Fabien Perrin & Nicolas Grimault, Laboratoire Cognition Auditive et Psychoacoustique, Centre de Recherche en Neurosciences de Lyon, Université Lyon 1 - CNRS

L'audition : Acoustique et Cerveau

Les fonctions sensorielles et perceptives sont à l'origine des représentations de l'environnement et de notre corps. Les organes sensoriels sont constitués de cellules spécialisées (récepteurs sensoriels) sensibles aux variations physiques et chimiques de l'environnement qu'elles «traduisent» (phénomène appelé transduction sensorielle) en activités électriques neuronales. Le message est alors transféré (on parle de conduction et de transmission neuronales) au système nerveux central. Schématiquement, le message «remonte» par la moelle épinière et/ou le tronc cérébral, puis traverse le thalamus et atteint le cortex. Les premières étapes du traitement de l'information (transduction et transmission au cortex) correspondent globalement au phénomène de sensation. La perception, quant à elle, peut être vue comme étant l'interprétation du message sensoriel, c'est-à-dire la comparaison du message sensoriel aux représentations acquises de notre environnement (les traces mnésiques). La perception serait plutôt le reflet de l'activité des cortex sensoriels et associatifs (ou cognitifs) qui fait suite au traitement sensoriel d'un stimulus.

L'audition

Parmi nos cinq sens, l'audition permet d'analyser et d'interpréter les sons. Son étude par la physiologie et la psychologie est rendue difficile par le fait qu'il n'y a pas totale analogie entre les attributs psychologiques et les attributs purement physiques. Si les aspects physiques du son peuvent être déterminés avec une grande précision, il est par contre beaucoup plus difficile de cerner précisément les impressions psychologiques qui en découlent. Dire par exemple que l'amplitude (mesurée en dB) correspond qualitativement à la sonie, ou que la fréquence tonale (mesurée en Hz) correspond à la hauteur tonale (grave, medium, aigu), n'est en fait pas exact. En effet, la sonie varie également selon les caractéristiques fréquentielles des sons (la différence de sonie entre deux sons de composition fréquentielle différente présentés à deux amplitudes différentes ne sera pas identique) et la composition fréquentielle d'un son détermine aussi le timbre (liée à la proportion d'harmoniques et aux relations d'amplitude de celles-ci).

Les voies nerveuses de l'audition sont assez bien connues, même s'il reste encore quelques zones d'ombre. La déformation de la membrane tympanique par les sons entraîne, via le mouvement des osselets et le pivotement des stéréocils des cellules ciliées de la cochlée, une modification du potentiel de membrane des cellules afférentes, c'est-à-dire des neurones qui constituent le nerf auditif. Celui-ci se projette sur le noyau cochléaire, qui est à l'origine de trois voies de transmission de l'information par le tronc cérébral. Ces voies, monaurales ou binaurales, passent notamment par le noyau olivaire et le noyau du lemnisque latéral, et convergent sur le colliculus inférieur, qui projette à son tour sur le corps genouillé médian du thalamus. Le thalamus projette quant à lui sur le cortex auditif primaire, situé anatomiquement au niveau du gyrus de Heschl, c'est-à-dire dans la partie supérieure du lobe temporal et en bordure de la scissure de Sylvius.

Les premières étapes du traitement de l'information auditive par les structures sous-corticales précédentes participeraient à la sélectivité fréquentielle, au codage de la sonie, à la localisation spatiale des sons, voire même aux premières étapes de l'analyse des scènes auditives.

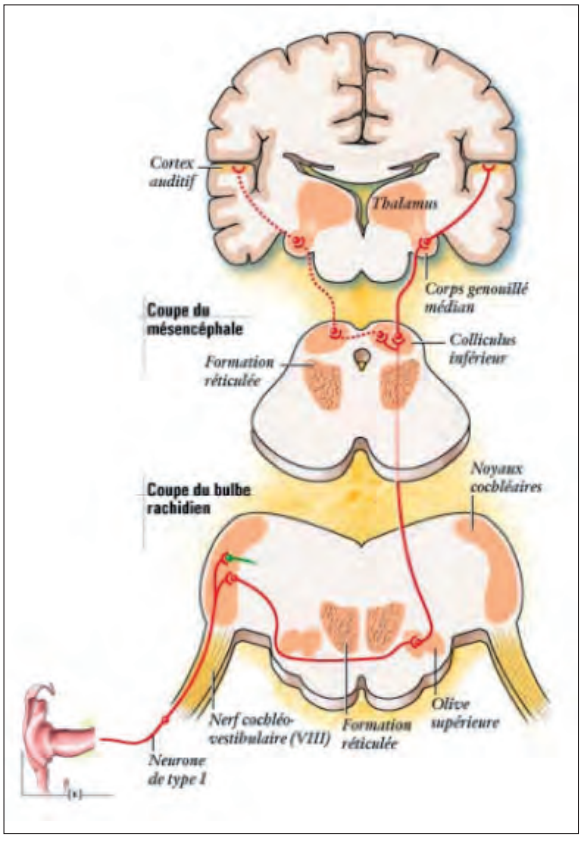

Fig 1 : Schéma simplifié des voies ascendantes du trajet de l'information sensorielle auditive. Source : Auteur : Gina Devau ; Conception technique et graphique : Alexis Gilliot et Olivier Ducos http://www.restice.univ-montp2.fr/anlo/co/ANLO-P4-CH1.html

Les neurones de l'olive supérieure médiane seraient de véritables détecteurs de coïncidence calculant les écarts temporels interauraux. La partie latérale de cette même structure, quant à elle, calculerait les différences interaurales d'intensité. Ces deux régions du tronc cérébral participeraient donc à notre capacité de localisation spatiale des sons. Dans le colliculus inférieur, on retrouverait plutôt des neurones à sélectivité tonale, c'est-à-dire des neurones qui déchargent spécifiquement pour certaines modulations d'amplitude et de fréquence (par exemple pour des sons de vocalisation). Ces neurones permettraient un codage périodotopique à l'origine de la perception de la hauteur. Enfin, le cortex auditif primaire présente une organisation tonotopique, comme la cochlée et les voies auditives (due à une projection point par point du thalamus), et peut-être même une organisation périodotopique. Certains auteurs évoquent aussi une spécialisation hémisphérique, le cortex auditif primaire gauche étant plus sensible aux variations de temps et le droit aux variations de fréquence.

