Percevoir la parole quand elle est produite différemment : étude des mécanismes de familiarisation multimodale/multisensorielle entre locuteurs tout-venants et locuteurs présentant un trouble de l'articulation

Alexandre Hennequin

Communauté
UNIVERSITÉ Grenoble Alpes

THÈSE

Pour obtenir le grade de

DOCTEUR DE LA COMMUNAUTE UNIVERSITE GRENOBLE ALPES

Spécialité : **CIA « Ingénierie de la Cognition, de l'interaction, de l'Apprentissage et de la création »**

Arrêté ministériel : 25 mai 2016

Présentée par

Alexandre HENNEQUIN

Thèse dirigée par **Marion DOHEN**
et codirigée par **Amélie ROCHET-CAPELLAN,** et **Jean-Luc SCHWARTZ**

préparée au sein du **Laboratoire Gipsa-lab**
dans **l'École Doctorale EDISCE – « Ingénierie pour la Santé, la Cognition et l'Environnement »**

Percevoir la parole quand elle est produite différemment : Étude des mécanismes de familiarisation multimodale/multisensorielle entre locuteurs tout-venant et locuteurs présentant un trouble de l'articulation

Thèse soutenue publiquement le **26 juin 2019**
devant le jury composé de :

Mme Marion DOHEN
Maître de conférence à Grenoble-INP, Directrice de thèse
Mme Valérie HAZAN
Professeure à University College London, Présidente du jury
Mme Martine HENNEQUIN
Professeure à Université de Clermont Auvergne, Examinatrice
Mme Christine MEUNIER
Chargée de recherche au CNRS, Rapporteure
Mme Amélie ROCHET-CAPELLAN
Chargée de recherche au CNRS, Co-directrice de thèse
M. Jean-Luc SCHWARTZ
Directeur de recherche au CNRS, Co-directeur de thèse
M. Rudolph SOCK
Professeur à Université de Strasbourg, Examinateur

Remerciements

Premièrement, je dois remercier mes parents. Tout simplement, pour tout ce qu'ils m'ont donné. C'est une évidence qui passe bien au dessus de la tête, mais je ne serais pas la personne que je suis aujourd'hui sans eux. Cette thèse est le point final à toutes mes années d'études, mais elle me permet de me rappeler le chemin parcouru. Parce que je vous l'ai pas assez dit : merci pour tout, je vous aime.

Ensuite, je souhaite remercier Marion, Amélie et Jean-Luc pour m'avoir accompagné et guidé au long de cette aventure de trois ans (et un peu plus). Merci pour vos conseils, votre patience, et de m'avoir donné les clés pour réussir. J'ai tellement appris dans le domaine de la parole, et avec plaisir grâce à vous trois. J'ai aussi appris sur moi-même, comme quoi il n'est jamais trop tard. Au cours de cette thèse, j'ai beaucoup changé d'un point de vue professionnel et même personnel. La thèse a été pour moi une course sur la durée, avec quelques embûches ; mais je savais que j'avais les bonnes personnes auprès de moi pour arriver jusqu'au bout.

J'aimerais remercier également toutes les personnes que j'ai pu croiser au Gipsa, et il y en a eu. J'ai apprécié toutes les conversations (plus ou moins sérieuses, plus ou moins scientifiques, plus ou moins longues) que j'ai pu avoir avec chacun d'entre vous. J'ai rencontré des personnes de tous les horizons, et allant chacun dans leur propre direction. Et même si ce n'étaient que quelques moments dans une vie, j'ai chacune de ces rencontres m'a reflété sur où j'en étais, et où je me dirigeais dans ma propre vie. Peut-être même que certains d'entre eux ne se reconnaitront pas dans cette description, mais certains méritent entièrement d'être mentionnés ici : Bharat (« my *brother from another mother* »), Cindy (qui a expressément demandé à figurer ici, mais qui n'aura pas la place n°1), tous les supers co-bureaux que j'ai eus (Marie-Lou, J-F, Jonathan, Hélène, ...), la bonne équipe pour sortir sur Grenoble (Sophie, Firas, Andrei, Raùl, Omar, Anne, et tant d'autres), aux plus anciens qui m'ont vus débarquer au Gipsa (Marion, Maël, Diandra, , Adela, ...), et finalement à tous ceux que j'oublie sûrement.

Finalement, j'aimerais remercier ceux qui me connaissent depuis très (trop ?) longtemps, mais qui n'ont sûrement compris ce que je faisais que le jour de ma défense. Sans avoir besoin de comprendre, ils étaient là avant, pendant, et encore après ma thèse. Je parle de la bande de Grenoble (Gaëtan – « c'était trop beau pour être vrai » , Ranc & Nelly, Gaby & Marine, Hoch', Pouni et Clem, ...) et de celle de Valence (« la bande à Hennequin », j'ai entendu dire une fois : Yann, Jim, Cloé & Baston, Soso, Chloé, Melvyn & Tam, Mitch & Claire, Loys & Manu, encore une autre Claire, les exilés parisiens & lyonnais qui manquent à l'appel ici, et l'exilé du Creusot). Des gens que j'ai pas regretté d'avoir comme amis – les gars (et meufs) sûrs – dont je sais ce qu'ils seront libres autour du 13 juillet de chaque année.

Ce manuscrit est le reflet de mon expérience scientifique et mon intérêt dans le domaine de la parole. Elle ne capture pas tous les moments que j'ai pu passer au long de cette aventure, mais ce sont justement ces moments qui ont rendu cette thèse une expérience unique.

Bonne lecture,

Alexandre Hennequin

Table des matières

1 Introduction

La parole est la modalité la plus utilisée par l'homme pour communiquer. Elle est à la fois un instrument et un besoin majeur de l'être humain qui lui permet de construire et de maintenir son identité sociale. En particulier, de la compétence de parole, sont souvent inférées les compétences intellectuelles et sociales. Pourtant, si la parole est omniprésente et semble venir « naturellement » à chacun d'entre nous, la recherche la révèle comme un phénomène complexe à la fois pour le locuteur – qui doit parvenir à articuler plusieurs sons par secondes de manière coordonnée afin d'exprimer son message – et pour l'auditeur – qui devra décoder les sons qu'il perçoit afin de comprendre ce qui lui a été dit par différentes personnes et dans des conditions qui parfois altèrent les signaux véhiculés par le locuteur (cf. environnement bruité, mouvements articulatoires plus ou moins temporairement altérés etc.).

Dans cette thèse, nous portons notre attention sur deux outils dont dispose l'auditeur afin d'améliorer la compréhension de la parole de son interlocuteur. Le premier outil est la multisensorialité de la perception de la parole. Lorsque nous parlons, nous émettons des sons qui sont perçus par les oreilles de notre interlocuteur. Cependant, nous transmettons aussi une information visuelle puisque les gestes articulatoires à l'origine des sons émis sont – au moins en partie et en situation de communication face-à-face – aussi perçus par les yeux de ce même interlocuteur. Dans certaines situations, par exemple lorsque nous discutons avec un ami dans un endroit bruyant, nous nous reposons en fait sur notre vision (en plus de notre ouïe) pour pouvoir le comprendre. Le second outil dont dispose un auditeur pour améliorer sa perception de la parole d'un locuteur donné est la familiarisation envers ce locuteur : plus nous sommes exposés à la parole d'une personne, mieux on la comprend. Pour illustrer ce phénomène, on peut par exemple penser aux enfants en bas âge qui, lorsqu'ils s'expriment, ne sont la plupart du temps compris que par leurs parents. D'une certaine manière, nous nous habituons à la façon dont parle un individu, dans le même but de promouvoir

la communication. Il semble de plus que le système moteur joue un rôle dans cet apprentissage perceptif en le favorisant : si nous imitons notre interlocuteur, nous le percevons ensuite mieux. Cette imitation favoriserait l'alignement des représentations motrices de l'auditeur avec celles du locuteur.

Dans cette thèse, nous nous intéresserons ainsi à la manière dont la vision, lorsqu'elle est disponible à l'auditeur, lui permet d'améliorer son expérience communicative via une meilleure perception de la parole qu'il reçoit. Nous porterons ensuite notre attention sur la façon dont un auditeur extrait et construit des connaissances à partir de la parole à laquelle il est familiarisé en interrogeant le rôle de l'implication du système moteur de l'auditeur lors de cette familiarisation. Nous évaluerons la mise en jeu ces deux mécanismes – de perception multisensorielle et d'apprentissage perceptif actif – dans la perception de la parole de locuteurs atypiques ayant des difficultés de parole liées à des spécificités anatomiques, physiologiques et cognitives. Plus particulièrement nous nous intéresserons à la perception de la parole de locuteurs avec trisomie T21 (T21). La T21 est une pathologie d'origine génétique affectant l'ensemble du phénotype. Plus particulièrement, la parole de l'individu avec T21 est altérée et son intelligibilité est diminuée. Ses capacités communicatives par la parole s'en trouvent ainsi réduites ce qui impacte son insertion dans la société. D'autant plus que le trouble en production de la parole chez les personnes avec une T21 est souvent réduit à la déficience intellectuelle alors qu'on sait aujourd'hui que ça n'est pas le cas puisque les personnes avec T21 comprennent bien mieux la parole qu'elles ne peuvent la produire. Améliorer la parole des personnes avec T21 constitue donc un enjeu crucial de leur intégration. La prise en charge orthophonique permet de palier ou de compenser certains problèmes mais repose uniquement sur une adaptation de la part de la personne qui est déjà en difficultés. La communication est pourtant un acte coopératif qui requiert une adaptation bidirectionnelle des deux interlocuteurs. L'adaptation de l'interlocuteur à son interlocuteur avec T21 représente ainsi une autre manière d'améliorer la communication des personnes avec T21 que nous explorerons dans cette thèse.

Les deux mécanismes communicatifs mentionnés précédemment sont ainsi des candidats potentiels pour atteindre ce but. On ne sait cependant pas si les spécificités des locuteurs avec T21 notamment motrices mais aussi anatomiques et physiologiques ne pourraient pas interférer avec ces mécanismes en les rendant inopérants. Si la perception visuelle de la parole et la familiarisation au locuteur reposent sur un alignement entre les représentations motrices du locuteur avec celles de son interlocuteur, on peut se demander si cet alignement est possible lorsque les deux interlocuteurs ont des systèmes de production de la parole et donc potentiellement des représentations motrices de la parole trop différents. Ainsi, cette thèse proposera deux études expérimentales qui permettront d'en savoir plus sur ce sujet en apportant notamment des éléments de réponses aux questions suivantes :

- Malgré des spécificités motrices, anatomiques et physiologiques, l'information visuelle de la parole des personnes avec T21 est-elle perceptible par des interlocuteurs tout-venant de la même manière qu'elle l'est lorsque le locuteur est lui aussi tout-venant ?
- L'écart d'intelligibilité auditive entre locuteurs avec T21 et tout-venant se réduit-il lorsque l'information visuelle est disponible en plus de l'information auditive ?
- Observe-t-on un phénomène d'apprentissage perceptif sur la parole de locuteurs avec T21 par des personnes tout-venant ?
- Recruter le système moteur en imitant les productions du locuteur avec T21 peut-il jouer un rôle dans ce potentiel apprentissage perceptif ?

Cette thèse s'articule en trois grandes parties. Le premier chapitre présente le contexte théorique dans lequel elle s'inscrit en décrivant d'abord en détails les mécanismes de perception multisensorielle (section 2) et d'apprentissage perceptif de la parole (section 2.2), en s'intéressant ensuite à l'implication du système moteur dans ces processus cognitifs (section 2.4), puis en décrivant le contexte spécifique de la parole des locuteurs avec T21 (section 2.5). Ensuite, un

second chapitre (section 3) est dédié à la description du travail expérimental produit lors de la thèse permettant de fournier des éléments de réponse aux questions abordées ci-dessus. Ce chapitre se scinde deux parties présentant chacune un article scientifique dont l'un a été publié dans le *Journal of Speech, Language, and Hearing Research* en 2018 (section 3.1) et dont l'autre a été soumis pour publication (section 3.2). Ces articles sont rédigés en anglais mais chacun est précédé d'un résumé détaillé de présentation en français. Finalement, le dernier chapitre (section 4) permet de discuter des apports de ce travail relativement à certaines limites et de le mettre en perspective au regard d'aspects cliniques et théoriques.

2 Partie théorique

2.1 La parole est multimodale

2.1.1 Entendre, voir et ressentir la parole

Percevoir la parole peut se concevoir comme la résolution d'un problème dont l'objectif est de retrouver, dans le signal sonore entendu, le message émis par le locuteur. La parole est capable de convoyer à la fois des idées, des sentiments et/ou des pensées propres à un individu. L'information contenue dans la parole doit alors être captée par l'interlocuteur, la perception de la parole ne comporte ainsi pas les mêmes défis lorsqu'on échange des nouvelles avec un ami dans un café au calme, ou dans un hall de gare à une heure d'affluence. Dans cette dernière condition, la perception du signal acoustique de parole est rendue difficile à cause de l'environnement bruité. Malgré la détérioration du signal par le bruit, l'interlocuteur est pourtant capable de le reconstituer. Intéressons-nous maintenant de manière plus approfondie à ce qu'est percevoir la parole et par quelles voies sensorielles on peut le faire (voir Figure 1).

La production de la parole est un phénomène complexe qui met en jeu le système respiratoire et l'intégralité du conduit vocal, de l'inspiration de l'air requis à la génération du son et son expulsion par les poumons, en passant par les cordes vocales puis jusqu'aux lèvres du locuteur. Le déplacement des articulateurs permet de moduler le son issu des cordes vocales pour produire une multitude de sons différents, mais résulte également dans des mouvements au moins en partie visibles ainsi que des phénomènes aérotactiles. Notre système auditif perçoit le signal acoustique de parole. Les mouvements articulatoires sont quant à eux en partie perçus par notre système visuel. Enfin, les phénomènes aérotactiles peuvent être captés par le système somatosensoriel. Ainsi, en conditions de communication face-à-face, la perception de la parole ne se réduit pas seulement à percevoir un son mais bien à capter la parole dans l'ensemble des modalités sensorielles (auditive, visuelle mais aussi dans certains cas tactile) qu'elle peut stimuler chez

l'interlocuteur selon le contexte. Dans chacune de ces modalités, la nature de l'information (respectivement acoustique, optique ou haptique) est captée par différents organes sensoriels. L'information est ensuite transmise au cerveau qui a pour tâche d'intégrer, de décoder et d'extraire un percept. La perception de parole est réussie si ce percept correspond au message émis par le locuteur (voir Figure 1).

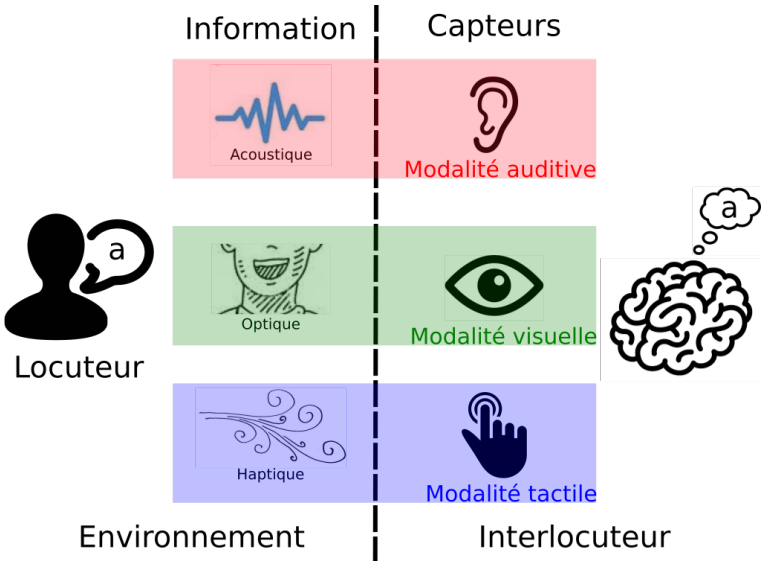


Figure 1 : La perception de la parole est multimodale. Un locuteur qui produit de la parole émet différentes informations qui sont à leur tour captées par l'interlocuteur. Cette information est traitée pour comprendre le message initial.

