. (**1992**). "Evaluation of a dual-channel full dynamic range compression system for people with sensorineural hearing loss", Ear Hear. **13**, 349–370.

Moore, B. C. J. et Moore, G. A. (**2003**). "Discrimination of the fundamental frequency of complex tones with fixed and shifting spectral envelopes by normally hearing and hearing-impaired subjects", Hear. Res. **182**, 153–163.

Moore, B. C. J. et Peters, R. W. (**1992**). "Pitch discrimination and phase sensitivity in young and elderly subjects and its relationship to frequency selectivity", J. Acoust. Soc. Am. **91**, 2881–2893.

Nooteboom, S. G., Brokx, J. P. L., et de Rooij, J. J. (**1978**). "Contributions of prosody to speech perception", *in Studies in the Perception of Language*, edité par W. J. M. Levelt et G. B. F. d'Arcais, 75–107 (Wiley and Sons, New-York, USA).

Patterson, R. D. (**1976**). "Auditory filter shapes derived with noise stimuli", J. Acoust. Soc. Am. **59**, 640–654.

Patterson, R. D. et Moore, B. C. J. (**1986**). "Auditory filters and excitation patterns as representations of frequency resolution", *in Frequency Selectivity in Hearing*, edité par B. C. Moore (Academic Press, London).

Patterson, R. D., Nimmo-Smith, I., Weber, D. L., et Milroy, R. (**1982**). "The deterioration of hearing with age : frequency selectivity, the critical ratio, the audiogram, and speech threshold", J. Acoust. Soc. Am. **72**, 1788–1803.

Peters, R. W., Moore, B. C., et Baer, T. (**1998**). "Speech reception thresholds in noise with and without spectral and temporal dips for hearing-impaired and normally hearing people", J. Acoust. Soc. Am. **103**, 577–587.

Pressnitzer, D. et Hupé, J.-M. (**2006**). "Temporal dynamics of auditory and visual bistability reveal common principles of perceptual organization", Curr. Biol. **16**, 1351–1357.

Qin, M. K. et Oxenham, A. J. (**2005**). "Effects of envelope-vocoder processing on f0 discrimination and concurrent-vowel identification", Ear Hear. **26**, 451–460.

Remez, R. E., Rubin, P. E., Berns, S. M., Pardo, J. S., et Lang, J. M. (**1994**). "On the perceptual organization of speech", Psychol. Rev. **101**, 129–156.

Richie, C., Kewley-Port, D., et Coughlin, M. (**2003**). "Discrimination and identification of vowels by young, hearing-impaired adults", J. Acoust. Soc. Am. **114**, 2923–33.

Roberts, B., Glasberg, B. R., et Moore, B. C. J. (**2002**). "Primitive stream segregation of tone sequences without differences in fundamental frequency or passband", J. Acoust. Soc. Am. **112**, 2074–2085.

Rogers, C. F., Healy, E. W., et Montgomery, A. A. (**2006**). "Sensitivity to isolated and concurrent intensity and fundamental frequency increments by cochlear implant users under natural listening conditions", J. Acoust. Soc. Am. **119**, 2276–2287.

Rose, M. M. et Moore, B. C. J. (**1997**). "Perceptual grouping of tone sequences by normally hearing and hearing-impaired listeners", J. Acoust. Soc. Am. **102**, 1768–1778.

Rose, M. M. et Moore, B. C. J. (**2005**). "The relationship between stream segregation and frequency discrimination in normally hearing and hearing-impaired subjects", Hear. Res. **204**, 16–28.

Schwartz, J.-L., Boe, L.-J., Vallee, N., et Abry, C. (**1997**). "The dispersion-focalization theory of vowel systems", J. Phonetics **25**, 255–286.

Shannon, R. V. (**1983**). “Multichannel electrical stimulation of the auditory nerve in man. I. basic psychophysics”, Hear. Res. **11**, 157–189.

Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., et Ekelid, M. (**1995**). “Speech recognition with primarily temporal cues”, Science **270**, 303–304.

Singh, P. G. (**1987**). “Perceptual organization of complex-tone sequences : a tradeoff between pitch and timbre ?”, J. Acoust. Soc. Am. **82**, 886–899.

Singh, P. G. et Bregman, A. S. (**1997**). “The influence of different timbre attributes on the perceptual segregation of complex-tone sequences”, J. Acoust. Soc. Am. **102**, 1943–1952.

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

Stainsby, T. H., Moore, B. C. J., et Glasberg, B. R. (**2004**a). “Auditory streaming based on temporal structure in hearing-impaired listeners”, Hear. Res. **192**, 119–130.

Stainsby, T. H., Moore, B. C. J., Medland, P. J., et Glasberg, B. R. (**2004**b). “Sequential streaming and effective level differences due to phase-spectrum manipulations”, J. Acoust. Soc. Am. **115**, 1665–1673.

Stickney, G. S., Assmann, P. F., Chang, J., et Zeng, F.-G. (**2007**). “Effects of cochlear implant processing and fundamental frequency on the intelligibility of competing sentences”, J. Acoust. Soc. Am. **122**, 1069–1078.