La cognition auditive

L'interprétation du message sensoriel, c'est-à-dire la comparaison du signal acoustique entendu et transduit aux connaissances (traces mnésiques élaborées par l'exposition préalable à l'environnement), serait effectuée en grande partie par des cortex associatifs.

Le cortex auditif secondaire (ou ceinture du cortex auditif) par exemple reçoit des projections du thalamus et du cortex auditif primaire. Il est constitué notamment du planum temporale et, à gauche, de l'aire de Wernicke (jonction du lobe temporal et du lobe pariétal). L'organisation tonotopique du cortex auditif secondaire est moins stricte que celle du cortex auditif primaire, au profit d'une analyse plus cognitive du signal. Par exemple, l'aire de Wernicke serait impliquée dans l'analyse des sons du langage (pour leurs aspects phonologiques, sémantiques, syntaxiques, etc.). Plus généralement, il a été évoqué une véritable spécialisation hémisphérique des structures cognitives de l'audition. L'hémisphère gauche (structures des lobes temporal, frontal et pariétal) serait ainsi davantage impliqué dans l'analyse des sons du langage et l'hémisphère droit serait plutôt impliqué dans le traitement de la musique et de la prosodie du langage.

La description des fonctions cognitives de l'audition ne peut se limiter à cette rapide synthèse. En effet, celles-ci font intervenir de nombreuses autres structures corticales et sous-corticales, en fonction de la nature de la stimulation, du contexte environnemental, des motivations de la personne qui perçoit et de divers paramètres physiologiques internes. Par exemple, les sons sont très souvent associés à une dimension visuelle (la parole aux lèvres) ce qui implique des synthèses multisensorielles (implication des cortex visuels associatifs). Les ressources attentionnelles disponibles vont également conditionner la compréhension d'un message sonore et la capacité d'extraction d'un son du bruit de fond (rôle des cortex frontaux et temporaux). Enfin, l'apprentissage d'un son avec ou sans mouvement (via un instrument de musique) lie plus ou moins automatiquement et durablement les fonctions auditives et motrices (cortex moteur et pré-moteur, ganglion de la base, cervelet).

Fig 2 : « Dessine moi l'intérieur de ta tête »: Evolution de la représentation du cerveau au cours du développement chez l'enfant [A].

Diverses interactions avec le système auditif peuvent être décrites et chacune d'entre elles va s'exprimer différemment chez chacun(e) de nous. Car la dimension plastique du système nerveux est une donnée importante pour comprendre le fonctionnement cérébral. Nous ne pouvons plus nous limiter à une description purement localisationniste du système nerveux et nous devons maintenant le considérer comme un ensemble de réseaux neuronaux, dont la connectivité effective et fonctionnelle varie au cours du temps, c'est-à-dire à l'épreuve des expériences sensori-motrices.

Les méthodes exploratoires

Outre les techniques d'exploration comportementales usuelles de la psychoacoustique, les techniques expérimentales permettant l'étude de l'audition ont continûment progressé. Aujourd'hui, quatre méthodologies principales sont utilisées :

- la tomographie par émission de positons (PETscan) consiste à injecter un produit radioactif dans le sang pour suivre sa diffusion dans les différentes zones cérébrales ;
- l'imagerie par résonance magnétique fonctionnelle (IRMf) consiste à mesurer les variations locales de débit sanguin cérébral, liées aux variations magnétiques induites par l'oxygène du sang ;
- la magnétoencéphalographie (MEG) consiste à enregistrer l'activité magnétique des neurones induite par son activité électrique ;
- enfin, l'électroencéphalographie (EEG) consiste à enregistrer directement l'activité électrique des neurones, parfois avec des électrodes intracérébrales (on parle alors de stéréo-EEG).

Chacune de ces méthodologies présentent des avantages et des inconvénients en privilégiant soit la localisation spatiale de l'activité cérébrale (PET, IRMf) soit *a contrario* les fluctuations temporelles rapides de cette activité cérébrale (MEG, EEG). Toutefois ces deux dernières techniques peuvent aussi bénéficier de reconstructions mathématiques qui permettent de localiser les sources génératrices de l'activité électrique/magnétique. Par ailleurs, certaines méthodologies comme l'IRMf sont extrêmement bruyantes et nécessitent, lorsqu'on étudie l'audition, de mettre en place des protocoles expérimentaux alternant périodes de stimulation et périodes d'enregistrement.

Enfin, les dernières techniques d'analyse permettent maintenant d'explorer la connectivité effective et fonctionnelle, en recherchant des corrélations temporelles (par exemple des synchronisations oscillatoires) entre des régions cérébrales différentes.

Dans ce numéro spécial, il sera abordé les principales thématiques actuelles sur la perception auditive. Ainsi les articles [1,3,4,5,6] concerneront l'étude des différentes catégories de sons : langage, musique, et timbre. Les articles [1,2,4,5,7] aborderont la perception en contexte : concurrence, séparation de sources, sons multiples et interactions audiovisuelles. Enfin, les articles [1, 2, 3, 4] discuteront les conséquences fonctionnelles ainsi que les perspectives de réhabilitation de différentes pathologies auditives.

Liste des articles

[1]- Etienne Gaudrain (pp. 7 à 12) : Percevoir la parole ? Dans le bruit ? En étant malentendant ?
[2]- Isabelle Viaud-Delmon (pp. 13 à 17) : Les environnements sonores à le rescousse de la cognition.
[3]- Daniele Schön & Barbara Tillmann (pp. 18 à 22) : Cognition musicale.
[4]- Christian Füllgrabe (pp. 23 à 26) : Le vieillissement de la perception de la parole - Le rôle de l'audition périphérique, de l'audition centrale et de la cognition.
[5]- Laurent Demany & Samuele Carcagno (pp. 27 à 30) : Le «rehaussement» auditif.
[6]- Daniel Pressnitzer, Trévor Agus, & Clara Suied (pp. 31 à 34) : La reconnaissance du timbre des sons.
[7]- Nicolas Grimault & Aymeric Devergie (pp. 35 à 37): Les mécanismes cognitif de l'analyse séquentielle des scènes auditives.

Références bibliographiques

[A] Savy (2005) « Comment des enfants de 5 à 11 ans dessinent ce qu'ils ont dans leur tête », Dissertation doctorale, Université Lyon 1.