Le biais qui consiste à réduire la perception de la parole à la simple modalité auditive est « naturel » : dans des conditions favorables, c'est-à-dire quand le signal acoustique est préservé, on peut percevoir la parole d'un locuteur sans avoir ni à le voir ni à utiliser notre sens tactile. Cette capacité à percevoir correctement grâce à l'information acoustique s'est renforcée dans la vie quotidienne par l'essor du téléphone et de la radio par exemple (Rosenblum, 2005). Or, s'il peut sembler évident que la parole est avant tout portée par la modalité auditive (et c'est le cas), l'intuition d'une parole uniquement auditive s'efface dès qu'on considère les personnes sourdes et malentendantes : la modalité auditive n'est plus ou que peu disponible, ces personnes sont néanmoins capables de capter, au moins partiellement, la parole via la modalité visuelle (Campbell, Dodd, & Burnham, 1998) : c'est ce qu'on appelle communément « lire sur les

lèvres ». Les individus tout-venant le font aussi même si chacun en a peu conscience : en fait, quand celle-ci est disponible, nous utilisons l'information visuelle pour percevoir la parole. Par exemple, Bernstein, Demorest et Tucker (2000) se sont intéressés à la perception purement visuelle de la parole par des participants tout-venant ou profondément malentendants. Les résultats, en accord avec le reste de la littérature, ont mis en évidence un taux de reconnaissance de mots variant de 10 à 30% chez l'ensemble des participants tout-venant. On note également un avantage en terme de performances chez les personnes avec un déficit auditif par rapport aux participants tout-venant. Ces performances en perception purement visuelle sont cependant assez faibles en comparaison à la perception via la modalité auditive chez les personnes tout-venant mais indiquent tout de même que la modalité visuelle contient une information pertinente pour la perception de la parole.

De même, les personnes privées à la fois de leur ouïe et de leur vue, sont capables de percevoir la parole par la voie tactile, comme en témoigne la méthode Tadoma (Alcorn, 1932). Cette méthode consiste à « lire » des indices articulatoires par le toucher. La personne place son pouce sur les lèvres du locuteur, et pose ses doigts sur sa joue et sa gorge. Il devient alors possible de capter les mouvements des lèvres, de la mâchoire et la vibration des cordes vocales du locuteur, informations qui permettent de percevoir la parole, au moins en partie. Pourtant, la méthode Tadoma est difficile à mettre en place dans la vie au quotidien, notamment pour des raisons de norme sociale et d'habituation. D'autre part, les évènements aérotactiles émis lors de la production de parole ne permettent pas de véhiculer l'ensemble des traits distinctifs de la parole et donc, n'offrent qu'une perception partielle. Cette méthode nous indique néanmoins que des sensations tactiles, renseignant directement sur la configuration des articulateurs de parole, peuvent contribuer à percevoir cette dernière.

Au quotidien, l'auditeur utilise l'ensemble des informations sensorielles disponibles pour percevoir la parole : la perception de la parole est multisensorielle. Cependant, bien qu'utilisant des informations provenant de plusieurs modalités sensorielles, cette perception doit faire émerger un percept

cohérent tenant compte de l'ensemble des informations relatives à chaque modalité : on parle de fusion ou d'intégration des modalités.

2.1.2 Percevoir la parole : intégrer différentes modalités

Dans la vie de tous les jours, la plupart des interactions se font en face-à-face, rendant disponible l'information visuelle en plus de l'information auditive. Comme nous l'aborderons ensuite, la recherche a porté son intérêt sur l'apport du visuel à la perception de parole depuis les années 1950. En revanche, la fusion des modalités auditive et tactile parait moins évidente. Cependant, chez l'auditeur tout-venant, une stimulation aérotactile peut influencer la perception auditive de la parole, comme l'ont montré Gick et Derrick (2009), via une tâche de catégorisation phonétique des sons /pa/-/ba/ ou /ta/-/da/. Ces paires de sons ont été choisies parce qu'en anglais, un, et un seul, des sons de chaque paire (/pa/ et /ta/) est aspiré (c'est-à-dire qu'il se produit en émettant une bouffée d'air). Lors de la perception de ces sons, les participants ont reçu, sans le savoir, une bouffée d'air sur une partie du corps (la main droite ou le cou) au même instant qu'ils percevaient le son à identifier. L'effet de la bouffée d'air reçue par les participants a modifié leur perception : ils catégorisaient plus souvent les sons comme aspirés lorsque la bouffée d'air était émise. De plus, alors que les participants ont rapporté ne pas avoir été conscients d'avoir reçu une quelconque information tactile, ils ont intégré cette information à l'information auditive, ce qui a guidé leur décision dans la tâche de catégorisation. L'intégration multisensorielle dans la perception de la parole s'observe aussi lorsque la perception visuelle de la parole est accompagnée d'une information tactile (Bicevskis, Derrick, & Gick, 2016).

La fusion des informations provenant de différentes modalités pour la perception de la parole a d'abord été étudiée pour les modalités auditives et visuelles. Les résultats de l'expérience de McGurk et MacDonald (1976), certainement la plus connue dans ce domaine, illustre bien cette fusion. Les auteurs ont demandé à des adultes de catégoriser une série de stimuli audiovisuels incohérents consistant en la combinaison d'un stimulus audio d'un locuteur produisant un /ba/ en synchronie avec le stimulus vidéo du même locuteur

articulant un /ga/. Ces stimuli audiovisuels ont été perçus comme /da/ par la plupart des participants, ce qui ne correspond ni au percept en modalité auditive seule, ni au percept en modalité visuelle seule. Ce percept serait issu de la fusion des informations auditive et visuelle. Ce résultat met en avant que : (1) l'information visuelle peut être aussi importante pour la perception de parole que l'information auditive : lorsque les informations entre modalités sont incohérentes, la fusion ne favorise pas nécessairement l'une ou l'autre des deux modalités ; (2) la fusion des informations multimodales en perception de parole semble automatique et notre cerveau utilise toute information perçue (même incohérente) afin de créer un percept valide à partir des informations qui nous sont présentées. Bien que la parole soit perçue par des capteurs unimodaux, la cooccurrence des informations captées aboutit à un percept qui intègre l'ensemble des informations relatives à chaque modalité.

Afin d'intégrer l'information portée par des stimuli multimodaux, on peut supposer que, dans une certaine mesure, les signaux relatifs à chaque modalité partagent une certaine information afin d'être associés, puis éventuellement intégrés. L'association entre modalités auditive et visuelle pour la perception de la parole est observable dès le plus jeune âge. Kuhl et Meltzoff (1982) ont réalisé une étude dans laquelle des bébés de 18 à 20 semaines étaient installés devant un écran sur lequel était affiché deux visages qui articulaient de manière synchrone chacun une voyelle différente (/a/ ou /i/). Les bébés percevaient simultanément un signal audio correspondant à une seule des deux voyelles prononcées par les visages. Une caméra était placée en face d'eux afin de savoir sur quel visage se portait leur attention (voir Figure 2).

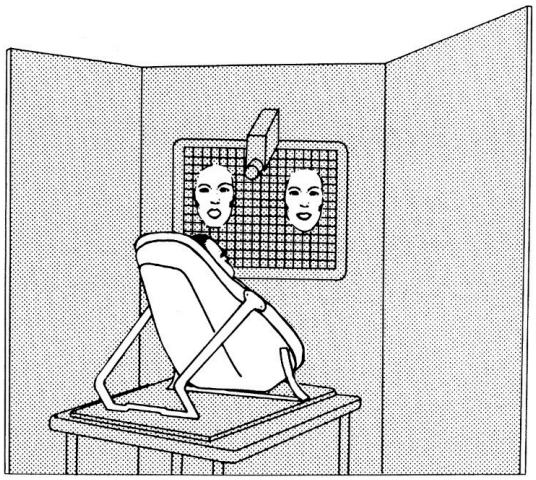


Figure 2 : Extrait de Kuhl & Meltzoff (1982). Illustration du dispositif expérimental utilisé dans leur expérience : le bébé est installé devant un écran avec deux visages. Une caméra le filme pour savoir où il regarde.

Les données ont montré que les bébés préféraient regarder le visage qui était cohérent avec le stimulus audio qu'ils entendaient, c'est-à-dire celui qui articulait la même voyelle que celle qu'ils percevaient auditivement. Dans une expérience supplémentaire, les auteurs ont manipulé le signal audio en supprimant l'information spectrale tout en conservant l'information temporelle. Plus simplement, les voyelles étaient remplacées par des tons de même durée et à la même fréquence fondamentale moyenne. Dans cette condition, les enfants observaient autant chaque visage présenté : il n'y avait plus de préférence pour le visage produisant la voyelle correspondante. Ces deux résultats montrent que l'enfant, dès son plus jeune âge, associe les modalités visuelle et auditive lorsqu'il perçoit la parole. De plus, la détection d'une telle correspondance est soumise à condition : la modalité auditive doit contenir l'information (spectrale) pertinente afin d'être associée avec ce qui est perçu visuellement (Rosen, Fourcin, & Moore, 1981).

2.1.3 Intégrer pour mieux percevoir

Le locuteur, afin d'être compris quel que soient l'environnement et l'auditeur, peut adopter un certain nombre de stratégies modifiant ses productions sonores (Cooke, King, Garnier, & Aubanel, 2014 ; Hazan, Tuomainen, Kim, Davis, Sheffield & Brungart, 2018). Dans la vie de tous les jours, simplement écouter son

interlocuteur permet de le comprendre. Cependant, nous avons vu que l'ajout d'une modalité, tactile ou visuelle, lors de la perception de parole peut changer le percept obtenu même dans une condition dans laquelle le stimulus auditif est parfaitement clair (Bicevskis et al., 2016 ; Gick & Derrick, 2009 ; McGurk & MacDonald, 1976). Désormais, notre intérêt se portera uniquement sur la perception audio-visuelle, notamment parce qu'elle représente la situation la plus écologique lors de la communication entre deux personnes sans déficits sensoriels.

Dans des conditions idéales, la perception auditive atteint un niveau de compréhension quasi-optimal. On parle alors d'« effet plafond » : la mesure observée atteint sa limite supérieure et on ne peut donc plus observer d'amélioration. Inversement, on parle d'« effet plancher » lorsqu'une mesure perceptive est tellement faible qu'il est impossible d'en observer une variation.

Pour étudier la contribution de la modalité visuelle dans la perception de la parole, une méthode consiste à réduire la qualité de l'information auditive afin de s'affranchir de l'effet plafond et ainsi pouvoir comparer la perception auditive à la perception audio-visuelle. C'est l'idée de l'expérience princeps présentée par Sumby et Pollack (1954). À différents niveaux de Rapport Signal sur Bruit (RSB ; allant de -30 dB jusque 0 dB par tranche de 6 dB), des groupes de participants avaient pour tâche de reconnaître des mots présentés parmi une liste de vocabulaire dont la taille (nombre de réponses possibles) variait (8, 16, 32, 64, 128 et 256 mots). Ils faisaient ça soit auditivement (A) soit audiovisuellement (AV). Les performances des participants (voir Figure 3) étaient dépendantes de la taille du vocabulaire, à savoir que les résultats (pourcentage de réponses correctes) étaient les meilleurs pour la plus petite liste. Plus la taille du vocabulaire était grande, moins bonnes étaient les performances. Ce résultat paraît assez logique puisque plus le vocabulaire est grand plus le nombre de réponses possibles est important. L'autre variable d'intérêt était le niveau de bruit : ainsi, la reconnaissance correcte des mots était au plus bas au niveau de bruit le plus défavorable (RSB = -30dB) avec environ 5% de bonnes réponses en modalité A et environ 40% en AV. Ces performances augmentent au fur et à mesure que le RSB croît et atteignent un maximum de 90-95% de mots reconnus à 0 dB dans les deux modalités A et AV,

signalant ainsi un effet plafond. Ce résultat nous indique que plus l'information auditive est claire, meilleure sont les performances des participants et ce dans les deux modalités observées. Le point important est que lorsque la perception auditive était dégradée par du bruit, la modalité visuelle apportait une information significative qui améliorait la perception : par exemple pour un RSB de -30dB, on passe de 5% de réponses correctes en A à 40% en AV.

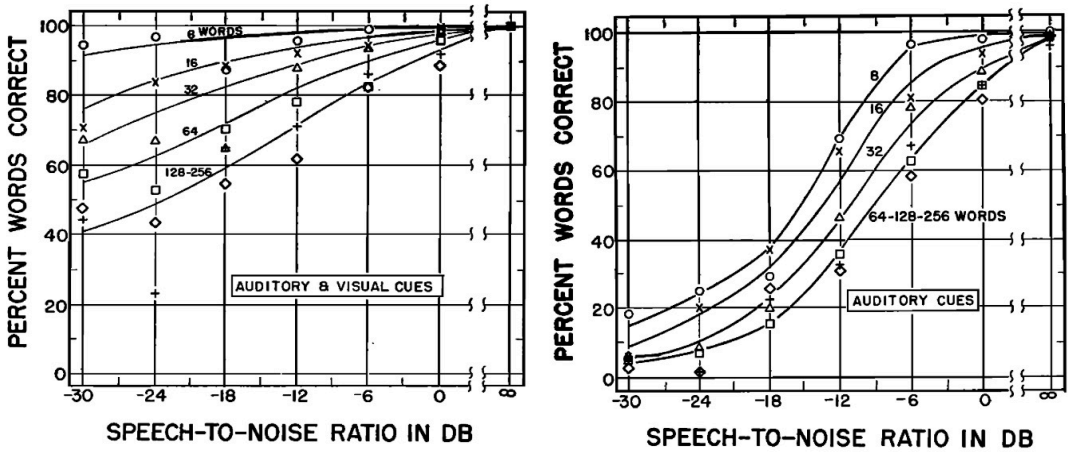


Figure 3 : Extraits de Sumby & Pollack (1954). Les graphes représentent le pourcentage de mots correctement identifiés (de 0 à 100%) en fonction du rapport signal sur bruit (SPEECH-TO-NOISE RATIO variant de -30 à ∞ dB), en modalité AV à gauche et en modalité A à droite. Chaque courbe représente un groupe de participants ayant effectué un test de reconnaissance de mots avec une liste de mots possibles de taille différente (8, 16, 32, 64, 128, 256).

Toujours dans le but d'observer un bénéfice de la modalité visuelle pour la perception de la parole, une autre méthode consiste à rendre la parole auditive plus difficile à *comprendre* plutôt que plus difficile à *entendre*. En d'autres termes, le signal acoustique est alors intact, mais le message convoyé est complexe. Cette complexité peut par exemple être manipulée grâce au contenu sémantique du message (*i.e.* relatif au sens), ou grâce à un accent non familier ou encore en utilisant de la parole prononcée par un locuteur non natif (Reisberg, McLean, & Goldfield, 1987). Une étude (Arnold & Hill, 2001) examine la compréhension du message parlé dans trois conditions : d'abord, des apprenants du français comme seconde langue (ayant l'anglais pour première langue) ont été testés sur leur perception du français ; ensuite, des natifs anglophones ont perçu la parole d'une personne avec un accent écossais ; et enfin, des participants natifs anglophones ont été testés sur leur compréhension de la lecture d'un texte syntaxiquement et

sémantiquement complexe. Dans chacune des expériences, la parole était présentée soit auditivement seulement (A) soit audiovisuellement (AV). Les participants devaient répondre à une série de questions pour tester leur compréhension de l'extrait perçu. Afin de s'assurer que la parole utilisée dans les trois expériences était parfaitement audible, d'autres participants ont eu pour tâche de répéter les phrases présentées : ils y sont parfaitement parvenus. Bien que le message soit parfaitement audible, les résultats en compréhension dans la modalité A n'atteignaient pas les valeurs plafonds, et l'ajout de la modalité visuelle améliorait les performances des participants. Les résultats de cette étude montrent un bénéfice en modalité AV par rapport à la modalité A lorsque la parole est clairement audible et sont concordants avec le reste de la littérature à ce propos (Davis & Kim, 2004 ; Reisberg et al., 1987). Même confronté à une parole acoustiquement parfaitement audible, on peut observer un bénéfice de la présence de l'information visuelle.

En complément, on peut se poser la question de savoir sur quoi porte l'attention d'un interlocuteur lorsqu'il perçoit la parole visuellement. En d'autres termes, que voit-on qui nous aide à mieux percevoir ? Deux études se sont intéressées à cette question (Bernstein, Auer, & Takayanagi, 2004 ; Summerfield, 1979) : des participants percevant des phrases plongées dans du bruit avaient pour tâche de les retranscrire dans 5 conditions dont une présentait la parole en modalité auditive seule, et 4 autres conditions en modalité audiovisuelle. Dans ces dernières conditions, les contenus visuels présentés étaient différents, à savoir : le visage du locuteur, les lèvres du locuteur, des points caractérisant les mouvements des coins et milieux des lèvres du locuteur, ou un cercle dont l'amplitude variait proportionnellement à l'intensité du signal audio. L'exactitude de transcription des mots a été évaluée par participant afin de comparer les différentes conditions. Les résultats montrent que seules les conditions avec une partie ou l'ensemble du visage du locuteur ont permis une meilleure perception par rapport à la condition auditive seule. C'est donc le fait de voir le visage ou même simplement les lèvres du locuteur qui permet à l'interlocuteur de mieux comprendre la parole. Dans les deux dernières conditions, il y a en effet des informations visuelles potentiellement

utiles à la perception mais celles-ci n'ont pas permis d'améliorer les performances par rapport à la condition auditive. De plus, il est important de noter que toutes les conditions avec une composante visuelle ont permis aux participants de mieux détecter la présence de parole dans le bruit (sans en comprendre son contenu) par rapport à la condition auditive seule. Même quand l'information visuelle n'aidait pas au décodage du message, elle a permis de mieux en détecter la présence. De plus, elle réduirait la charge cognitive mise en œuvre afin de percevoir la parole (Strand, Brown, & Barbour, 2018).