Stickney, G. S., Zeng, F.-G., Litovsky, R., et Assmann, P. (**2004**). “Cochlear implant speech recognition with speech maskers”, J. Acoust. Soc. Am. **116**, 1081–1091.

Studebaker, G. A. (**1985**). “A “rationalized” arcsine transform”, J. Speech Hear. Res. **28**, 455–462.

Summers, V. et Leek, M. (**1998**). “F0 processing and the separation of competing speech signals by listeners with normal hearing and with hearing loss”, J. Speech Lang. Hear. Res. **41**, 1294–1306.

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

Tessier, E. (**2001**). “Étude de la variabilité de l’indice de localisation pour la caractérisation de sources de parole interférentes”, Thèse de doctorat, Institut National Polytechnique de Grenoble, France.

Thomas, I. B., Hill, P. B., Carroll, F. S., et Garcia, B. (**1970**). “Temporal order in the perception of vowels”, J. Acoust. Soc. Am. **48**, 1010–1013.

Tong, Y. C. et Clark, G. M. (**1985**). “Absolute identification of electric pulse rates and electrode positions by cochlear implant patients”, J. Acoust. Soc. Am. **77**, 1881–1888.

Townshend, B., Cotter, N., Compernolle, D. V., et White, R. L. (**1987**). “Pitch perception by cochlear implant subjects”, J. Acoust. Soc. Am. **82**, 106–115.

Trainor, L. et Trehub, S. (**1989**). "Aging and auditory temporal sequencing : ordering the elements of repeating tone patterns", Percept. Psychophys. **45**, 417–26.

Tsuzaki, M., Takeshima, C., Irino, T., et Patterson, R. D. (**2007**). "Auditory stream segregation based on speaker size, and identification of size-modulated vowel sequences", *in Hearing - From Sensory Processing to Perception*, edité par B. Kollmeier, G. Klump, V. Hohmann, U. Langemann, M. Mauermann, S. Uppenkamp, et J. Verhey, chapitre 31, 285 (Springer).

van Noorden, L. P. A. S. (**1975**). "Temporal coherence in the perception of tones sequences", Thèse de doctorat, Eindhoven University of Technology, The Netherlands.

Vliegen, J., Moore, B. C. J., et Oxenham, A. J. (**1999**). "The role of spectral and periodicity cues in auditory stream segregation, measured using a temporal discrimination task", J. Acoust. Soc. Am. **106**, 938–945.

Vliegen, J. et Oxenham, A. J. (**1999**). "Sequential stream segregation in the absence of spectral cues", J. Acoust. Soc. Am. **105**, 339–346.

Warren, R. M., Obusek, C. J., Farmer, R. M., et Warren, R. P. (**1969**). "Auditory sequence : Confusion of patterns other than speech or music", Science **164**, 586–587.

Zeng, F. G. (**2002**). "Temporal pitch in electric hearing", Hear. Res. **174**, 101–106.

Zeng, F. G. et Shannon, R. V. (**1994**). "Loudness-coding mechanisms inferred from electric stimulation of the human auditory system", Science **264**, 564–566.

Rôle de la ségrégation séquentielle pour la séparation de voix concurrentes

La ségrégation séquentielle n'a été que très peu étudiée avec des signaux de parole et il est donc difficile d'évaluer l'implication de ce mécanisme dans la séparation de voix concurrentes. Il est aussi difficile d'identifier les indices qui sont pertinents pour la séparation de signaux de parole, ces indices ayant surtout été étudiés isolément à l'aide de signaux très simples. L'objet de cette thèse est d'étudier les indices perceptifs impliqués dans la ségrégation séquentielle de voyelles différant par leur hauteur fondamentale. Les investigations menées montrent que les indices spectraux de hauteur jouent un rôle dans ce phénomène. En revanche, la périodicité de l'enveloppe temporelle ne semble pas constituer un indice de hauteur exploitable dans le cas de la parole. Enfin, il semble que la capacité des sujets à percevoir ces deux types d'indices ne permette pas d'expliquer l'ensemble des variations de performances des sujets dans une tâche de ségrégation séquentielle de voyelles.

Psychoacoustique — Mots clés : perception sonore, analyse des scènes auditives, ségrégation séquentielle, voyelles, malentendants.

Role of auditory streaming in the separation of concurrent voices

Only very few studies about auditory stream segregation have used speech, and it is then difficult to evaluate how this mechanism is involved in concurrent speech perception. It is also difficult to identify the cues that are relevant for the separation of speech because these cues have been mainly studied isolated in some very simple signal. The matter of this thesis is to study the perceptual cues that are involved in the sequential segregation of vowels with alternating pitch. The performed investigations show that the spectral cues do play a role in this phenomenon. Constrastively, the periodicity of the temporal envelope does not seem to be a consistent pitch cue for speech. Finally, it appears that the ability of the subjects to perceive these two kind of cues does not explain all the observed variations in the performances of the subjects in the vowel sequential segregation task.

Psychoacoustics — Keywords : auditory perception, auditory scene analysis, streaming, vowels, hearing-impairment.