[B] Purves D, Augustine GJ, Fitzpatrick D, Hall WC, LaMantia A-S, White LE (2005) Neurosciences. De Boeck Université, 4e édition, Bruxelles.

Glossaire

Coefficients cepstraux : Le cepstre d'un signal est une transformation de ce signal du domaine temporel vers un autre domaine analogue au domaine temporel. Pour rappeler le fait que l'on effectue une transformation inverse à partir du domaine fréquentiel, les dénominations des notions sont des anagrammes de celles utilisées en fréquentiel. Ainsi le spectre devient le cepstre, la fréquence une quéfrence, un filtrage un liftrage…

Corrélats neuronaux : Les chercheurs en neurosciences cognitives ont entrepris de dresser des ponts entre des états mentaux (perçus, ressentis, et donc subjectifs) et des états neuraux. Ces programmes de recherche tentent d'identifier les «corrélats neuronaux de la conscience», c'est-à-dire des processus qui surviennent dans les circuits du cerveau lors d'une expérience particulière.

Diotique : Se dit d'une stimulation strictement identique et simultanée dans chacune des deux oreilles.

Formant : Fréquence de résonance du conduit vocal dont la valeur dépend de la configuration des cavités buccale et pharyngale propre à chaque articulation. En première approximation, les voyelles et certaines consonnes se définissent acoustiquement par leurs formants.

Hédonicité : Qui se rapporte à l'émotion. L'hédonicité positive d'un signal se réfère ainsi au plaisir qu'il procure.

Modification tonotopique : Se réfère à une variation de la correspondance entre la fréquence du signal utilisée pour stimuler et les zones corticales ou sous corticale excitées.

Percept : Entité cognitive, constituée d'un ensemble d'informations sélectionnées et structurées en fonction de l'expérience antérieure, et qui sont mobilisées dans une perception particulière.

Proprioception : Perception de son propre corps et de son activité : kinesthésique (sens du mouvement) et posturale.

Prosodie de la parole : Les changements de hauteur de notre voix lorsque nous parlons. Ces changements portent des informations sur l'intonation ainsi que sur le locuteur (genre, émotion...).

Schéma de Bregman : Ce sont des traces sensorielles auditives acquises par l'expérience et utilisées pour analyser une scène auditive.

Percevoir la parole ? Dans le bruit ? En étant malentendant ?

Etienne Gaudrain
University Medical Center Groningen, ENT Department
University of Groningen
Research School of Behavioral Cognitive Neuroscience
Hanzeplein 1
9713 GZ Groningen
Pays-Bas
E-mail : e.p.c.gaudrain@umcg.nl

Résumé
La perception de la parole dans le bruit reste un challenge pour les personnes atteintes de pertes auditives, et ce, malgré le progrès des prothèses auditives. L'étude de ce phénomène passe d'une part par la compréhension de ce qu'est le signal de parole lui-même, et comment il est perçu, et d'autre part par la compréhension des mécanismes perceptifs et cognitifs qui permettent de séparer un locuteur cible du bruit de fond, et de comprendre ce qu'il dit. Ces deux aspects sont passés en revue, ainsi que l'effet d'une perte auditive sur chacun d'entre eux.

Abstract
Speech perception in noisy environments remains a challenge for hearing-impaired listeners despite the most recent improvements of auditory prostheses. The study of this phenomenon requires (i) to understand what the speech signal is itself, and how it is perceived ; and (ii) to understand the perceptual and cognitive mechanisms allowing the separation of a target speaker in a background noise, and to get what they say. These two aspects are reviewed here, along with the effect of hearing loss on each of them.

Comprendre un locuteur unique, dans un environnement calme, lorsqu'on est jeune et en bonne santé, ne pose généralement pas de problème. Notre audition n'est cependant pas toujours aussi parfaite que chez ces auditeurs de référence, par exemple à la suite d'un traumatisme acoustique, ou par conséquence naturelle du vieillissement (voir l'article de Füllgrabe page 23). Dans cette situation, les personnes souffrant de déficits auditifs, peuvent avoir besoin d'une aide auditive pour atteindre des niveaux de compréhension comparables à ceux des «normo-entendants». Dans les environnements bruyants, en revanche, alors que la situation devient plus ardue pour les normo-entendants, l'emploi d'une prothèse auditive ne permet généralement pas aux malentendants de retrouver une audition normale. Cette limite à l'efficacité des prothèses a non seulement un effet sur la vie sociale (qui a généralement lieu dans des lieux bruyants) des personnes appareillées, mais est aussi un facteur majeur de disqualification des prothèses auditives chez les personnes souffrant de troubles auditifs[1].

La perception de la parole est un phénomène complexe : le signal de parole n'est pas particulièrement simple et les mécanismes perceptifs et cognitifs qui conduisent à sa compréhension sont intriqués et peu connus. Pour étudier l'effet d'une perte auditive ou celui de la présence de bruit dans l'environnement de l'auditeur, il est prudent de décomposer ce problème en plusieurs problèmes plus simples. Dans un premier temps, on peut s'intéresser au signal de parole lui-même : quelles sont ses dimensions principales, et comment sont-elles perçues ? Dans un second temps, lorsque ce signal est mélangé à d'autres signaux concurrents, on peut s'intéresser aux mécanismes qui permettent d'extraire le signal cible du bruit ambiant, et de l'interpréter.

Attributs perceptifs du signal de parole

Le signal de parole trouve son origine dans la vibration des cordes vocales qui produisent des séries d'impulsions acoustiques, périodiques, qui se propagent dans la cavité buccale, avant de rayonner à l'ouverture de la bouche (voir Figure 1). La cavité buccale agit comme un résonateur, ou plus exactement comme un ensemble de résonateurs. Pour cette raison, le système vocal est classiquement décrit comme un système source-filtre [2], [3] dont les cordes vocales constituent la source, et la cavité buccale, le filtre. Les caractéristiques de ces résonateurs peuvent être manipulées par le locuteur en modifiant la configuration de sa langue, de sa mâchoire et de ses lèvres. Ceci a pour effet de créer des pics, appelés formants, sur l'enveloppe spectrale du signal de parole, dont les positions correspondent aux fréquences de résonance de ces cavités.

1-Seul un quart des personnes âgées de 65 à 84 ans rapportant une perte de l'audition utilisent une correction [1].