La question suivante est celle des mécanismes permettant d'intégrer les informations visuelle et auditive. Summerfield (1987) suggère que les mouvements des lèvres lors de la production de parole pourraient être reliés à des propriétés acoustiques du signal auditif produit. Les informations perçues auditivement et visuellement seraient, à un certain degré, reliées l'une à l'autre. Dans cette direction, Grant et Seitz (2000) ont mené une expérience dans laquelle 10 participants tout-venant avaient pour tâche de détecter des phrases, dans du bruit, dans trois modalités : auditive seule (A), auditive accompagnée d'information visuelle cohérente (AVc) et auditive accompagnée d'information visuelle incohérente (AVi). Les résultats montrent que les participants détectent mieux les phrases dans la condition AVc par rapport aux conditions A et AVi. La présence d'une information visuelle améliore la détection de la phrase énoncée seulement si elle est cohérente avec l'information auditive. Les auteurs se sont ensuite intéressés à la corrélation entre informations auditive et visuelle. Ils ont trouvé une relation entre l'aire d'ouverture de la bouche et les propriétés spectrales du signal. Conjointement aux résultats comportementaux précédents, la facilitation de la perception de la parole ne serait possible que dans le cas d'une information cohérente entre modalités auditive et visuelle. L'interlocuteur serait alors guidé par l'information visuelle de manière temporelle, à savoir que les mouvements articulatoires sont temporellement proches de changements dans le signal acoustique ; au niveau spectral, l'information visuelle pourrait informer sur les propriétés du signal acoustique, et notamment sur les fluctuations de son enveloppe au cours du temps. Schwartz, Berthommier et Savariaux (2004) se sont

intéressés à cette question en menant une expérience de catégorisation perceptive des voyelles /y/ et /u/ en modalités auditive (A), visuelle (V) et audiovisuelle (AV). L'information visuelle ne permet pas de catégoriser ces deux voyelles et les résultats en V le confirment puisqu'ils ne sont pas différents du hasard. Pourtant, les auteurs trouvent que les performances en AV sont significativement meilleures que celles en A. Ils interprètent cela en montrant que les modifications articulatoires (aire intéro-labiale) précèdent les modifications acoustiques. Bien que non pertinente en tant que telle, l'information visuelle permettrait donc de savoir à quel moment l'information acoustique pertinente va apparaître et ainsi de mieux la percevoir.

Si la redondance entre informations auditive et visuelle joue un rôle important dans les mécanismes de fusion et détection de parole, la nature très différente de ces deux informations pourrait suggérer que les indices à extraire dans chaque modalité puissent en partie être complémentaires. De plus, certaines situations de communication pourraient privilégier un canal de communication plutôt qu'un autre. Si certains auteurs ont supposé que l'information visuelle serait d'autant plus utile que l'information auditive est appauvrie (Erber, 1969 ; Sumby & Pollack, 1954), cette idée a été ensuite contestée (Ross, Saint-Amour, Leavitt, Javitt, & Foxe, 2007). En effet, Ross et collègues ont étudié les perceptions auditive seule (A), visuelle seule (V) et audio-visuelle (AV) de mots monosyllabiques présentés à différents niveaux de rapport signal sur bruit (RSB ; de -24 à 0 dB par pas de 4 dB). Les pourcentages de réponses correctes (cf. Figure 4) en condition A varient de 0 à 85% et de 20% à 95% en condition AV. Les auteurs se sont ensuite intéressés au gain apporté par la multisensorialité, qu'ils ont défini comme la différence entre les performances en modalités AV et A (cf. Figure 4). Ils observent un pic de cette mesure à un RSB de -12 dB qui correspond à la valeur moyenne des RSB pour lesquels : 1. les participants ne peuvent plus extraire aucune information acoustique du signal de parole et ne peuvent donc plus que se reposer sur la modalité V, c'est-à-dire -24dB (puisque, pour ce RSB, le score en modalité A est de 0%) et 2. les participants pourraient ne se reposer que sur la modalité A, c'est-à-dire 0dB (puisque, pour ce RSB, le score en modalité A est proche du plafond de

100%). Contrairement à l'idée soumise par Sumby et Pollack (1954), le bénéfice apporté par l'intégration des modalités est maximal non pas lorsque l'une des deux modalités A ou V est la plus dégradée, mais lorsque les deux modalités peuvent partiellement contribuer à percevoir correctement. Enfin, la perception en modalité AV n'est pas égale à la somme des performances en modalités A et V : la performance brute en présence d'information multimodale à un RSB de -12 dB est en effet de 45%, ce qui est supérieur à la somme des performances en modalités unimodales A (20%) et V (<10%).

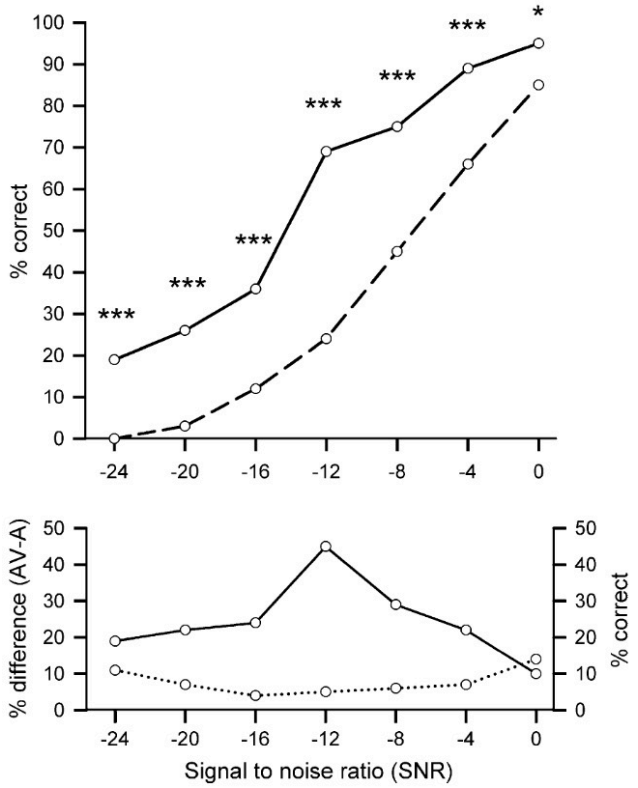


Figure 4 : Extrait de Ross et al. (2007). En haut : pourcentage de réponses correctes en fonction du RSB en conditions A (pointillés) et AV (trait plein), les étoiles illustrent une différence significative entre modalités pour un RSB donné. En bas : différences entre les conditions A et AV en pourcentage (trait plein, ordonnée à gauche) et pourcentage de réponses correctes en condition V (pointillés, ordonnée à droite) en fonction du RSB.

Cette observation s'interprète par le fait que les modalités sont non seulement redondantes mais aussi complémentaires (Binnie, Montgomery, & Jackson, 1974). Pour quantifier et illustrer ce phénomène, Summerfield (1987) a analysé les résultats de données sur les perceptions auditive et visuelle des consonnes de l'anglais. Pour la perception auditive, il a repris les données de

plusieurs études de discrimination des consonnes à différents niveaux de RSB. Cela a permis de construire plusieurs matrices de confusions, qui ont été traduites en un classement arborescent des consonnes (à droite, voir Figure 5). Au bas de l'arbre (aux RSB les plus défavorables), les consonnes sont largement confondues et donc regroupées en une seule classe (le « tronc » de l'arbre). À chaque étape, et au fur et à mesure que le RSB devient plus favorable, les consonnes sont de plus en plus facilement discernables et se séparent dans l'arborescence. À la dernière étape, toutes les consonnes sont clairement séparées les unes des autres dans les « branches » ultimes. Plus deux consonnes se séparent tard dans l'arbre et plus elles sont difficiles à différencier. L'auteur reprend également des données permettant de proposer un tel arbre pour la perception visuelle. Le procédé ne pouvant être le même (pas de différents RSB en visuel), le classement par arborescence a été obtenu à partir de l'unique matrice de confusions en utilisant une classification hiérarchique dans laquelle chaque étape dissocie la consonne la moins confondue avec toutes les autres.

À l'issue de la construction d'une telle arborescence, il est notable que les consonnes sont ordonnées selon des traits articulatoires. Il est alors possible d'identifier dans chaque modalité les traits qui sont les plus robustes, c'est-à-dire pour lesquels la séparation de deux consonnes distinctes par ce trait se fait au plus tôt. Pour la modalité auditive, c'est le mode d'articulation qui apparaît comme le trait le plus saillant, alors que c'est le lieu d'articulation pour la modalité visuelle. Ainsi, deux consonnes facilement confondues dans une modalité peuvent être éventuellement facilement discriminées dans l'autre. Par exemple, /b/-/p/ sont bien discriminées auditivement, mais pas visuellement, parce qu'elles sont caractérisées par le même geste de fermeture des lèvres, et inversement /b/ et /d/ sont plus facilement confondues dans le bruit puisqu'elles partagent le même mode d'articulation, mais se séparent clairement visuellement par la présence ou non d'un geste d'occlusion labiale. De même, les voyelles du français /i/, /a/ et /u/ ne sont pas également reconnaissables selon la modalité dans laquelle elles sont perçues (Benoit, Mohamadi, & Kandel, 1994). Indépendamment du contexte, /a/ est mieux reconnue en modalité auditive seule suivie de /i/ puis /y/. Par contre, en

modalité visuelle seule, c'est /y/ qui est la mieux reconnue, puis /a/ et enfin /i/. L'information extraite dans chacune des modalités peut être complémentaire : par exemple, le voisement, qui est déterminé par la vibration des cordes vocales, sera un trait plus facile à entendre qu'à voir. L'effet du bruit impacterait donc différemment les traits phonétiques en fonction de leur saillance dans la modalité auditive (Miller & Nicely, 1955). De plus, la détection d'un trait phonétique dans une modalité pourrait être améliorée par la présence d'un autre trait plus saillant dans une autre modalité. Par exemple, la présence de voisement aide à la perception du lieu d'articulation des consonnes (Alm & Behne, 2008 ; Alm, Behne, & Wang, 2009). Inversement, une information visuelle sur le lieu d'articulation peut influencer la perception du voisement (Eg & Behne, 2009).

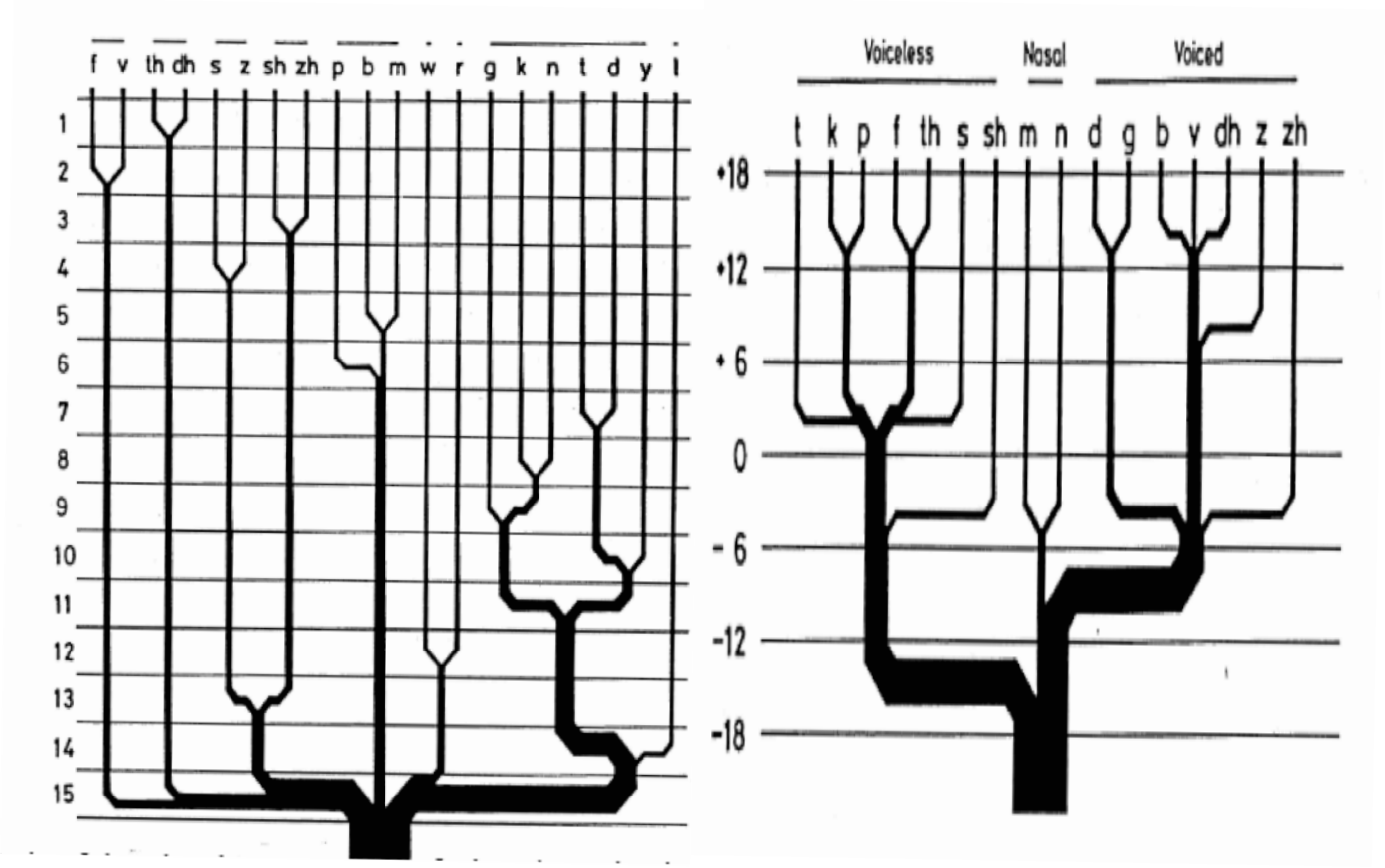


Figure 5 : Extrait de Summerfield (1987). À gauche, le classement arborescent obtenu en modalité visuelle, et à droite celui issu de la modalité auditive. Les détails concernant ces figures sont décrits dans le texte.

2.1.4 Modéliser l'intégration des modalités

Le processus d'intégration multimodale prend en entrée un signal de parole, perçu à travers différentes modalités, pour aboutir à un percept final. Différents phénomènes ont été précédemment décrits afin de comprendre les bénéfices de ce processus. Il est également important de tenter de théoriser ce processus afin de mieux comprendre et prévoir son fonctionnement.

Notamment, l'intégration elle-même peut avoir lieu à plusieurs stades dans le processus de perception et avoir des implications importantes dans la façon d'appréhender ce qu'est percevoir, même de manière unimodale. Ainsi, Peelle et Sommers (2015) dénombrent (en reprenant des propositions classiques) trois architectures d'intégration possibles (voir Figure 6). La première, nommée « *late integration* » (intégration tardive) suggère que les informations sont d'abord traitées unimodalement pour ensuite être intégrées. Dans un premier temps, il n'y aurait donc pas d'interaction lors du traitement des informations mais l'intégration tardive permettrait d'ajuster la combinaison des deux modalités en fonction de la pertinence de chacune. La seconde architecture, nommée « *early integration* » (intégration précoce) propose que le traitement des informations unimodales peut être influencé par celui d'une autre modalité. Classiquement, l'information auditive est prise comme référence – bien que nous ayons vu que la modalité auditive n'est pas systématiquement dominante sur les autres modalités. Néanmoins, ce modèle permet d'expliquer comment l'information visuelle aide à mieux détecter un signal audio (Grant & Seitz, 2000). Finalement, la dernière proposition est une combinaison des deux premières : ainsi, cette version bénéficie des avantages de chaque structure en termes d'adaptabilité (*late integration*) et de détection précoce (*early integration*).