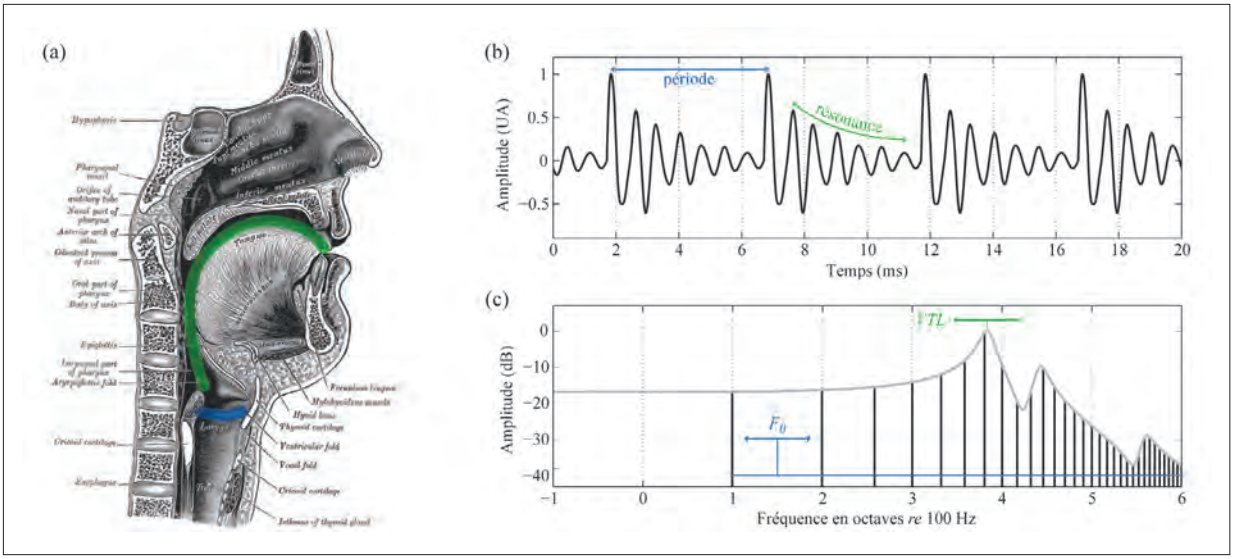

Fig. 1 : (a) Coupe sagittale de l'appareil vocal (d'après [36]). Les cordes vocales (la source) sont surlignées en bleu, et le tractus vocal (le filtre) est surligné en vert. (b) Forme d'onde de la voyelle /a/. La période des impulsions glottiques est mise en évidence par une flèche bleue. La durée de la résonance dans le tractus vocal est montrée par une flèche verte. (c) Spectre du même son. Les raies harmoniques sont montrées en noir, l'enveloppe spectrale est représentée en gris. La fréquence fondamentale (F_0) est indiquée en bleu, et l'effet de la longueur du tractus vocal (VTL) est illustré en vert. Adapté de [6].

Le système de production de la parole peut donc être caractérisé par deux dimensions principales : la fréquence de pulsation des cordes vocales (abrégé GPR pour *glottal-pulse rate*) et la longueur du tractus vocal (noté VTL pour *vocal-tract length*). Ces dimensions ont des conséquences directement visibles sur la forme d'onde et sur le spectre du signal acoustique produit (voir Figure 1). La GPR, responsable de la nature périodique du signal, se traduit directement par la fréquence fondamentale (F_0). La longueur du tractus vocal se traduit par la durée de la résonance de l'impulsion glottique, ou, dans le spectre, par la position de l'enveloppe spectrale sur un axe de fréquence logarithmique. Quand un locuteur grandit de l'enfance à l'âge adulte, l'enveloppe spectrale de ses productions vocales se translatent progressivement vers les basses fréquences, préservant les relations entre formants [4].

Ces deux dimensions, F_0 et VTL, produisent aussi des percepts très distincts. La fréquence fondamentale donne directement lieu au percept de hauteur fondamentale. La hauteur fondamentale est un percept extrêmement clair, auquel nous sommes très sensibles : le seuil de détection d'une différence de hauteur dans un son de parole est de l'ordre de 2% (ou un quart de ton) [5]. Le nom du percept associé à la VTL reste malheureusement encore à inventer mais celui-ci correspond essentiellement à la taille perçue, ou apparente, du locuteur. Le concept se généralise aux instruments de musique et distingue alors les instruments qui couvrent les différents registres d'une famille d'instruments, par exemple le violon et le violoncelle [6]. La VTL bénéficie, elle aussi, d'une excellente sensitivité, avec des seuils de détection variant de 5 à 7% (un peu plus d'un demi-ton) [7].

Plusieurs indices poussent à croire que F_0 et VTL sont bien des dimensions fondamentales des sons de parole au niveau perceptif.

D'abord, nous sommes capables de comprendre des sons de parole couvrant un des combinaisons de F_0 et VTL qui dépassent largement le domaine normalement couvert par les humains. Par exemple, la reconnaissance de voyelles n'est que faiblement affectée lorsque l'on simule un locuteur de plus de 3 m de hauteur, ou de moins de 50 cm [7]. Alors même que nous sommes très sensibles à ces deux dimensions, nous sommes aussi parfaitement capable de les évaluer séparément et de les ignorer le cas échéant. Confirmant cette idée, des études utilisant des techniques de neuro-imagerie ont montré que les nouveaux-nés étaient sensibles à ces dimensions [8], et ont mis en évidence l'existence d'aires cérébrales dédiées au traitement de chacune de celles-ci [9], [10], [11].

Donc, en tant qu'auditeurs, nous utilisons F_0 et VTL pour évaluer la taille d'une personne à partir de sa voix [5]. Ce qui permet ainsi d'estimer son âge, ou du moins de distinguer les enfants des adultes. En outre, lorsque nous avons affaire à des interlocuteurs adultes, F_0 et VTL nous permettent d'estimer s'il s'agit d'un homme ou d'une femme [12] : les hommes tendent à être plus grands que les femmes, et ont donc un tractus vocal plus long ; leurs cordes vocales sont plus lourdes (un résultat de la montée de testostérone qui survient à l'adolescence) et vibrent donc plus lentement. La voix plus «grave» des locuteurs masculins se caractérise donc non seulement par une hauteur fondamentale plus basse que celle des femmes, mais aussi par des formants positionnés à des fréquences sensiblement plus basses.

De manière générale, ces deux dimensions permettent de distinguer une voix d'une autre, et donc, un locuteur d'un autre. Dans ce contexte, nous utilisons les connaissances que nous avons accumulées sur ces deux dimensions (de façon implicite) pour interpréter correctement ce que nous percevons.

En particulier, nous avons montré que lorsqu'un auditeur doit décider si deux voix proviennent de la même personne, la F_0 ne devenait un indice fiable que lorsque la différence entre ces deux voix excédait 3,8 demi-tons (soit bien plus que le seuil de détection d'un quart de ton). En revanche, 2,2 demi-tons étaient suffisants suivant la dimension de VTL, qui a pourtant un seuil de détection d'un demi-ton plus grand que pour la F_0 [13]. L'étude des fluctuations naturelles de F_0 dans la voix révèle des déviations standards autour de la moyenne de +/- 3,6 demi-tons [14], directement comparable à la valeur de 3,8 demi-tons obtenue dans notre étude. Cette différence entre F_0 et VTL peut s'expliquer par le fait que les locuteurs sont capables de modifier la F_0 de leur voix sur une très grande plage de valeurs que les auditeurs connaissent, alors que modifier la longueur de leur tractus vocal est beaucoup plus ardu.

Puisque F_0 et VTL permettent de distinguer les locuteurs les uns des autres, il est tout naturel que ces dimensions jouent un rôle primordial dans la perception de la parole dans le bruit. Lorsque deux syllabes sont prononcées exactement au même instant, de petites différences de F_0 et/ou de VTL (2,0 et 3,2 demi-tons, respectivement) facilitent leur perception [15]. Cet effet ne se cantonne pas aux syllabes, mais se généralise à la perception de phrases simultanées [16].

La perception de la F_0 chez les malentendants a été largement décrite dans la littérature comme «dégradée» ou «faible» [17]. Cette faible représentation de la hauteur semble avoir des conséquences directes sur la capacité des malentendants à séparer perceptivement des voix qui diffèrent selon cette dimension [18], [19], [20]. La dégradation de la perception de la VTL est en revanche beaucoup moins documentée [21]. En particulier, chez les implantés cochléaires, il semble que la dimension soit totalement inaccessible [22].

Mécanismes de séparation de voix concurrentes

Les indices perceptifs présentés ci-dessus, s'ils sont accessibles à l'auditeur, sont exploités par des mécanismes perceptifs qui assurent la séparation de voix concurrentes. On peut catégoriser ces mécanismes en trois grandes classes [23] : les mécanismes de ségrégation simultanée, les mécanismes de ségrégation séquentielle, et les mécanismes de restauration.

Les mécanismes de ségrégation simultanée séparent les évènements sonores qui surviennent simultanément, par exemple deux voyelles prononcées au même moment par deux locuteurs différents. Les mécanismes de ségrégation séquentielle assemblent des évènements sonores successifs pour former un flux continu, par exemple les syllabes successives prononcées par un locuteur, excluant celles prononcées par d'autres locuteurs.

Enfin, les mécanismes de restauration sont, eux, responsables de l'interprétation des signaux acoustiques, c'est-à-dire d'inférer le contenu sémantique associé aux sons.

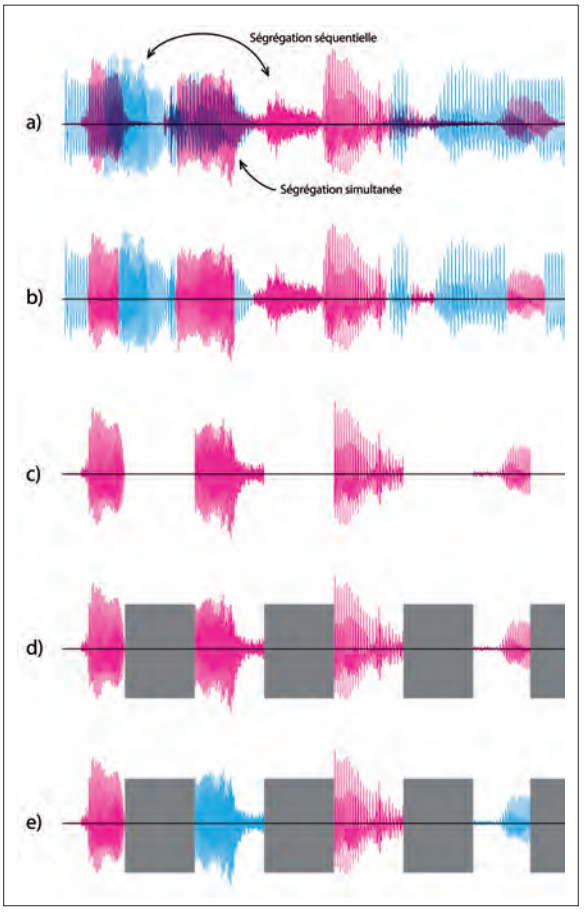

Fig. 2 : (a) Formes d'onde de deux portions de phrases concurrentes. L'une des phrases est représentée en magenta, tandis que l'autre est représentée en cyan. Les deux signaux sont représentés en transparence l'un au dessus de l'autre. Lorsqu'il y a recouvrement, le signal est représenté en bleu foncé. Les flèches pointent des segments où les deux mécanismes de ségrégation sont clairement présents. (b) Les deux mêmes phrases mélangées en utilisant la méthode de parole zébrée. (c) Une des deux phrases périodiquement interrompue par des silences. (d) La même phrase, interrompue par des bruits, représentés par des rectangles gris. (e) La même phrase, interrompue par des bruits, mais où les segments sont prononcés tour à tour par un homme et une femme.

Pour comprendre le rôle de chacune de ces classes de mécanismes, on peut décrire le scénario simpliste suivant : Considérons deux locuteurs prononçant deux phrases différentes au même moment, comme illustré dans la figure 2.a.

Lorsque les deux locuteurs prononcent deux syllabes au même moment, l'auditeur doit avoir recours à des mécanismes de ségrégation simultanée pour les séparer perceptivement, c'est-à-dire pour entendre l'un ou l'autre des locuteurs. Le concept ne se limite pas aux syllabes, mais s'applique à tout segment de signal où les deux locuteurs produisent de l'énergie. Les segments successifs provenant d'un locuteur donné doivent ensuite être assemblés pour former un flux continu et cohérent. Cette opération n'est triviale qu'en apparence. Le signal de parole présente, naturellement, des silences et des pauses (qui ont une valeur informationnelle), ce qui fait que les segments successifs de parole peuvent être disjoints dans le temps.