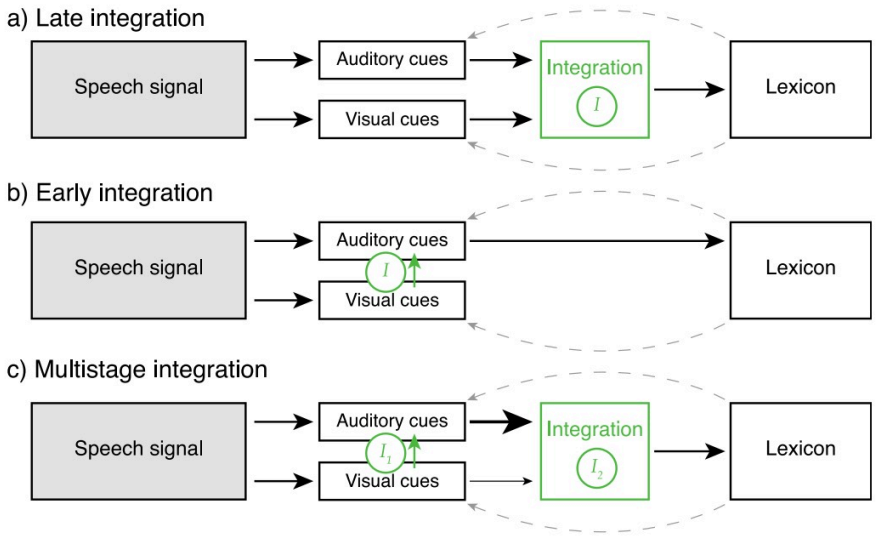


Figure 6 : Extrait de Peelle & Sommers (2015). Schémas décrivant les trois types de modèles d'intégration des signaux multimodaux de parole (les flèches en pointillés font référence à des boucles de rétroaction qui ne nous intéressent pas ici). *Late integration* = intégration tardive – *early integration* = intégration précoce – *multistage integration* = intégration multi-étapes – *speech signal* = signal de parole – *auditory cues* = indices auditifs – *visual cues* = indices visuels – *lexicon* = lexique.

Ces architectures rentrent dans une taxinomie plus globale des modèles de perception multimodale (Schwartz, Robert-Ribes, & Escudier, 1998). Parmi les modèles de cette taxinomie, plutôt qu'une dominance d'une modalité sur l'autre, l'idée soutenue est celle d'une parole multimodale par nature : la parole perçue dans une modalité serait dérivée d'un « objet parole », lui-même amodal (Rosenblum, 2005, 2008). Cet objet amodal aurait des propriétés sensorielles variées (auditives, visuelles, tactiles), et c'est en référence aux propriétés amodales de l'objet que l'intégration s'effectuerait.

Ceci a conduit à une étude visant à tester si l'information fournie par un locuteur dans une modalité spécifique peut aider à mieux envisager la parole d'un même locuteur dans une autre modalité (von Kriegstein et al., 2008). Des participants ont d'abord suivi un entraînement dans lequel leur étaient présentés six locuteurs, chacun pour une durée inférieure à deux minutes. Cet entraînement présentait la parole associée au nom des locuteurs, mais le visage de ceux-ci n'était présenté que pour la moitié des locuteurs ; les autres locuteurs étaient visuellement représentés par une image fixe arbitraire. Dans un second temps, les participants ont effectué une tâche de reconnaissance de locuteur ainsi qu'une

tâche de reconnaissance de parole. Les résultats montrent un bénéfice dans ces deux tâches pour les locuteurs dont les visages avaient été observés, en comparaison des autres. Les participants ont donc utilisé l'information visuelle du locuteur afin de mieux le percevoir dans deux tâches purement auditives.

Ce phénomène correspond à un transfert entre les modalités : l'information recueillie dans une modalité est à la fois utile et bénéfique pour résoudre un problème dans une autre. Dans une autre étude (Rosenblum, Miller, & Sanchez, 2007), des participants ont réalisé une tâche de reconnaissance de mots par lecture labiale sur un locuteur (pendant une durée d'une heure) en ne voyant que la moitié basse du visage. À la suite de cette tâche, ils ont participé à un test purement auditif dans lequel ils entendaient soit le même locuteur soit un nouveau locuteur et devaient répéter les mots perçus inclus dans des phrases bruitées. Les performances des participants dans la seconde tâche étaient meilleures lorsque le locuteur qu'ils entendaient était le même que celui qu'ils avaient vu dans la première tâche. L'expérience miroir a également été concluante (Sanchez, Dias, & Rosenblum, 2013) : des participants ayant préalablement entendu un locuteur ont été plus performants dans une tâche de lecture labiale pour ce même locuteur que pour un nouveau locuteur. La connaissance d'un locuteur dans une modalité améliore donc sa perception dans l'autre modalité. En d'autres termes, ce l'on a entendu précédemment d'un locuteur permet de mieux identifier visuellement ce qu'il dit maintenant (et inversement).

Des effets de transfert s'observent aussi dans une même modalité : l'entraînement ou la familiarisation à percevoir la parole produite par un locuteur *via* la modalité auditive peut faciliter la compréhension d'autres énoncés produit par ce même locuteur. Cet aspect est particulièrement intéressant quand il s'agit d'améliorer l'intelligibilité.

2.2 Apprendre à percevoir (ou à mieux percevoir) la parole

2.2.1 Percevoir des sons, entendre un phonème

Un phonème est une représentation mentale de la catégorie associée à un ensemble de sons. Les phonèmes d'une langue donnée sont distribués en un ensemble fini de catégories distinctes (Liberman, Harris, Hoffman, & Griffith, 1957). Cependant, ces catégories sont construites sur un domaine continu (typiquement celui des paramètres acoustiques) : par exemple, un son proche de la limite des catégories /b/ ou /d/ est néanmoins perçu comme appartenant soit à l'une soit à l'autre catégorie. Bien que le son soit ambigu, il est perçu comme un phonème et un seul. Cette perception en catégories distinctes est également attestée dans la perception des voyelles (Pisoni, 1975). L'association son-phonème ainsi que les limites entre phonèmes sont plutôt claires (voir Figure 7, issue de Peterson & Barney, 1952), et ce malgré les variabilités entre locuteurs et auditeurs.

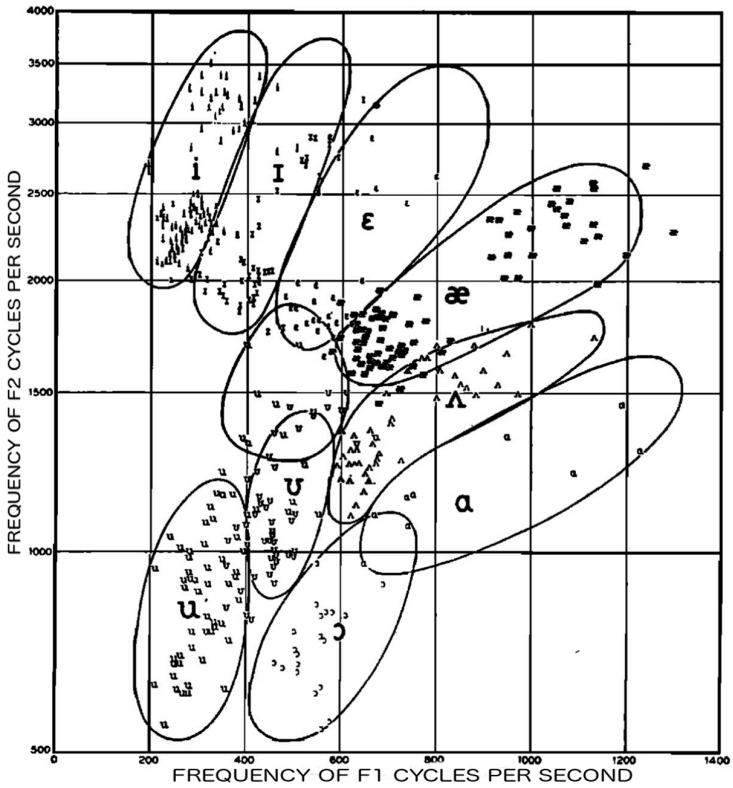


Figure 7 : Extrait de Peterson & Barney (1952). Fréquence du second formant (en ordonnée) en fonction de celle du premier formant (en abscisse). Chaque point représente une voyelle produit par un des locuteurs (homme, femme, ou enfant ; n = 76). Les labels associés à chaque point ont été décidés sur la base de l'accord d'une cohorte d'auditeurs (n = 70).

Les variations dans la production des sons dépendent de facteurs intra- mais aussi inter-locuteurs tels que l'âge, le sexe ou le dialecte (Kraljic, Brennan, & Samuel, 2008). On parle d'allophones pour désigner les variantes de la réalisation sonore d'un phonème : la *jota /r/* et le son /ʁ/ sont deux phonèmes en espagnol, mais deux allophones du seul phonème existant dans le répertoire de l'auditeur français – même si un auditeur français entendra une différence entre les sons, il les catégorisera comme un /ʁ/. Parmi l'ensemble des sons que l'humain est capable de produire, il utilise un nombre restreint de phonèmes afin de communiquer. L'apprentissage phonétique peut être décrit comme le fait de cartographier cet ensemble de phonèmes, spécifique et nécessaire à l'auditeur, qui en retour influe sur sa façon de percevoir les sons à travers différentes langues (Strange, 1995). Ce processus, dérivé de l'expérience de l'auditeur, est unique à chacun. Pourtant, il permet de communiquer avec tous. Ce chapitre se concentrera sur cette relation sons-phonèmes et les mécanismes qui interagissent avec elle pour assurer la communication.

2.2.2 Qu'est-ce qu'être un bon auditeur ?

Le système de perception de parole se développe très tôt dans la vie d'un être humain. Déjà dans le ventre de sa mère, le nourrisson perçoit sa voix et montre une préférence pour celle-ci par rapport à d'autres voix après la naissance (DeCasper & Spence, 1986). Au cours de leur première année de vie, le système perceptif des nourrissons évolue. Lorsqu'ils ont été exposés à une paire de phonèmes qui constitue un contraste phonologique en hindi mais pas en anglais, des nourrissons anglais âgés de 7 mois ont été capables de discriminer les deux sons hindi, tout comme des locuteurs adultes de l'hindi (Werker, Gilbert, Humphrey, & Tees, 1981). Par contre, des locuteurs adultes anglais échouent à discriminer ce contraste qui n'existe pas dans leur langue. Une autre étude Werker et Tees (1984) ont reproduit ce résultat pour un autre contraste d'une autre langue (Salish) et ont montré que cette capacité à discriminer des phonèmes d'une autre langue disparaît à la fin de la première année de vie. Un nourrisson de moins d'un an est-il alors un meilleur auditeur qu'un adulte ? Dans une tâche de discrimination

de sons non-natifs : oui. Mais la perte de sensibilité à discriminer deux sons qui advient par la suite sera compensée par une spécialisation à distinguer les phonèmes d'une langue particulière. C'est l'expérience linguistique qui permettra à l 'enfant d'acquérir une sensibilité accrue aux sons de sa langue (Jusczyk, Friederici, Wessels, Svenkerud, & Jusczyk, 1993). Par la suite, cette sensibilité lui permettra d'apprendre sa langue et de communiquer via le langage oral (Werker & Yeung, 2005).

Un des critères pour être un « bon » auditeur de parole serait donc de pouvoir associer un son donné à la bonne catégorie phonologique, qu'on suppose partagée avec le locuteur. Chaque auditeur se spécialiserait très tôt dans sa vie à percevoir un nombre fini de phonèmes, via l'expérience de la langue. Cela lui permettrait de parvenir à associer une multitude de sons très variables à un nombre limité de catégories pertinentes pour sa langue maternelle. Cette spécialisation est associée à une discrimination réduite à l'intérieur des catégories phonétiques pertinentes, à l'inverse de la perception aux limites de ces mêmes catégories (Kuhl & Iverson, 1995). La perception d'un son de parole est façonnée par les connaissances de l'auditeur. Ces connaissances peuvent être phonétiques, mais également lexicales, comme le montre l'expérience de Ganong (1980). Dans cette expérience, un continuum de sons a été préparé entre le mot anglais *dash* (en français, « tiret ») et le pseudo-mot *tash* (sans signification). Des participants devaient catégoriser quel mot ils entendaient. Les stimuli au milieu du continuum, habituellement perçus comme les plus ambigus dans les expériences de perception catégorielle, étaient plus souvent perçus comme étant le mot plutôt que le pseudo-mot. Les sons ambigus ont donc été plus souvent associés au phonème en accord avec la connaissance lexicale de l'auditeur. Les connaissances préalables (appelées aussi « *priors* ») de l'auditeur semblent façonner sa perception de la parole. Dans une autre étude (Niedzielski, 1999), des participants résidant dans la même ville ont participé à un test perceptif dans lequel ils entendaient des stimuli synthétisés. Ils avaient pour tâche d'identifier parmi ces stimuli synthétiques ceux qui correspondaient le mieux à la parole d'un locuteur réel. La moitié des participants croyaient que le locuteur était originaire de la même région qu'eux alors que

l'autre moitié pensait qu'il venait d'une autre région d'un pays frontalier. Les participants écoutaient les mêmes stimuli mais avaient donc un *prior* différent quant à la région d'où venait le locuteur. Les résultats montrent que les deux groupes de participants n'ont pas choisi les mêmes réponses parmi l'ensemble des stimuli présentés. Bien qu'il s'agisse strictement des mêmes sons, des stéréotypes liés aux attentes entrent en jeu dans la perception du locuteur. Ainsi, on sait que l'information géolinguistique relative au locuteur constitue un *prior* influant sur la perception de la parole.

Tout comme la production des sons varie d'un locuteur à l'autre, la capacité à les percevoir varie en fonction de l'auditeur. La perception de la parole n'est pas déterminée uniquement par le locuteur ou l'auditeur, mais bien par la relation entre les deux. Certains travaux ont d'ailleurs montré qu'un auditeur, lorsqu'il est confronté à la même liste de stimuli, les perçoit mieux lorsqu'ils sont produits par un unique locuteur plutôt que plusieurs (Creelman, 1957 ; Mullennix, Pisoni, & Martin, 1989). L'auditeur adapterait son système perceptif à la parole du locuteur, et cet avantage, bien qu'assez faible, se perdrait lorsque l'auditeur est aussi confronté aux stimuli d'autres locuteurs (voir Figure 8). Néanmoins, cet effet de congruence entre auditeur et locuteur peut produire des effets de généralisation, puisque des auditeurs bénéficient d'un avantage d'intelligibilité lors de la perception d'un locuteur produisant de la parole dans une langue étrangère lorsqu'ils partagent la même langue maternelle (Bent & Bradlow, 2016). De la même manière, la perception d'une langue seconde requiert plus d'efforts à l'auditeur que lorsqu'il perçoit sa langue maternelle (Borghini & Hazan, 2018).

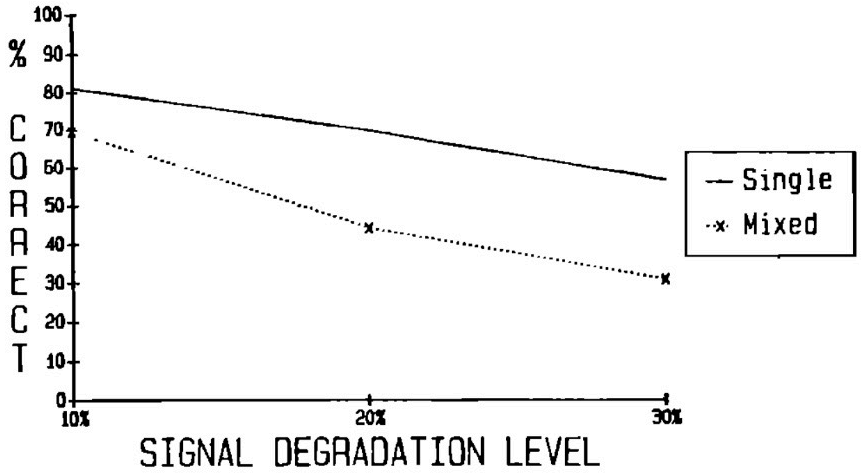


Figure 8 : Extrait de Mullennix et al. (1989). Pourcentages de réponses correctes (en ordonnée) en fonction du niveau de dégradation du signal de parole (en abscisse) et du nombre de locuteurs présentés (trait plein : un seul / pointillés : plusieurs).

2.3 Nos représentations phonologiques s'adaptent à la parole qu'on perçoit

L'apprentissage perceptif est défini comme un mécanisme d'adaptation qui se base sur l'information fournie par l'environnement pour mieux appréhender ce même environnement (Goldstone, 1998). Un grand nombre d'études ont démontré des effets d'apprentissage perceptif dans le domaine de la parole (pour une revue, voir Samuel & Kraljic, 2009). Pour observer les changements apportés par cet apprentissage, il existe deux types de paradigmes expérimentaux. Le premier paradigme s'intéresse directement à l'étude des conditions d'amélioration de la perception, il est dénommé « familiarisation ». Dans ce paradigme, un auditeur est confronté à des stimuli d'apprentissage d'un certain type de parole et on cherche à savoir si cela l'aide ensuite à mieux comprendre cette parole. Le second paradigme est le « recalibrage phonétique » au cours duquel le système phonétique d'un individu est modifié par apprentissage. Ce phénomène met en avant la malléabilité des représentations internes de l'auditeur lors de la perception de parole. L'adaptation à percevoir de nouveaux stimuli est observée ; ce paradigme diffère du précédent dans le sens où il se concentre sur les processus de plasticité du système perceptif, plutôt que sur l'objectif de gain de performance.

Ces deux paradigmes explicitent deux comportements distincts mais pourraient mettre en jeu des processus communs d'apprentissage. Goldstone (1998) a listé quatre types de processus sous-jacents à l'apprentissage perceptif, à savoir :

- « *Attention weighting* » - Pondération de l'attention. Parmi l'ensemble des caractéristiques de l'information perçue, certaines sont plus critiques que d'autres selon la tâche. La modulation de l'attention permet de hiérarchiser ces caractéristiques afin de mieux répondre à la tâche demandée ;
- « *Imprinting* » - Imprégnation. Ce terme est relié au mot « empreinte » qui est en fait un motif, une structure perçue de manière répétée, autour de laquelle les récepteurs captant l'information issue de l'environnement peuvent s'organiser ou se réorganiser ;
- « *Differentiation* » - Différentiation. Il s'agit de catégoriser deux stimuli comme différents alors qu'auparavant ils auraient été classés dans la même catégorie ;
- « *Unitization* » - Groupement. C'est principalement l'opposé de la différentiation : deux stimuli classifiés auparavant dans deux catégories différentes peuvent être regroupés dans une seule catégorie.

Un exemple de pondération de l'attention lors de l'apprentissage est mis en évidence dans l'expérience de Francis, Baldwin et Nusbaum (2000) dans laquelle les auteurs présentaient à des participants des stimuli contenant deux indices acoustiques cibles pertinents pour correctement détecter le lieu d'articulation du son entendu (contenus dans les transitions formantiques ou bien le *burst* lors de la production de consonnes /b/, /d/, /g/). Les consonnes artificiellement créées à partir de chacune des combinaisons de l'information acoustique issue des transitions formantiques et du *burst* pour chacune des trois consonnes. Ainsi, il y avait un stimulus dont les deux indices étaient en coopération (par exemple, une transition formantique d'un /b/ associée à un *burst* d'un /b/), alors que les deux autres stimuli contenaient des indices non cohérents (une transition formantique d'un /b/ associée à un *burst* de /d/ ou /g/). Les participants percevaient ensuite

ces stimuli dans un contexte CV (Consonne-Voyelle), et l'information phonologique relative au son qu'ils entendaient leur était donnée. Celle-ci ne pouvait cependant être en accord qu'avec un seul des deux indices acoustiques manipulés. Les résultats montrent que les participants ont alors concentré leur attention sur l'indice acoustique pertinent dans la tâche d'écoute, et ce même pour des stimuli qu'ils n'avaient pas entendu auparavant. Ce phénomène lié à l'apprentissage est un exemple de pondération de l'attention : l'auditeur privilégie l'information pertinente à la détection correcte du phonème parmi plusieurs indices.

D'autres données suggèrent que l'on s'habitue très rapidement à une parole prononcée avec un accent : environ 1 min d'exposition suffit pour arriver à traiter la parole accentuée aussi rapidement que des productions sans accent (Clarke & Garrett, 2004). Il est probable que, pour ce faire, nous ajustions nos connaissances phonologiques afin de mieux percevoir cet accent (Maye, Aslin, & Tanenhaus, 2008). Nous serions donc capables de modifier très rapidement notre système perceptif pour reconnaître des sons qui sont produits avec une prononciation atypique. Il est suggéré que, bien que la différence phonétique soit perceptible par l'auditeur, il adopterait une stratégie de double association phonétique (appliquée à la fois au son issu de son propre dialecte et au son accentué) comme des représentations d'un même phonème (Samuel & Larraza, 2015). Cela correspond à un mécanisme de groupement : des stimuli perçus comme appartenant à deux catégories distinctes sont finalement fusionnés en une seule.

Cette capacité d'association peut aussi être modulée lexicalement : on apprend même en regardant un film avec des sous-titres (Mitterer & McQueen, 2009). Dans cette expérience, des participants néerlandais avaient pour tâche de répéter des passages de parole produite dans une langue étrangère, l'anglais, avec deux accents possibles (anglais écossais ou australien). Au préalable, on leur présentait des extraits audiovisuels de cette langue étrangère, dans l'un des deux accents et dans trois conditions correspondant à trois groupes de participants : sans sous-titres, ou avec des sous-titres écrits soit dans la langue native des auditeurs (néerlandais) soit dans la langue étrangère considérée (anglais). Conformément aux expériences précédentes, les groupes de participants exposés à

la parole étrangère dans un accent donné ont mieux réussi à répéter les stimuli avec cet accent que les stimuli avec l'autre accent, sur lesquels ils n'avaient pas été entraînés. Mais le point important est que les participants qui ont perçus les extraits audiovisuels sous-titrés dans la langue cible (anglais) ont présenté de meilleures performances que ceux entraînés sans sous-titres, et ce même avec des mots qu'ils n'avaient pas entendus auparavant. Cependant, l'entraînement avec des sous-titres dans la langue d'origine (néerlandais) a au contraire diminué les performances. Les auteurs suggèrent que l'information phonologique extraite des sous-titres dans la langue cible renforcent le contenu phonétique cible perçu, mais qu'à l'inverse les sous-titres traduits gêneraient ou induiraient des erreurs lors de l'apprentissage, puisque le contenu phonétique de ces sous-titres (dans la langue maternelle) serait différent du contenu phonétique entendu (dans la langue étrangère). Cela expliquerait la différence obtenue entre les groupes ayant des langues de sous-titres différentes, notamment dans la tâche de répétition des nouveaux stimuli. Les auditeurs non-natifs pourraient en effet subir une compétition lors de la perception de parole, provoquée par l'activation de candidats lexicaux issus de leur langue native en écoutant une parole non-native (Broersma & Cutler, 2008). Ces changements perdurent dans le temps (Kraljic & Samuel, 2005) et affectent bien les connaissances au niveau phonétique puisque l'apprentissage peut se généraliser à des nouveaux stimuli (Francis et al., 2000 ; Kraljic, Samuel, & Brennan, 2008 ; Maye et al., 2008 ; Mitterer & Mcqueen, 2009).

Il apparaît donc que l'information lexicale intervient dans les processus d'ajustement des catégories phonétiques d'un auditeur en fonction de son expérience perceptive. Pour étudier la capacité et la rapidité d'un locuteur à s'adapter, il est possible d'utiliser cette information pour diriger ces ajustements (Clarke-Davidson, Luce, & Sawusch, 2008). Ainsi, Norris, McQueen et Cutler (2003) ont fait écouter à des participants 100 mots dont 20 étaient contrôlés dans le sens où ils comprenaient, en position finale, une fricative ambiguë entre /f/ et /s/. La moitié des participants ont perçu des stimuli ambigus qui étaient des mots existants seulement si le son final était interprété comme /f/, et l'autre moitié seulement si le son final était interprété comme /s/. Nous avons vu que

l'information lexicale induit une perception biaisée d'un son ambigu (Ganong, 1980), et, dans cette expérience, chaque participant a perçu le même son ambigu, mais n'a appris à l'associer qu'à un seul des deux phonèmes (/f/ ou /s/). La seconde partie de l'expérience consistait en une tâche de catégorisation d'un continuum /f/-/s/, et les participants ont alors catégorisé plus souvent les sons ambigus comme appartenant à la catégorie à laquelle ils avaient été entraînés. La catégorie phonétique spécifiée par l'information lexicale a inclus les sons ambigus, décalant en conséquence la frontière entre les catégories phonologiques /f/ et /s/ des auditeurs, et cet effet a perduré dans la tâche de catégorisation (voir Figure 9). Cet effet d'adaptation rapide est également possible sur la base de l'information visuelle (Bertelson, Vroomen, & De Gelder, 2003 ; Mitchel, Gerfen, & Weiss, 2016) si un son ambigu est apparié avec un visage articulant le son correspondant à l'une des catégories impliquées. Dans ce cas, les participants sont influencés par l'information visuelle pour modifier les frontières de leurs catégories phonologiques. Ces changements nécessitent très peu de stimuli et suggèrent qu'en tant qu'auditeur, nous sommes perpétuellement et rapidement en train de nous habituer à la parole du locuteur auquel nous sommes confrontés.

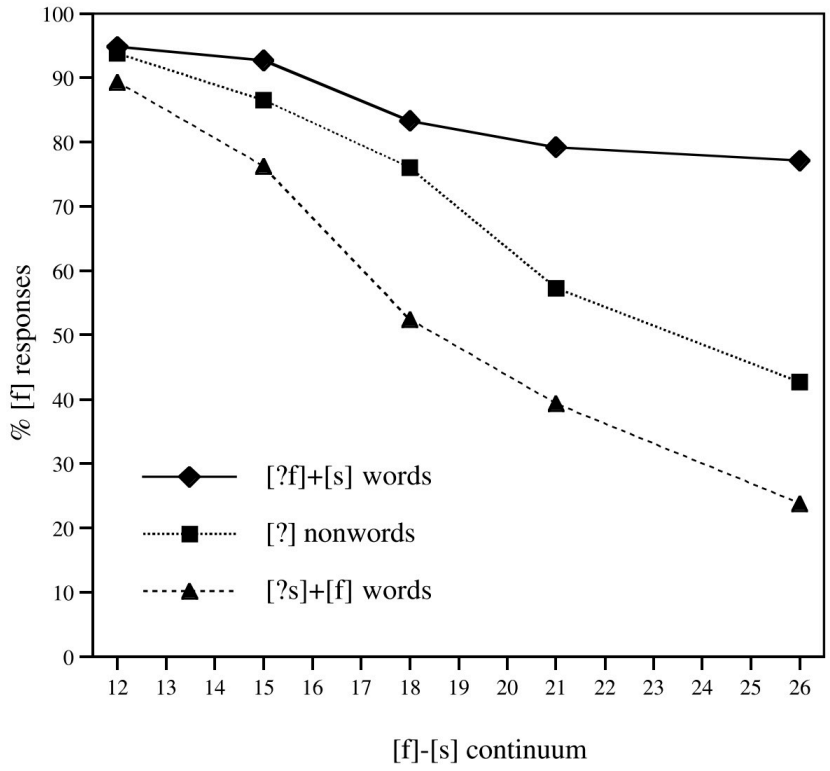


Figure 9 : Extrait de Norris et al. (2003). Pourcentage de réponses /f/ (*% [f]* responses) en fonction des conditions d'apprentissage. La première condition (*[?f] + [s] words,* trait uni et losanges) correspond au groupe ayant appris à associer les sons ambigus au phonème /f/, à l'inverse de ceux qui l'ont associé au phonème /s/ (troisième condition : *[?s] + [f]*, trait en tirets). Dans la seconde condition (*[?] nonwords*, trait en pointillés), les participants ont perçu les sons ambigus mais apprennent sur des non-mots et ne bénéficient ainsi d'aucune information lexicale.

La perception et l'apprentissage perceptif seraient spécifiques au locuteur, et l'auditeur s'adapterait aux idiosyncrasies de chacun (Nygaard & Pisoni, 1998 ; Nygaard, Sommers, & Pisoni, 1994 ; Souza, Gehani, Wright, & McCloy, 2013). La familiarité avec le locuteur permet de mieux le reconnaître et également de mieux le comprendre. Cet apprentissage est relatif au locuteur et ne se transfère pas aux stimuli produits par un nouveau locuteur (Eisner & McQueen, 2005). Pourtant, l'auditeur n'a pas besoin de reconnaître la voix du locuteur pour bénéficier de l'apprentissage perceptif (Holmes, Domingo, & Johnsrude, 2018) : cela suggère que la capacité à s'adapter à la parole d'un individu est, du moins en partie, indépendante de la capacité à attribuer une identité au locuteur.

Pourtant, une généralisation suite à l'apprentissage est observée pour la familiarisation aux dialectes (Kraljic, Brennan, et al., 2008 ; Reinisch & Holt, 2014). Les bilingues, par exemple, peuvent sélectionner et utiliser leurs connaissances

propres à chacune de leurs langues afin d'ajuster leurs systèmes perceptifs selon le contexte linguistique (Gonzales, Byers-Heinlein, & Lotto, 2019). L'efficacité de l'apprentissage pourrait dépendre de la variabilité acoustique des stimuli perçus (Wade, Jongman & Sereno 2007), et cette variabilité permettrait à l'auditeur d'inférer les ajustements acoustiques-phonétiques nécessaires pour mieux percevoir. C'est par exemple le cas lors de l'apprentissage d'une parole avec accent produite par plusieurs locuteurs plutôt qu'un seul (Bradlow & Bent, 2008) : l'auditeur montre un bénéfice à la compréhension du locuteur qu'il a appris, mais apprendre à partir de la parole de plusieurs locuteurs permet de mieux percevoir ensuite un locuteur jamais rencontré auparavant. L'information recueillie chez l'ensemble des locuteurs serait également présente dans la parole du nouveau locuteur et pourrait représenter un invariant dans leur dialecte commun. Ceci démontre l'existence d'un processus de généralisation par une approche d'apprentissage multi-locuteurs.

Certaines études ont tenté d'explorer l'existence de possibles phénomènes de blocage de l'apprentissage. En suivant le protocole de l'expérience d'apprentissage par un biais lexical (Norris et al., 2003) décrite plus tôt, une étude (Zhang & Samuel, 2014) a pu montrer que l'apprentissage perceptif de la parole d'un locuteur était perturbé lorsque la parole était bruitée, mais à l'inverse qu'il restait efficace lorsque le participant avait une charge cognitive plus lourde (avec une tâche secondaire pendant l'apprentissage : la mémorisation d'une séquence de nombres, ou la recherche visuelle d'une lettre). La capacité d'apprentissage reposerait davantage sur la qualité du signal acoustique et la capacité à retrouver les informations critiques à l'apprentissage. Dans ce même paradigme, le bénéfice de l'apprentissage disparaît si un facteur indépendant du locuteur peut expliquer l'ambiguïté de sa prononciation : par exemple, si les stimuli critiques pour l'apprentissage étaient accompagnés d'une vidéo sur laquelle les participants voyaient un stylo dans la bouche du locuteur, alors l'effet d'apprentissage s'annulait (Kraljic, Brennan, et al., 2008 ; Kraljic & Samuel, 2011). Des processus jugeant de la « pertinence » de certains stimuli empêcheraient alors

l'apprentissage des caractéristiques du signal si elles sont apparemment induites par une cause spécifique et non transférable hors de la condition d'apprentissage.

En résumé, l'apprentissage perceptif permet de mettre à jour les connaissances phonétiques et phonologiques associées aux idiosyncrasies de production, et ce pour le bien de la communication. Les mécanismes d'apprentissage utilisent toute information disponible (*e.g.* visuelle, lexicale) pour réaliser ces modifications. Les représentations internes de l'auditeur sont rapidement mises à jour, même face à une exposition limitée lors de l'apprentissage. Comme nous allons le voir ensuite, l'apprentissage perceptif induit des changements qui ne se limitent pas au système perceptif de l'auditeur.

2.3.1 Le lien entre perception et production dans les processus d'adaptation

Le terme d'« apprentissage » réfère non seulement à l'évolution des frontières catégorielles entre phonèmes, mais aussi à l'acquisition de nouvelles connaissances phonologiques, par exemple lorsqu'un auditeur est confronté à une langue qui ne possède pas les mêmes catégories phonologiques que sa langue maternelle (Jamieson & Morosan, 1986 ; Logan, Lively, & Pisoni, 1991). Les bénéfices de l'apprentissage perceptif demeurent plusieurs mois après l'entraînement et se généralisent à d'autres mots, mais restent limités au locuteur sur lequel a été effectué l'entraînement (Logan, Lively, & Pisoni, 1991). Le lien entre la capacité à percevoir et à produire a aussi été exploré lors de l'apprentissage. Bradlow, Pisoni, Akahane-Yamada, et Tohkura (1997), ont montré que des participants qui parvenaient à mieux discriminer deux catégories phonologiques étaient également capables de les produire plus précisément. Ce bénéfice à la fois en production et en perception peut aussi être obtenu avec des stimuli audiovisuels et ce de manière plus efficace qu'avec des stimuli présentés auditivement (Hazan, Sennema, Iba, & Faulkner, 2005 ; Inceoglu, 2016).

Les changements en perception n'entraîneraient pas systématiquement des adaptations en production (Kraljic, Brennan, et al., 2008). Ainsi, lors d'une tâche d'imitation d'un continuum phonétique, des participants natifs de deux pays ont

réussi à reproduire finement les variations des sons qu'ils percevaient lorsqu'ils appartenaient à leurs catégories phonologiques natives respectives mais pas les variations de sons n'appartenant pas à leur langue (Olmstead et al., 2013). Si chacun est limité par son répertoire phonétique en perception, il en va de même en production.