De plus, pendant ces pauses, l'autre locuteur peut lui-même produire des segments contenant de l'énergie et du contenu phonétique. Les mécanismes de ségrégation séquentielle doivent donc non seulement assembler les segments de parole qui proviennent d'un certain locuteur, mais aussi exclure ceux qui proviendraient de l'autre locuteur. Cet aiguillage est, en outre, compliqué par le fait qu'il s'opère non seulement sur des segments clairs de parole (c'est-à-dire prononcés par un locuteur pendant que l'autre est silencieux) mais aussi sur des segments issus de la ségrégation simultanée et qui peuvent donc être dégradés si cette dernière est imparfaite. Le flux obtenu est ainsi constitué de segments incomplets de parole, et en inférer le contenu sémantique est aussi peu trivial, d'où le terme «restauration» utilisé pour identifier ce processus d'inférence.

On verra plus loin que ces mécanismes ne s'enchaînent pas nécessairement dans cet ordre, et qu'ils entretiennent en réalité une relation plus intriquée. Cependant leur distinction reste utile comme outil d'étude, notamment car ces différentes familles de mécanismes exploitent différents aspects du signal de parole, et sont affectés différemment par les effets d'une perte auditive.

Ségrégation simultanée

La méthode de prédilection pour étudier la ségrégation simultanée est de présenter deux voyelles, ou deux syllabes, commençant au même instant, et terminant en même temps. Les deux voyelles diffèrent non seulement par leur nature phonologique, mais aussi suivant une autre dimension perceptive, comme la hauteur fondamentale [24] ou la VTL [15]. Lorsque les spectres des deux sons utilisés présentent une zone substantielle de recouvrement, il est possible qu'une partie du spectre d'un des sons «masque» une partie de l'autre son. C'est-à-dire que les fibres nerveuses codant cette zone fréquentielle ne répondent qu'au son masquant, le plus fort dans cette plage de fréquence, et non au son masqué, plus faible. Comme la réponse en fréquence du système auditif a une résolution limitée, ce phénomène de masquage s'étend autour du masque, rendant l'identification du son masqué d'autant plus difficile.

Une des conséquences de la perte auditive concerne justement la résolution spectrale. Les pertes neurosensorielles, comme la presbyacousie, se caractérisent généralement par une diminution de la résolution spectrale, un effet qui ne peut être compensé par les prothèses auditives actuelles. La réduction de résolution spectrale est encore plus spectaculaire chez les personnes munies d'un implant cochléaire. Chez ces sujets, deux effets se combinent : d'une part, le renforcement des indices perceptifs, comme la F_0, utilisés pour séparer les deux sons, et d'autre part, le renforcement de l'effet masquant d'un son sur l'autre.
On peut donc dire sans surprise, que les pertes auditives ont un effet direct sur l'efficacité de ce mécanisme [20], [25].

Ségrégation séquentielle

La ségrégation séquentielle surtout a été étudiée par des méthodes reposant sur les signaux artificiels, très différents de la parole.

Initialement, une séquence de sons purs, dont la fréquence alternait entre deux valeurs, était présentée à des sujets qui devaient indiquer s'ils percevaient un ou deux «flux» [26]. Lorsque les sons purs ont des fréquences proches, les sujets tendent à entendre la séquence comme provenant d'une seule source sonore. Au contraire, lorsqu'une grande différence de fréquence sépare les différents sons purs, la séquence est entendue comme provenant de deux sources sonores distinctes.

Alors que certains théoriciens ont émis des doutes quant à l'utilité de ce type de mécanisme pour les signaux de parole [23], [27], des études récentes ont montré que les séquences de voyelles étaient soumises à la même organisation en flux auditifs, par exemple sur la base de la F_0 [19] ou de la VTL [28].

De manière générale, une revue de la littérature sur la ségrégation séquentielle met en évidence que tout indice suffisamment saillant peut être responsable d'une création de flux auditifs [29]. Lorsque l'indice responsable de l'organisation perceptive de la séquence devient moins saillant, comme à la suite d'une perte auditive, la ségrégation séquentielle devient moins automatique, moins systématique. C'est le cas, par exemple, pour la hauteur fondamentale qui est moins saillante chez les malendendants que chez les normo-entendants, résultant en un déficit de ségrégation séquentielle [18], [19].

En revanche, contrairement à la ségrégation simultanée, la perte auditive (et la perte de résolution fréquentielle qui en résulte) n'affecte pas directement l'intelligibilité des éléments de la séquence. En effet, puisque les voix concurrentes ne se chevauchent pas dans le temps, une perte de résolution fréquentielle n'engendre pas d'accroissement du phénomène de masquage dans la ségrégation séquentielle. Autrement dit, la perte auditive altère surtout l'organisation perceptive séquentielle, mais par les éléments de parole constituant la séquence, alors que la ségrégation simultanée souffre à la fois de la perte de saillance des indices perceptifs et du masquage accru dû à la perte de résolution spectrale.

Récemment, nous avons pu vérifier que ce principe d'inégalité entre ségrégation séquentielle et simultanée face à la perte de résolution spectrale s'appliquait effectivement au cas de la parole. Pour cela nous avons développé une méthode appelée «parole zébrée» (*zebra-speech*) qui consiste à obtenir le mélange purement séquentiel de deux phrases concurrentes en ne gardant que le signal le plus intense à chaque «instant» [30]. Un exemple illustrant la parole zébrée est représenté Figure 2.b. En comparant l'intelligibilité de la parole zébrée à celle du mélange classique où les deux signaux sont simplement sommés (comme dans la Figure 2.a), nous avons observé que lorsque la résolution spectrale diminuait, l'intelligibilité de la parole zébrée décroissait moins vite que celle du mélange classique.

Inférence et restauration

La troisième famille de mécanismes concerne la capacité d'un locuteur à extraire le contenu sémantique d'un flux de parole, dégradé ou non. La difficulté de cette étape dépend du succès des deux autres types de mécanismes.