D'autre part, si nous ajustons notre perception à notre interlocuteur, nous lui ajustons en fait aussi notre production. Ainsi, lorsqu'un participant est exposé à la parole d'un interlocuteur, que ce soit en situation conversationnelle ou en simple écoute, ses productions présentent parfois un phénomène de convergence, qui se traduit par le fait qu'elles subissent des modifications qui la rapprochent des productions de son interlocuteur (Gambi & Pickering, 2013 ; Pickering & Garrod, 2004), et ce à plusieurs niveaux (*e.g.* vocabulaire, pauses, intensité de la parole, etc.). Il est suggéré que s'aligner sur les représentations de son interlocuteur faciliterait les échanges : le dialogue entre deux personnes serait donc conçu comme une activité coopérative dans laquelle les deux acteurs-interlocuteurs devraient conjointement s'adapter. Au sein de ce phénomène de convergence, on peut se focaliser, dans le contexte de cette thèse, sur la convergence phonétique entre deux interlocuteurs (Pardo, 2013). La convergence phonétique indique bien un lien entre perception et production mais pas nécessairement un effet miroir (Pardo, 2012 ; et voir plus loin) : ce mécanisme serait guidé à la fois par des processus automatiques et par des processus conscients, notamment socio-psychologiques (Giles, Ogay, & Gallois, 2006). Ainsi, la convergence se ferait à l'intérieur de son propre espace phonologique natif et serait modulée par certains facteurs, comme la présence d'une image représentant le locuteur ainsi que le degré d'attraction envers ce locuteur d'après le jugement de l'auditeur (Babel, 2012).

Une autre manière d'observer une adaptation en ligne d'un locuteur est de perturber sa propre parole. En effet, une seule personne peut être à la fois l'auditeur et le locuteur des schémas d'adaptation précédents. C'est par exemple le cas des expériences de « *lip-tube* » dans lesquelles des locuteurs subissent une perturbation les empêchant d'arrondir leurs lèvres dans la production du son /u/.

Or, cette voyelle nécessite, en situation normale en français, d'arrondir les lèvres pour atteindre des valeurs de formants adéquates. Certains participants mettent alors en place une stratégie de recul de la langue leur permettant de compenser en partie cette perturbation et finalement de produire un /u/ relativement acceptable (Savariaux, Perrier, & Orliaguet, 1995). Cette expérience met en avant la capacité d'un locuteur à trouver une stratégie adaptée afin que la cible acoustique qu'il/elle essaie de produire corresponde à sa connaissance du son et à la perception de sa propre production.

Un autre type de perturbation, cette fois-ci par manipulation acoustique, a mis en avant un mécanisme d'adaptation à sa propre parole (Lametti, Krol, Shiller, & Ostry, 2014 ; Shiller, Sato, Gracco, & Baum, 2009). Ces expériences ont manipulé en temps réel le retour auditif qu'avaient les participants de leurs propres productions. Ainsi, les participants prononçaient un mot contenant un phonème (par exemple le son /ɛ/ dans le mot *head* - tête en français), mais entendaient un phonème immédiatement acoustiquement altéré pour ressembler à un phonème différent (le phonème /a/ dans *had* – eu en français, dérivé du phonème /ɛ/ de *head* par un décalage du premier formant). Suite à la perturbation de leurs voyelles produites, les participants ont compensé et adapté leurs productions, et leur perception du contraste qui avait été manipulé en a également été modifiée. Ces résultats confortent l'idée d'un lien entre production et perception de la parole. D'une part, l'altération du retour perceptif a modifié directement la production des sons, afin d'être conforme aux connaissances phonologiques du locuteur. D'autre part, la mise à jour de ces connaissances motrices a rejailli sur la perception, en retour. L'effet de l'apprentissage se dissipe peu à peu pour revenir à l'état initial de la production et de la perception ; au-delà de la plasticité, il existe une stabilité des connaissances phonologiques du locuteur (Lametti, Rochet-Capellan, Neufeld, Shiller, & Ostry, 2014).

Désormais, nous portons notre attention sur le rôle potentiel du système moteur dans la perception de la parole qui constitue une problématique centrale dans la recherche sur la perception de la parole et qui pourrait expliciter le lien entre production et perception que nous venons de mettre en avant.

2.4 Le rôle du système moteur lors de la perception de parole

Dans le cadre de ce manuscrit, la parole a été présentée comme un « objet » pouvant être perçu au travers de différentes modalités, et s'adapter, par apprentissage, aux conditions d'interaction. L'objectif de cette partie est d'abord de prendre du recul dans le but de mieux interpréter les fondements sous-jacents au fonctionnement des mécanismes présentés précédemment. Les théories que nous allons aborder maintenant s'appuient sur un large ensemble de données expérimentales et tentent de répondre à des problèmes posés par la perception de parole. Ces approches très différentes, comme on le verra, abordent toutes la nature des processus de perception et le fait qu'un stimulus est traduit vers un percept (pour une revue, voir Samuel, 2011). Après avoir exposé les théories de la perception de la parole que nous avons jugées pertinentes pour contextualiser notre travail, nous aborderons des travaux empiriques interrogeant l'implication du système moteur dans la perception de la parole et l'apport potentiel de son implication dans la familiarisation au locuteur.

2.4.1 Théories de la perception de la parole

Les théories de la perception de la parole sont plurielles et varient notamment dans le rôle qu'elles accordent aux systèmes auditif et moteur.

On peut d'abord mentionner les théories auditives de la parole (Diehl & Kluender, 1989 ; Diehl, Lotto, & Holt, 2004 ; Stevens & Blumstein, 1978). Celles-ci défendent l'idée que l'information nécessaire pour accéder à un percept réside entièrement dans le signal acoustique et ne nécessite pas de connaissances sur les mécanismes (essentiellement articulatoires ou moteurs) qui ont conduit à la formation de ce signal. Afin de pouvoir différencier un percept d'un autre et également de percevoir un même son dans différents contextes, le signal acoustique incorporerait des invariants acoustiques préparant l'émergence des percepts. Cette hypothèse est au cœur d'une des variantes des théories auditives de la perception de la parole : la « théorie quantique » (Blumstein & Stevens, 1979 ; Stevens & Blumstein, 1978). Les variations des paramètres articulatoires ne reflètent pas celles de la perception acoustique. Un changement subtil au niveau

articulatoire peut faire varier fortement le résultat acoustique et un tel changement peut constituer la frontière entre deux catégories phonologiques. Inversement, les sons intra-catégoriels aboutissent à une perception stable malgré des changements articulatoires qui peuvent être conséquents parce que malgré ces changements les paramètres acoustiques changent peu (voir Figure 10).

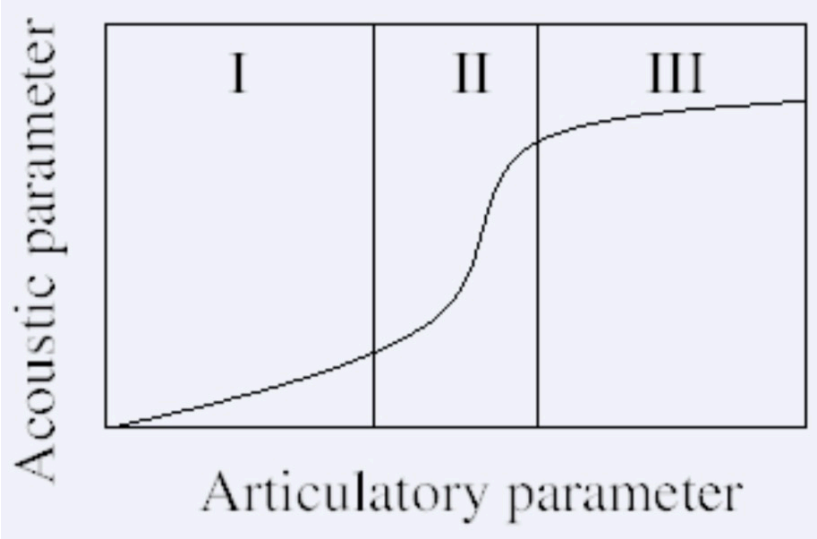


Figure 10 : Extrait de Stevens (1989). Variations d'un paramètre acoustique (en ordonnée) en fonction d'un paramètre articulatoire (abscisses). On distingue trois zones : les zones I et III sont stables sur le plan acoustique malgré une variation importante sur le plan articulatoire : ce sont des régions correspondant potentiellement à des phonèmes. À l'inverse, la zone II représente une frontière entre deux catégories, puisqu'un changement acoustique grand et rapide est observé pour un faible changement articulatoire.

À l'inverse, les théories motrices de la parole soutiennent l'hypothèse que la perception de parole se baserait sur la perception des gestes articulatoires planifiés à l'origine de cette parole (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967 ; Liberman & Mattingly, 1985). Ainsi, le signal acoustique (et autres signaux sensoriels, notamment visuels) sont conçus comme le résultat de la réalisation d'évènements moteurs planifiés. L'auditeur, lui-même capable de produire de la parole, retrouverait lors de la perception les plans moteurs du locuteur. Ainsi, il existerait un lien entre perception et production de la parole. Chaque humain aurait ses propres représentations motrices internes, qu'il recruterait lorsqu'il perçoit les réalisations issues des représentations internes d'un autre individu. L'hypothèse d'un lien entre perception et production a été confortée par la découverte des « neurones miroirs » (Rizzolatti & Craighero, 2004

; Rizzolatti, Fogassi, & Gallese, 2001). D'abord mis en évidence chez les macaques, les neurones miroirs sont un type de neurones qui s'activent à la fois lorsque le macaque produit une certaine action et lorsqu'il observe quelqu'un (macaque ou humain) produire cette même action. Chez l'humain, il a pu être mis en évidence des activations des mêmes zones motrices à la fois lors de la perception et de la production de sons nécessitant les mêmes articulateurs (Pulvermüller et al., 2006), et notamment pour la perception et la production de la parole (voir une revue dans Skipper, Devlin, & Lametti, 2017). Une théorie adjacente aux théories motrices appelée « Perception Réaliste Directe » (Galantucci, Fowler, & Turvey, 2006) propose que l'auditeur percevrait directement les mouvements effectués plutôt que de référer à des représentations internes motrices. Selon cette théorie, la perception de la parole ne serait pas spécifique : comme les autres actes de perception, elle renverrait au monde physique produisant les événements perceptifs – ici, le conduit vocal et les gestes articulatoires.

Un troisième type de théories de la perception de la parole, dites sensori-motrices, proposent de faire converger les idées précédentes vers une conception de la perception de parole comme dépendant à la fois des traitements perceptifs issus de la perception d'un signal acoustique de parole et des connaissances motrices du locuteur. Par exemple, la Théorie de la Perception pour le Contrôle de l'Action (Perception-for-Action-Control Theory, PACT – Schwartz, Basirat, Ménard, & Sato, 2012) présente un cadre théorique de la perception de la parole combinant les représentations motrices et les informations sensorielles. L'intégration des informations et connaissances sensorielles et motrices conditionnerait les unités distinctives de la parole.

L'existence de ces différentes théories amène à discuter l'implication du système moteur dans les deux mécanismes abordés précédemment : la perception multisensorielle et l'apprentissage perceptif.

2.4.2 Résultats empiriques suggérant l'implication du système moteur dans la perception, notamment de parole

Dans la littérature, un nombre conséquent d'études mettent en évidence l'implication du système moteur dans la perception de mouvements du corps humain (Wilson & Knoblich, 2005). Une personne percevant une action active ses propres représentations motrices pour réaliser cette action. Pourtant, cette planification de l'action semble être sans résultat effectif puisque la perception de l'action est passive : aucune action n'est ouvertement effectuée. Pour justifier ces activations motrices sans acte moteur, il est suggéré d'abord qu'elles augmenteraient l'attention vers les mouvements. Notons en effet qu'il semblerait que notre attention soit plus rapidement attirée vers des silhouettes de corps humain (Downing, Bray, Rogers, & Childs, 2004) ou des visages (Ro, Russell, & Lavie, 2001) que vers des objets et des formes de main. D'autre part, ces activations pourraient être le reflet d'une résonance perceptive sur l'action perçue, que l'on pourrait considérer comme le produit d'une imitation inconsciente.

Les processus impliqués lors de la perception et la planification d'action partageraient un espace de représentation commun (Prinz, 1997). L'imitation inconsciente des actions d'un autre humain permettrait une meilleure interprétation de celles-ci : l'imitation pourrait avoir un rôle prédictif qui pourrait en retour ajuster le système perceptif en jeu. Les simulations mentales pourraient ainsi prédire un résultat précis à l'issue de l'action en cours, ou bien simuler le résultat d'une nouvelle action (Wilson & Knoblich, 2005), permettant ainsi de mieux comprendre les intentions de l'individu qui l'a produite (Blakemore & Decety, 2001). De cet fait, la personne dont nous serions le plus à même de comprendre les actions serait nous-même. Nous sommes à la fois capables de mieux identifier nos propres actions (Repp & Knoblich, 2004) mais également de mieux prédire le résultat de ces actions (Knoblich & Flach, 2001). En conséquence, les capacités perceptives pourraient bénéficier de l'expertise en production de certains mouvements spécifiques, par exemple : dans le sport (Aglioti, Cesari, Romani, & Urgesi, 2008) ou la musique (Sherwin & Sajda, 2013). Notre expérience

permet à la fois de mieux prédire le résultat d'une action observée grâce à des simulations internes plus fines, ou encore apprendre à porter son attention sur les mouvements critiques afin de récupérer l'information la plus pertinente à fournir aux processus de traitement (Abernethy, Zawi, & Jackson, 2008). D'une certaine manière, on peut considérer un adulte comme un expert en perception de parole. Mais pour chaque locuteur que l'on rencontre, nous sommes confrontés à une parole spécifique à laquelle notre système perceptif s'ajuste. Ainsi, notre propre parole est celle dont nous sommes le plus expert, et celle qu'on reconnait le mieux, par exemple dans une tâche de perception de parole visuelle (Tye-Murray, Spehar, Myerson, Hale, & Sommers, 2013).

Une manière de détecter la présence d'une activité du système moteur lors de la perception passive de parole est d'avoir recours à l'imagerie cérébrale. De nombreuses études proposent ainsi que certaines zones cérébrales activées lors de la production le seraient également lors de la perception de la parole. L'investissement du système moteur est particulièrement important en situation de perception audiovisuelle de la parole (pour une revue, voir Skipper et al., 2017).

Un ensemble de paradigmes expérimentaux ont également permis de mettre en avant un rôle fonctionnel du système moteur lors de la perception de parole chez l'adulte. Parmi ceux-ci, on peut citer : la perturbation de l'état du système moteur lors de la perception de parole en modalité auditive seule (D'Ausilio, Bufalari, Salmas, & Fadiga, 2012 ; D'Ausilio et al., 2009) et audiovisuelle (Sato, Buccino, Gentilucci, & Cattaneo, 2010), la fatigue motrice due à la répétition d'un geste moteur spécifique (Sato et al., 2011), l'adaptation du système perceptif à une altération somatosensorielle (Ito, Tiede, & Ostry, 2009 ; Nasir & Ostry, 2009) ou bien du retour auditif (Houde & Jordan, 1998 ; Shiller et al., 2009) lors de la production de parole, et une rapidité de traitement renforcée grâce à l'emploi du lien perceptuo-moteur en modalités auditive (Fowler, 2003) et audiovisuelle (Scarbel, Beautemps, Schwartz, & Sato, 2014).

Au-delà d'un avantage en termes de vitesse de traitement, le système moteur influe sur la perception, et ce dès le plus jeune âge. Dans une expérience qui a eu un retentissement important (Bruderer, Danielson, Kandhadai, & Werker,

2015), des bébés de 6 mois avaient un jouet à mettre dans la bouche qui contraignait leur langue à avoir une position spécifique. Dans un paradigme de préférence visuelle, ils écoutaient un contraste non-natif. Ce contraste nécessitait un changement de la position de la langue lors de son articulation. Le fait qu'il s'agissait d'un contraste non-natif permet de rejeter l'influence de toute connaissance préalable des bébés sur ce qui pourrait différencier les deux sons. Les résultats montrent que si la langue du bébé est libre, il parvient à distinguer les deux sons dans un paradigme classique de temps de regard. Dans un premier temps, les auteurs ont donc montré que les bébés étaient capables de percevoir le contraste testé. Cependant, avec un jouet en bouche, cette capacité de distinction des deux sons disparaît mais uniquement si le jouet contraint les mouvements de la langue d'une manière qui empêche la production du son perçu. Ces résultats suggèrent que, sans connaissances linguistiques particulières, les bébés utilisent l'information sensorimotrice pour distinguer deux sons. La capacité à construire des catégories perceptives serait liée aux capacités en production de l'enfant (Vilain, Dole, Lœvenbruck, Pascalis & Schwartz, 2013). Une autre étude (Yeung & Werker, 2013) a confirmé le rôle de mécanismes articulatoires dans la perception multisensorielle chez le bébé dès 4,5 mois : dans un paradigme d'attention audiovisuelle, l'attention des bébés se détournait du visage qui correspondait aux sons qu'ils percevaient lorsqu'ils avaient la même posture de lèvres nécessaire pour produire ces sons. Ces deux études mettent en évidence un lien entre production et perception de la parole dès les premières phases du développement, même si ce lien pourrait être plus complexe qu'une simple relation bidirectionnelle d'équivalence (Pardo, 2012 ; Schütz-Bosbach & Prinz, 2007). De fait, les capacités en perception ne reflètent pas celles en production : ceci est évident chez le bébé, mais c'est aussi le cas chez l'adulte : on peut prendre l'exemple des contrastes non natifs (Bradlow et al., 1997). Ainsi, il apparaît qu'au plus tôt dans le développement, le système moteur joue un rôle dans la perception de parole, et ce particulièrement en situation d'intermodalité. Même dans une tâche de perception visuelle pure (Turner, McIntosh, & Moody, 2015), des enfants âgés de 7 ans en moyenne dont les mouvements faciaux étaient restreints ont été moins

performants dans une tâche de lecture labiale que d'autres enfants libres dans leurs mouvements faciaux.

Le système moteur serait donc requis lors de la perception de parole. Les connaissances motrices propres à l'auditeur, confrontées à celles du locuteur perçu, lui permettent d'ajuster son système perceptif afin de mieux le percevoir. Le rôle du système moteur est particulièrement important dans le contexte de la multisensorialité (notamment de la parole audiovisuelle). Dans le cas où le locuteur serait également l'auditeur, alors cet ajustement serait maximal, et les performances associées meilleures que pour toute autre parole. À l'inverse, une perturbation du système moteur gênerait un tel effet. Si le système moteur est impliqué dans la perception de la parole alors, on peut supposer que son recrutement pourrait jouer un rôle dans l'apprentissage perceptif.

2.4.3 Recruter le système moteur dans l'apprentissage perceptif

Le lien entre production et perception de parole lors de l'apprentissage perceptif a déjà été mentionné (voir Chapitre 2.3.1). Comme nous venons de le voir, l'auditeur semble s'adapter à la parole qu'il perçoit grâce à un certain nombre de processus imitatifs, le plus souvent implicites. Les ajustements qui en résultent serviraient, en fin de compte, à mieux percevoir la parole. Afin de spécifiquement tester ce lien lors de l'apprentissage, Adank, Hagoort et Bekkering, (2010) ont évalué si l'imitation explicite de la parole perçue permettait à des participants de mieux la comprendre dans un second temps. L'apprentissage perceptif s'est fait sur un accent néerlandais produit par une locutrice native, créé artificiellement grâce à des règles phonologiques inventées construites spécifiquement pour l'expérience. Pour estimer les performances des participants – eux aussi néerlandais natifs – deux tests ont été réalisés. Dans chacun de ces tests, les participants avaient pour tâche de répéter cent phrases présentées dans le bruit. Chaque phrase comportait quatre mots cibles qui permettaient d'évaluer quantitativement les performances. Si les participants étaient capables de répéter plus de la moitié des mots cibles, le niveau de bruit de la phrase suivante était augmenté, rendant la tâche plus facile. Inversement, si les participants répétaient correctement moins de la moitié des

mots cibles, alors le niveau de bruit était diminué, rendant la tâche plus difficile. Le niveau de bruit ne changeait pas d'une phrase à la suivante s'ils répétaient exactement deux mots corrects sur quatre. Ainsi, le niveau de bruit au cours de l'expérience s'adaptait à la performance du participant La performance au cours d'un test était mesurée grâce au niveau de bruit évoluant au cours de l'expérience (Plomp & Mimpen, 1979) : il s'agit de la moyenne du niveau de bruit sur l'ensemble des phrases présentées (« Speech Reception Threshold », SRT, d'autant plus bas que la performance perceptive est grande). Entre les deux tests, les participants étaient assignés à différents groupes d'entraînement perceptif. Le groupe « *Baseline* » (Contrôle), ne bénéficiait pas d'entraînement mais participait simplement aux deux tests, afin de vérifier qu'il n'y ait pas d'apprentissage spontané lors de la passation des tests. Les autres groupes perçevaient cent autres phrases présentées clairement (sans bruit). Les groupes avaient chacun une tâche spécifique, à savoir : le groupe « *Listening* » (Écoute) devait simplement écouter les stimuli d'entraînement, le groupe « *Repeating* » (Répétition) devait répéter – sans reproduire l'accent – les stimuli, le groupe « *Transcription* » (Transcription) retranscrivait orthographiquement les phrases perçues, le groupe « *Imitation* » (Imitation) devait imiter les phrases présentées, et enfin les participants du groupe « *Imitation + Noise* » (Imitation dans le bruit) imitaient les phrases présentées sans être capables d'entendre leurs propres imitations (casque diffusant du bruit dans les oreilles). Les résultats sont présentés sur la Figure 11 qui illustre les différences entre les performances, mesurées par le SRT, du premier test (avant l'entraînement) et celle du second (après l'entraînement) pour les différents groupes. Ainsi, une valeur positive (SRT plus élevé dans le premier test) indique une meilleure performance dans le second test par rapport au premier. On peut noter un effet d'apprentissage passif lors de l'entraînement dans la condition Contrôle. Une autre étude (Kreitewolf, Mathias, & von Kriegstein, 2017) rapporte un tel effet, qui pourrait pointer sur un mécanisme d'apprentissage implicite et systématique lors de la perception de parole. Cependant, ce mécanisme (s'il n'en est qu'un) est mal compris, notamment dans les interactions quotidiennes (Case, Seyfarth, & Levi, 2018).

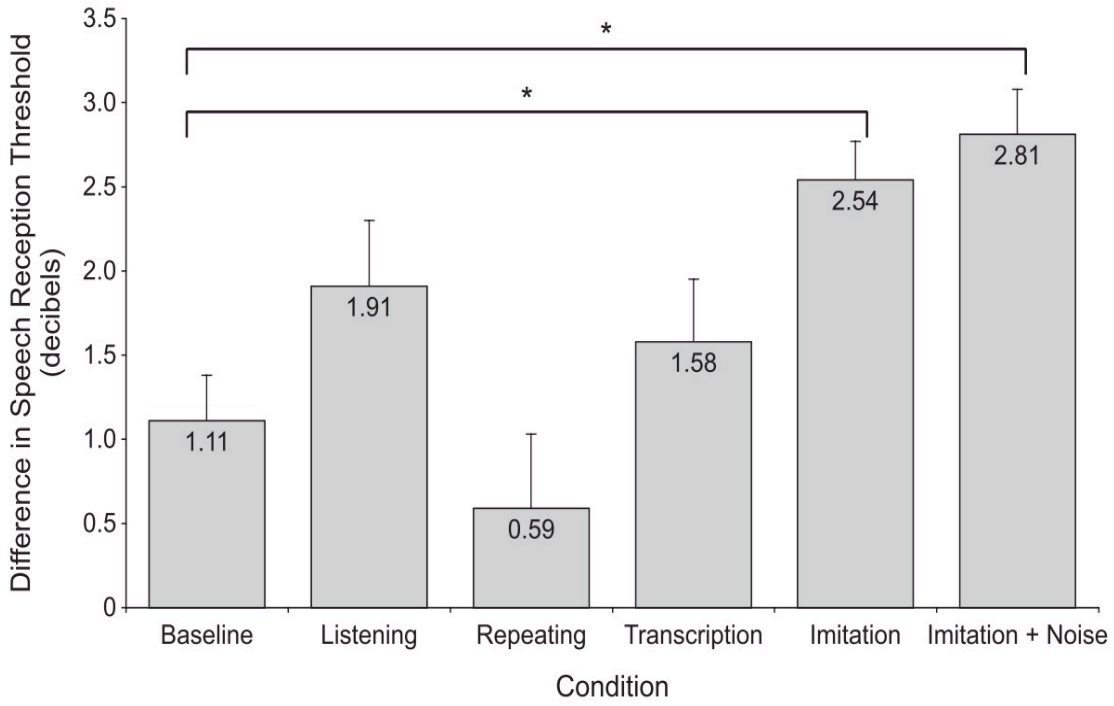


Figure 11 : Extrait de Adank et al. (2010). Différence de SRT (Speech Reception Threshold) entre le premier et le second test pour chacun des groupes de participants [*Baseline* = Contrôle, *Listening* = Écoute, *Repeating* = Répétition, *Transcription* = Transcription, *Imitation (+ Noise)* = Imitation (avec bruit)]

Les performances des groupes Écoute, Transcription et Répétition ne sont pas significativement différentes de celles du groupe Contrôle. On observe même une tendance de performances légèrement plus faibles pour le groupe « Répétition » : il est suggéré que la production gênerait lors de l'apprentissage perceptif, comme ce serait le cas pour apprendre des mots ou sons dans une langue étrangère (Baese-Berk & Samuel, 2016). On pourrait également suggérer que les représentations motrices impliquées lors de la répétition sont celles de l'auditeur et non du locuteur, et qu'elles pourraient entrer en conflit dans la tâche perceptive, contrairement à celles impliquées dans l'imitation qui se rapprocheraient de celles du locuteur. En effet, les seuls groupes ayant une performance significativement différente du groupe Contrôle sont ceux ayant eu recours à l'imitation pendant l'apprentissage. Le gain en intelligibilité observé est maintenu même si le participant n'est pas en mesure d'entendre ses propres imitations (condition Imitation avec bruit). L'effet de l'imitation ne se borne donc pas à doubler l'entrée auditive créée lors de la production de parole par l'auditeur. L'appel explicite aux

processus imitatifs mis en jeu lors de la perception de parole permettrait à l'auditeur de mieux ajuster son système perceptif à la parole qu'il perçoit.

Le système moteur est évidemment impliqué dans la production de parole mais également lors de la perception de parole. Conséquemment, il joue un rôle dans la perception multisensorielle et l'apprentissage perceptif, auxquels nous nous sommes intéressés plus tôt. Ces deux processus cognitifs perceptifs seraient modulés par la configuration du système de production de parole de l'auditeur.

2.5 Percevoir une parole différente : le cas de la parole des personnes avec trisomie 21

2.5.1 Qu'est-ce que la trisomie 21 ?

La trisomie 21 (T21) est une anomalie génétique provoquée par la présence d'un chromosome supplémentaire dans la 21ème paire du génome de son hôte. Le taux d'apparition de la T21 a été estimé à 1/370 grossesses, alors qu'on dénombre un cas de T21 chez 1/1500 nouveau-nés en France en 2011 (Haute Autorité de Santé, 2015). Aux États-Unis, ce même chiffre est estimé à environ 1/700 nouveau-nés (Parker et al., 2010). Le phénotype associé à la T21 comporte un certain nombre de spécificités morphologiques qui retardent l'acquisition de certaines compétences motrices permettant l'autonomie (marcher, manger, parler, écrire, lire, etc.) et qui contribuent à stigmatiser les personnes (de Graaf, Levine, Goldstein, & Skotko, 2019). Si les individus avec T21 partagent des spécificités phénotypiques, celles-ci s'expriment de manières très différentes d'un individu à l'autre (Karmiloff-Smith et al., 2016). L'incidence de la T21 a été reliée à l'âge de la mère à la naissance de l'enfant, avec une augmentation des risques avec l'augmentation de l'âge (1/500 à 20 ans allant jusque 1/100 à 40 ans, Haute Autorité de Santé, 2015). En revanche, elle est indépendante de facteurs ethniques, sociaux-économiques ou géographiques (Arumugam et al., 2015 ; Haute Autorité de Santé, 2015). Malgré l'incidence élevée de la T21 sur le fœtus, le nombre de naissances d'enfants avec T21 a diminué dans les sociétés ayant mis en place un dépistage intra-utero, plus ou moins précoce, la T21 rentrant dans le spectre de

l'avortement thérapeutique (voir Figure 12). Le nombre de dépistages positifs prénataux a augmenté avec le temps, notamment du fait de l'augmentation de l'âge maternel moyen sur la période 1990-2009 (Loane et al., 2013, voir aussi Figure 12).

Jusqu'à récemment, le dépistage prénatal consistait en un test combiné à la fin du premier trimestre de grossesse évaluant un facteur de risque sur la base de paramètres échographiques et sanguins associés à l'âge de la mère (pour une description détaillée, voir Haute Autorité de Santé (HAS), 2015, p. 17). En théorie, cette méthode de dépistage permet de détecter 85 à 90% des fœtus avec T21 avec un taux de faux positifs de 5% (Haute Autorité de Santé, 2015). À partir d'un certain niveau de risque évalué (1/250), l'évaluation du caryotype était proposé, *via* amniocentèse ou choriocentèse, tests invasifs induisant un risque de fausse couche (0,5% à 1% des cas ; Haute Autorité de Santé, 2015). Plus récemment, sont apparus des tests de détection par analyse de l'ADN fœtal libéré dans la circulation sanguine maternelle. L'introduction de ces tests fiables à ~99% a été validée par le rapport de la HAS (2015) : depuis 2016, ce test est proposé et remboursé en cas de test combiné positif (risque > 1/250), mais il est accessible à tous pour ~600 euros. L'introduction de ce test va probablement contribuer à diminuer encore plus le nombre de naissances de bébés avec T21.

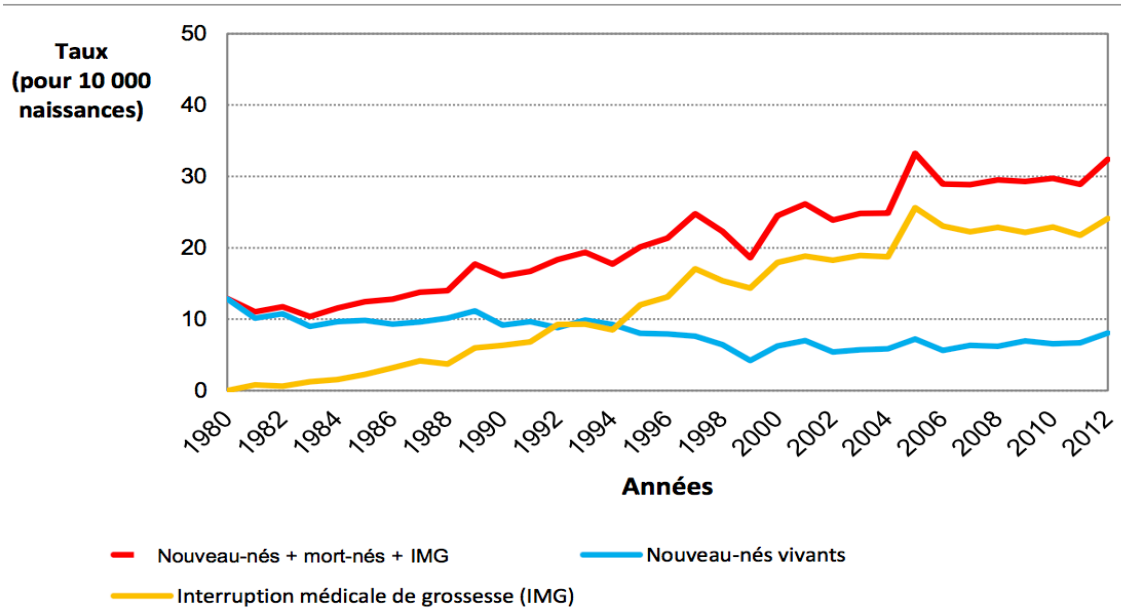


Figure 12 : Évolution du taux d'interruptions médicales de grossesses concernant des fœtus avec T21 et du nombre de naissances d'enfants avec T21 (pour 10 000) de 1980 à 2012 en France (extrait de Haute Autorité de Santé, 2015, p. 14).

La décision de l'interruption de grossesse pourrait en partie être motivée par une mauvaise connaissance de ce qu'est la T21 ou une évaluation biaisée de la difficulté d'élever un enfant avec une déficience (Rubeis & Steger, 2019), en lien avec un protocole d'annonce purement médicale du diagnostic potentiellement traumatisant (Skotko, Capone, & Kishnani, 2009 ; Skotko, Kishnani, & Capone, 2009). Les parents concernés estiment d'ailleurs que les difficultés d'élever leur enfant avec T21 viendraient avant tout du manque de support social et d'infrastructures adaptées pour répondre aux besoins spécifiques de leur enfant, plutôt que de ses besoins spécifiques. Ces éléments pointent vers une information déséquilibrée qui déforme la conception de la part du public sur ce qu'est réellement la T21 et la part qu'elle peut prendre dans la qualité de vie d'un individu.

La T21 est sans doute la source de déficience cognitive génétiquement identifiée la plus connue au monde (Katz & Lazcano-Ponce, 2008) . Cependant, bien que les traits d'une personne avec T21 soient très reconnaissables, le public connait très peu en quoi consiste réellement les difficultés que rencontrent ces personnes et sous-estiment le rôle que peut jouer l'environnement. Les capacités

cognitives des personnes avec T21 sont très variables, allant d'une déficience légère à profonde. Les personnes avec T21 peuvent également souffrir de problèmes de santé à prendre en compte lors du suivi de ces individus (Määttä, Tervo-Määttä, Taanila, Kaski, & Iivanainen, 2007). On note en particulier des cardiopathies très fréquentes – mais dont l'effet a été réduit par les progrès de la chirurgie cardiaque – ainsi qu'un risque plus élevé de maladie d'Alzheimer (Katz & Lazcano-Ponce, 2008 ; Määttä, Tervo-Määttä, Taanila, Kaski, & Iivanainen, 2007) – risque apparu récemment avec l'augmentation de l'espérance de vie des personnes (de 12 ans en 1940 à ~60 ans aujourd'hui : Bittles, Bower, Hussain, & Glasson, 2007).