Si les mécanismes de ségrégation parviennent à extraire le signal de parole cible de ce que l'auditeur entend, l'inférence sera triviale, et ne mérite pas d'être appelée «restauration». En revanche, si, comme c'est le cas dès qu'une autre source interfère avec le locuteur ciblé, les mécanismes de ségrégation ne peuvent fournir qu'une représentation partielle du signal de parole original, l'auditeur doit utiliser toutes les ressources à sa disposition pour tenter d'interpréter le message, et éventuellement restaurer les parties manquantes.

Deux méthodes sont classiquement utilisées pour étudier ce phénomène : la parole interrompue et le paradigme de restauration phonémique. La parole interrompue consiste en une phrase modulée par un signal périodique carré, c'est-à-dire qu'elle est périodiquement interrompue par un silence (voir Figure 2.c). Comprendre la phrase dans ces circonstances nécessite de deviner les éléments manquants du signal. Lorsque le taux d'interruption est rapide (plusieurs dizaines de Hertz), l'effet sur l'intelligibilité est minime [31]. En revanche, lorsque le taux d'interruption se rapproche du nombre moyen de syllabes par seconde (autour de 5 ou 6 Hz), l'intelligibilité décroît rapidement (restant néanmoins entre 30 et 50%, ou même 60% suivant les phrases utilisées). Les pertes auditives modérées ne semblent pas avoir un effet substantiel sur l'intelligibilité de la parole interrompue. Cependant la perception de la parole interrompue est soumise à un fort effet d'âge [32]. La plupart des sujets malentendants étant aussi âgés, il est difficile de séparer ces deux variables. En revanche, une forte détérioration de la résolution spectrale comme celle qui survient dans les implants cochléaires, affecte aussi la capacité des sujets à s'accommoder des interruptions [33].
Le paradigme de restauration phonémique consiste à mesurer l'intelligibilité de phrases interrompues, comme ci-dessus, puis de répéter l'opération avec des phrases dans lesquelles les interruptions ont été remplies par du bruit (voir Figure 2.d). Pour certains taux d'interruption, l'ajout de bruit, au lieu de réduire l'intelligibilité, résulte en un signal mieux compréhensible [34]. Pour ce phénomène de restauration phonémique, l'effet délétère d'une perte auditive semble relativement clair [35] tandis que l'effet de l'âge est moins clair. Bien que les mécanismes sous-jacents soient encore mal compris, la plupart des études mentionnent qu'en présence du bruit, les sujets perçoivent un son plus continu qu'en son absence. Ces observations suggèrent, parfois implicitement, que la présence de bruit facilite le groupement perceptif des deux segments de parole se trouvant de chaque côté de l'interruption, un phénomène qui n'est pas sans rappeler le mécanisme de ségrégation séquentielle, sans que le lien ait toutefois été formellement établi.

Conclusion

Comme le suggère cette dernière observation, il est probable que ces trois classes de mécanismes interagissent. Dans une étude récemment menée dans notre laboratoire, nous avons mis au point une situation de restauration phonémique où les segments de parole étaient prononcés successivement par deux voix différant par leur F_0 et leur VTL de façon à simuler un homme et une femme (Figure 2.e). Alors même que le signal était perçu comme provenant de deux locuteurs différents, les phrases étaient tout aussi intelligibles et l'ajout de bruit résultait bien en une augmentation des scores de compréhension. Autrement dit, malgré la présence d'indices perceptifs clairs indiquant à l'auditeur que deux locuteurs différents étaient présents, les sujets groupaient les segments successifs de la phrase afin d'inférer le sens de cette dernière.

Pour aller plus loin, on peut même supposer que le fait que les segments successifs appartiennent à la même phrase et forment donc un sens cohérent ait été directement utilisé comme indice de groupement, poussant le sujet à ignorer la différence de voix. Il s'agirait alors d'une interaction directe entre le mécanisme de restauration et celui de ségrégation séquentielle. Cette hypothèse fait écho aux travaux de Grimault et Devergie présentés dans l'article page 35 sur le rôle des connaissances dans la séparation de séquences musicales. En généralisant cette hypothèse, on peut donc se demander si les grandeurs perceptives de F_0 et de VTL se contentent de piloter les mécanismes de ségrégation, ou bien si elles sont elles-mêmes issues d'un processus d'analyse des scènes auditives.

Références bibliographiques

[1] « L'état de santé de la population en France - Suivi des objectifs annexés à la loi santé publique - Rapport 2009-2010 », Direction de la Recherche, des Études, de l'Évaluation et des Statistiques, http://www.drees.sante.gouv.fr/IMG/pdf/etat_sante_2009-2010.pdf, 2010.

[2] T. Chiba et M. Kajiyama, The vowel, its nature and structure. Tokyo-Kaiseikan Pub Co., Tokyo, 1941.

[3] G. Fant, Acoustic Theory of Speech Production. The Hague: Mouton De Gruyter, 1960.

[4] G. E. Peterson et H. L. Barney, « Control Methods Used in a Study of the Vowels », J. Acoust. Soc. Am., vol. 24, no 2, pp. 175-184, mars 1952.

[5] D. T. Ives, D. R. R. Smith, et R. D. Patterson, « Discrimination of speaker size from syllable phrases », J. Acoust. Soc. Am., vol. 118, no 6, pp. 3816-3822, déc. 2005.

[6] R. D. Patterson, E. Gaudrain, et T. C. Walters, « The Perception of Family and Register in Musical Notes », in Music Perception, M. R. Jones, R. R. Fay, et A. N. Popper, Éd. Springer, 1st Edition., vol. 36, pp. 13-50, 2010,.

[7] D. R. R. Smith, R. D. Patterson, R. Turner, H. Kawahara, et T. Irino, « The processing and perception of size information in speech sounds », J. Acoust. Soc. Am., vol. 117, no 1, pp. 305-318, janv. 2005.

[8] M. D. Vestergaard, G. P. Háden, Y. Shtyrov, R. D. Patterson, F. Pulvermüller, S. L. Denham, I. Sziller, et I. Winkler, « Auditory size-deviant detection in adults and newborn infants », Biol. Psychol., vol. 82, no 2, pp. 169-175, oct. 2009.

[9] K. von Kriegstein, D. R. R. Smith, R. D. Patterson, D. T. Ives, et T. D. Griffiths, « Neural Representation of Auditory Size in the Human Voice and in Sounds from Other Resonant Sources », Curr. Biol., vol. 17, no 13, pp. 1123-1128, juill. 2007.

[10] K. von Kriegstein, D. R. R. Smith, R. D. Patterson, S. J. Kiebel, et T. D. Griffiths, « How the human brain recognizes speech in the context of changing speakers », J. Neurosci., vol. 30, no 2, pp. 629-638, janv. 2010.

[11] P. Belin, P. E. G. Bestelmeyer, M. Latinus, et R. Watson, « Understanding voice perception », Br. J. Psychol., vol. 102, no 4, pp. 711-725, nov. 2011.

[12] D. R. R. Smith et R. D. Patterson, « The interaction of glottal-pulse rate and vocaltract length in judgements of speaker size, sex, and age », J. Acoust. Soc. Am., vol. 118, no 5, pp. 3177-3186, nov. 2005.

[13] E. Gaudrain, S. Li, V. Ban, et R. Patterson, « The role of glottal pulse rate and vocal tract length in the perception of speaker identity », Interspeech 2009, vol. 1-5, pp. 152-155, 2009.

[14] R. E. Kania, D. M. Hartl, S. Hans, S. Maeda, J. Vaissiere, et D. F. Brasnu, « Fundamental frequency histograms measured by electroglottography during speech: a pilot study for standardization », J. Voice, vol. 20, no 1, pp. 18-24, mars 2006.

[15] M. D. Vestergaard, N. R. C. Fyson, et R. D. Patterson, « The interaction of vocal characteristics and audibility in the recognition of concurrent syllables », J. Acoust. Soc. Am., vol. 125, no 2,p p. 1114-1124, févr. 2009.

[16] C. J. Darwin, D. S. Brungart, et B. D. Simpson, « Effects of fundamental frequency and vocal-tract length changes on attention to one of two simultaneous talkers », J. Acoust. Soc. Am., vol. 114, no 5, pp. 2913-2922, nov. 2003.

[17] B. C. J. Moore et R. P. Carlyon, « Perception of pitch by people with cochlear hearing loss and by cochlear implant users », in Pitch: neural coding and perception, C. J. Plack, A. J. Oxenham, R. R. Fay, et A. N. Popper, Éd. New-York, NY: Springer/Birkhäuser, pp. 234-277, 2005.

[18] E. Gaudrain, N. Grimault, E. W. Healy, et J.-C. Béra, « Streaming of vowel sequences based on fundamental frequency in a cochlear-implant simulation », J. Acoust. Soc. Am., vol. 124, no 5, pp. 3076-3087, nov. 2008.

[19] E. Gaudrain, N. Grimault, E. W. Healy, et J.-C. Béra, « Effect of spectral smearing on the perceptual segregation of vowel sequences », Hear. Res., vol. 231, no 1-2, pp. 32-41, sept. 2007.

[20] V. Summers et M. R. Leek, « FO processing and the separation of competing speech signals by listeners with normal hearing and with hearing loss », J. Speech Lang. Hear. Res., vol. 41, no 6, pp. 1294-1306, déc. 1998.

[21] C. L. Mackersie, J. Dewey, et L. A. Guthrie, « Effects of fundamental frequency and vocal-tract length cues on sentence segregation by listeners with hearing loss », J. Acoust. Soc. Am., vol. 130, no 2, pp. 1006-1019, 2011.

[22] Q.-J. Fu, S. Chinchilla, G. Nogaki, et J. J. Galvin 3rd, « Voice gender identification by cochlear implant users: the role of spectral and temporal resolution », J. Acoust. Soc. Am., vol. 118, no 3 Pt 1, pp. 1711-1718, sept. 2005.

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

[24] A. de Cheveigné, « Waveform interactions and the segregation of concurrent vowels », J. Acoust. Soc. Am., vol. 106, no 5, pp. 2959-2972, nov. 1999.

[25] X. Luo, Q.-J. Fu, H.-P. Wu, et C.-J. Hsu, « Concurrent-vowel and tone recognition by Mandarin-speaking cochlear implant users », Hear. Res., vol. 256, no 1-2, pp. 75-84, oct. 2009.

[26] L. P. A. S. van Noorden, « Temporal coherence in the perception of tones sequences », Ph.D. thesis, Eindhoven University of Technology, The Netherlands, 1975.

[27] R. E. Remez, P. E. Rubin, S. M. Berns, J. S. Pardo, et J. M. Lang, « On the perceptual organization of speech », Psychol. Rev., vol. 101, no 1, pp. 129-156, janv. 1994.

[28] C. Takeshima, M. Tsuzaki, et T. Irino, « Perception of vowel sequence with varying speaker size », Acoust. Sci. & Tech., vol. 31, no 2, pp. 156-164, 2010.

[29] B. C. J. Moore et H. Gockel, « Factors Influencing Sequential Stream Segregation », Acta Acust. united Ac., vol. 88, no 3, pp. 320-333, 2002.

[30] E. Gaudrain et R. P. Carlyon, « Using Zebra-speech to study sequential and simultaneous speech segregation in a cochlear-implant simulation », J. Acoust. Soc. Am., vol. 133, no 1, pp. 502-518, janv. 2013.

[31] P. B. Nelson et S.-H. Jin, « Factors affecting speech understanding in gated interference: cochlear implant users and normal-hearing listeners », J. Acoust. Soc. Am., vol. 115, no 5 Pt 1, pp. 2286-2294, mai 2004.

[32] M. Bergman, V. G. Blumenfeld, D. Cascardo, B. Dash, H. Levitt, et M. K. Margulies, «Age-related decrement in hearing for speech. Sampling and longitudinal studies », J. Gerontol., vol. 31, no 5, pp. 533-538, sept. 1976.

[33] M. Chatterjee, F. Peredo, D. Nelson, et D. Baskent, « Recognition of interrupted sentences under conditions of spectral degradation », J. Acoust. Soc. Am., vol. 127, no 2, pp. EL37-41, févr. 2010.

[34] G. L. Powers et J. C. Wilcox, « Intelligibility of temporally interrupted speech with and without intervening noise », J. Acoust. Soc. Am., vol. 61, no 1, pp. 195, 1977.

[35] D. Baskent, « Phonemic restoration in sensorineural hearing loss does not depend on baseline speech perception scores », J. Acoust. Soc. Am., vol. 128, no 4, pp. EL169-174, oct. 2010.

[36] H. Gray, Anatomy of the human body. Philadelphia: Lea & Febiger, 1908.