De manière générale, les individus avec T21 rapportent eux-mêmes être heureux et sont positifs à propos de leur vie (Skotko, Levine, & Goldstein, 2011). De plus, d'après Skotko et collègues (2011), l'estime de soi et le bonheur des participants à leur étude n'étaient pas corrélés avec leurs capacités fonctionnelles : bien qu'un individu avec T21 soit plus ou moins autonome, il n'en est pas moins heureux. L'opinion générale du public envers la T21, et plus généralement envers la déficience intellectuelle, est souvent biaisée et dépréciative, ce qui pourrait être un frein supplémentaire à l'intégration des individus concernés.

Bien que les personnes avec T21 accusent un déficit de performance dans la plupart des étapes du développement quand on les positionne sur une trajectoire développementale typique, leurs parents jugent qu'ils sont néanmoins capables de vivre en autonomie, ou encore de voyager seuls et de travailler (de Graaf et al., 2019). La plupart des différences avec l'individu typique sont communément imputées strictement aux capacités cognitives réduites de l'individu avec T21. Cependant les caractéristiques du phénotype de l'individu avec T21 seraient tout autant à mettre en cause surtout pour l'exécution de certaines tâches, notamment la production de parole (Silverman, 2007). Les capacités langagières de l'individu avec T21 sont globalement altérées (Smith & Stoel-Gammon, 1983 ; Stoel-Gammon, 1997) avec spécifiquement un déficit de production de la parole qui contraste avec les compétences en compréhension et avec les capacités d'expression via d'autres canaux et notamment via la gestualité manuelle (Kumin,

2012). Dans un questionnaire adressé à 2658 parents résidant aux États-Unis ou aux Pays-Bas, leur demandant d'évaluer les compétences de leurs enfants avec T21, de Graaf et collègues (2019) ont observé que 3 parents sur 4 jugent que leur enfant parlent « raisonnablement bien » à l'âge adulte. Cependant, les parents sont habitués à la parole de leur enfant et seraient plus à même de le comprendre et donc auraient tendance à surévaluer ses performances par rapport à une évaluation par d'autres personnes non familières (Souza et al., 2013). Les parents évaluent d'ailleurs l'intelligibilité de leur enfant comme étant meilleure quand il s'adresse à des personnes connues que non connues (Kumin, 1994). La capacité à se faire comprendre par la parole joue un rôle important dans la vie de tous les jours et dans la qualité de vie de l'individu. En effet, les déficiences qui touchent à la parole de l'individu impactent sa qualité de vie, et ses relations sociales et professionnelles (Smith, Verdolini, Gray, Nichols, Lemke, Barkmeier, Dove, & Hoffman, 1996). Les personnes souffrant de telles déficiences rapportent des difficultés à parler au téléphone, ou bien la nécessité de répéter afin de pouvoir être comprises. Concernant spécifiquement les locuteurs avec T21, les difficultés de parole ne sont pas réductibles à la déficience intellectuelle. Elles s'inscrivent d'abord dans des problématiques spécifiques au niveau de la phonation et de la sphère orofaciale sur les plans moteurs, anatomiques et sensoriels (Cleland, Wood, Hardcastle, Wishart, & Timmins, 2010). La parole produite par l'individu avec T21 serait ainsi altérée depuis les poumons jusqu'à la sortie du conduit vocal (pour une revue extensive, voir Arumugam et al., 2015 ; Kent & Vorperian, 2013). L'articulation des sons de la parole pour les locuteurs avec T21 et un vrai défi et passe par une prise en charge très précoce, débutant par des exercices du développement de la sensibilité sensorielle orale et péri-orale dès la naissance.

2.5.2 Les spécificités motrices, anatomiques, physiologiques, et sensorielles des systèmes phonatoire et orofacial pouvant impacter la parole des personnes avec T21

La T21 se caractérise par des spécificités anatomiques, physiologiques, motrices et sensorielles des systèmes phonatoire et orofacial qui sont plus ou

moins marquées d'un individu à l'autre. On note de plus une prévalence plus importante de problèmes de santé liés à la sphère orofaciale, telle que rapportée par les parents, chez les personnes avec T21 que chez leurs frères et sœurs tout-venant (Hennequin, Allison, & Veyrune, 2000).

La T21 est souvent associée à la dysarthrie et/ou l'apraxie de la parole dès l'enfance (Rupela, Velleman, & Andrianopoulos, 2016). En résultent des profils de production de parole complexes, desquels peuvent rendre compte chacun des troubles moteurs mentionnés ci-dessus qui peuvent servir de cadre afin de mieux comprendre les difficultés de l'individu avec T21 pour parler. La dysarthrie désigne un trouble moteur d'origine neurologique perturbant la production de parole et qui peut-être reliée à différentes étiologies (Liss, Spitzer, Caviness, Adler, & Edwards, 2000). Elle empêche la bonne transmission des messages neuraux vers les muscles du corps impliqués dans la production de parole. Il s'agit donc d'un trouble lié à la transmission des commandes motrices qui touche spécifiquement la parole et l'altère à plusieurs niveaux (systèmes phonatoires et articulatoires) ce qui résulte ainsi en un déficit d'intelligibilité. Cette pathologie fait l'objet d'une recherche attentive afin de comprendre ce qui différencie la parole typique de celle pathologique et de mettre le jour sur les processus moteurs nécessaires à une bonne communication (Fougeron et al., 2010). Cependant, le trouble de la production de la parole dans la T21 ne se réduit pas à la dysarthrie : c'est un trouble complexe qui est parfois plutôt identifié comme « apraxique » (déficience neurologique perturbant la planification des mouvements moteurs de parole) mais qui présente aussi des troubles moteurs non spécifiques (Rupela et al., 2016). On notera aussi des difficultés de coordination entre les lèvres et la bouche ainsi qu'une tendance à une ouverture verticale excessive (Hennequin, Faulks, Veyrune, & Bourdiol, 1999 ; Spender, Dennis, Stein, Cave, & Percy, 1995).

D'autre part, da Silva et collègues (2010) ont mis en évidence des capacités pulmonaires réduites chez un groupe de 15 individus avec T21 par rapport à un groupe tou de même taille (da Silva et al., 2010). Les personnes avec T21 auraient aussi plus de mal à contrôler la pression intra-orale lors de la production de parole que des personnes typiques (Rosin, Swift, Bless, & Kluppel Vetter, 1988). Un bon

contrôle du système respiratoire est pourtant crucial pour produire la parole puisque c'est l'air expulsé par les poumons qui permet, lors de son passage à travers le larynx, de produire le voisement.

La T21 se caractérise également par une anatomie spécifique de la sphère orofaciale. Bien que la langue aurait une taille moyenne comparable à celle observée chez les locuteurs typiques (Guimaraes, Donnelly, Shott, Amin, & Kalra, 2008 ; Macho, Andrade, Areias, Coelho, & Melo, 2014), la taille réduite de la cavité orale (Borghi, 1990) résulterait en une macroglossie relative (Guimaraes et al., 2008 ; Hennequin et al., 1999), c'est-à-dire une langue trop grosse relativement à l'espace buccal disponible et donc gênante pour articuler précisément et correctement les sons de parole. Si la taille et le volume pharyngaux ne semblent pas altérées (Xue, Kaine, & Ng, 2010), les personnes avec T21 présentent généralement un palais de forme atypique, dite en escalier (Kent & Vorperian, 2013), accentuant les limites articulatoires imposées par la langue. D'après le rapport des Centers for Disease Control and Prevention (2006) qui ont répertorié 18 types de déficiences à la naissance et leur prévalence aux Etats-Unis, la déficience la plus fréquente serait la présence d'une fente palatine (l'absence partielle ou complète de la voûte buccale séparant la cavité orale de la cavité nasale) suivie de la T21, Il souligne de plus un risque de comorbidité entre ces deux déficiences relativement important. Plus généralement, on associe la T21 à un palais plus haut, plus court et plus étroit que chez les individus typiques (Hennequin et al., 1999). De telles spécificités peuvent provoquer des anomalies occlusales (Macho et al., 2014) et gêner certaines fonctions comme la mastication, la déglutition mais aussi la production de parole. Lors de l'articulation des sons de parole, notamment de certaines consonnes, la réalisation d'une constriction entre la langue et le palais est cruciale ; les anomalies combinées du palais et de la langue nécessitent alors que le locuteur procède à un certain nombre d'ajustements lors de la production de parole. Cela peut entraîner une plus grande variabilité acoustique dans les sons de parole produits, par rapport à celle observée chez des locuteurs tout-venant ainsi que des réajustements de stratégies articulatoires dus à un espace d'articulation consonantique restreint, pouvant provoquer la

neutralisation de certains contrastes, tels que celui entre les consonnes [d] et [g] par exemple comme c'est le cas chez les enfants avec fente palatine (Béchet, Hirsch, Fauth, & Sock, 2012).

La T21 se caractérise aussi par des anatomies dentaire et musculaire atypiques (Hennequin et al., 1999 ; Kent & Vorperian, 2013). La disposition des dents ainsi que leur nombre parfois atypique (Hennequin et al., 1999) peuvent gêner la langue dans sa position de repos mais aussi pour l'articulation de certaines consonnes telles que les apico-dentales. Concernant les muscles orofaciaux, il est noté que certains seraient peu différenciés ou même inexistants en comparaison avec la population tout-venant (Kent & Vorperian, 2013). Au delà d'une configuration différente des muscles des articulateurs de parole, un autre trait caractéristique de la T21 est une hypotonie générale, potentiellement liée à un seuil d'activation musculaire plus élevé (Connaghan & Moore, 2013 ; da Silva et al., 2010 ; Latash, Wood, & Ulrich, 2008 ; Morris, Vaughan, & Vaccaro, 1982). Ainsi, activer un muscle et le maintenir étiré serait plus difficile pour les personnes avec T21 que pour les personnes tout-venant. On constate aussi un mauvais contrôle de la pression de la langue (Hashimoto, Igari, & Hanawa, 2014). Notons que certaines des anomalies anatomiques citées ci-dessus résultent directement de l'expression phénotypique de la particularité chromosomique alors que d'autres sont des conséquences de mécanismes adaptatifs que les individus mettent spontanément en place pour compenser certaines difficultés fonctionnelles (e.g. Hashimoto et al., 2014 ; Hennequin et al., 1999). Par exemple, l'hypotonie générale a un impact direct sur la posture des individus qui en retour peut avoir tendance, combinée à d'autres facteurs, à une mauvaise position des voies aériennes supérieures impactant la respiration. Pour améliorer l'accès des voies aériennes, les individus auraient ainsi tendance à abaisser leur mâchoire inférieure et sortir leur langue (Hennequin et al., 1999).

La plupart des individus avec T21 ont de plus des problèmes auditifs (Kent & Vorperian, 2013). Ceci aurait pour conséquence un appauvrissement du retour sensoriel auditif nécessaire pour la production de parole (Levelt, 1989, 1995), notamment lors de son acquisition, et contribuerait donc à un retard du

développement des capacités langagières orales en comparaison aux individus normo-entendants.

2.5.3 La parole de l'individu avec T21 et son intelligibilité

En partie à cause de toutes les spécificités évoquées ci-dessus, on dénote des profils de communications atypiques lors du développement de l'individu avec T21 et sa production de parole s'en trouvera affectée tout au long de sa vie (Bunton & Leddy, 2011 ; Kent & Vorperian, 2013 ; Moura et al., 2008 ; Rosin et al., 1988 ; Rupela et al., 2016). Des questionnaires auprès des parents (937 résidents des États-Unis, Kumin, 2006 ; 329 résidents en Turquie, Toğram, 2015) permettent de caractériser l'intelligibilité des personnes avec T21 sur des critères qualitatifs. Très peu de parents décrivent leur enfant comme « complètement intelligible » (Kumin : 1,5% des parents ; Toğram : 6%). Les parents rapportent aussi une plus grande difficulté pour leur enfant à produire les consonnes que les voyelles. Rappellons néanmoins que les parents jugent que leurs enfants parlent "relativement bien" à l'âge adulte (Graaf et al., 2019).

Les caractéristiques des sons de parole des personnes avec T21 ont également été étudiées en utilisant des évaluations perceptives et/ou par le biais de tests standardisés permettant de comparer les productions des enfants avec T21 avec celles d'enfants tout-venant. Dès l'enfance, les personnes avec T21 apparaissent ainsi avoir plus de mal à produire certains sons de parole que les enfants tout-venant. On notera des difficultés plus accentuées à produire les consonnes notamment les liquides, les nasales et les occlusives, des élisions consonantiques fréquentes notamment en fin de mot et dans les clusters (Barnes et al., 2009 - N = 34, âge = 4-16 ans ; Crosley & Dowling, 1989 - N = 22, âge moyen = 9 ans 8 mois ; J. Roberts et al., 2005 - N= 32, âge = 4-13 ans ; Sommers et al., 1988 - N = 45, âge = 13-22), On notera d'ailleurs que ces difficultés s'observent non seulement par rapport à des individus tout-venant mais aussi par rapport à d'autres individus ayant une autre maladie génétique impactant leurs facultés cognitives mais pas leurs articulateurs orofaciaux (Barnes et al., 2009 ; Roberts et al., 2005). Il apparaît cependant que l'intelligibilité augmente avec l'âge (Bunton &

Leddy, 2011). Dans une étude systématique des erreurs de production de 5 locuteurs adultes avec T21 (Bunton, Leddy, & Miller, 2007), l'intelligibilité des locuteurs a été mesurée dans deux tâches perceptives dont une comportait des réponses à choix multiples, et l'autre une retranscription correspondant à un choix ouvert. Les auteurs mettent en évidence en premier lieu la simplification des clusters consonantiques remplacés par des singletons en début et fin de mot, ainsi que des erreurs sur les voyelles et le lieu d'articulation des consonnes plosives et fricatives.

Deux études se sont plus particulièrement intéressées aux mouvements de la langue, mesurés par électro-palato-graphie, de locuteurs avec T21 âgés de 8 à 19 ans lors de la production des sons /t/ (Kumin, 2012) et /ʃ/ (Mahler & Jones, 2012) . Ces études ont pu notamment montré que bien que les électro-palato-grammes des individus avec T21 sont systématiquement atypiques par rapport à ceux d'individus tout-venant appariés cognitivement, certains d'entre eux parviennent malgré tout à produire les sons étudiés de façon perceptivement satisfaisante. L'étude des caractéristiques acoustiques des voyelles (durée, fréquence fondamentale et formants) montrent une plus grande variabilité dans la production des voyelles de 8 locuteurs adultes avec T21 francophones par rapport à un groupe de locuteurs tout-venant appariés en genre et âge (Rochet-Capellan & Dohen, 2015). L'espace vocalique formantique des personnes avec T21 serait potentiellement atypique (Rosin et al., 1988).

Globalement, la parole de l'individu avec T21 est altérée et son intelligibilité est réduite par rapport à celle de la parole de locuteurs tout-venant. Malgré une variabilité entre locuteurs, certains sons posent problème à l'ensemble de la population en relation directe avec certaines anomalies spécifiques qu'on retrouve chez les individus avec T21. Nous explorons par la suite plusieurs pistes de prise en charge se concentrant sur l'amélioration de ces difficultés de production de parole.

2.5.4 Prise en charge et méthodes pour améliorer l'intelligibilité des personnes avec T21

Comme on l'a vu la T21 est une condition complexe qui s'exprime différemment chez chaque individu. Comme pour chaque personne présentant des troubles de la parole, chaque prise en charge doit s'adapter aux besoins et limites de l'individu en prenant soin de ne pas sous-estimer ses limites. De l'évaluation des caractéristiques de la parole, jusqu'à l'intervention aux différents niveaux de compétences linguistiques, il existe chez les individus avec T21 une grande variabilité intra- et inter- individuelles. Rappelons de plus que les origines des troubles observables sont très variées puisque des spécificités peuvent être observées à tous les niveaux : sensoriel, moteur, anatomique et cognitif.

Il est conseillé de sensibiliser les parents à la nécessité d'une intervention la plus précoce possible pour favoriser le développement du langage et de la parole (Borrie, 2015 ; Hustad, Dardis, & McCourt, 2007 ; Keintz, Bunton, & Hoit, 2007). Il semble notamment très important d'effectuer des stimulations intra-orales précoces (Kumin, 2012). La prise en charge orthophonique doit ensuite se poursuivre le plus fréquemment possible. Elle passera évidemment par un entraînement à la production des différents sons de parole. Une forme de prise en charge peut concerner la perte d'audition observée chez les individus avec T21 (Mahler & Jones, 2012) : pousser l'individu à parler plus fort lui permet de mieux s'entendre. Comme un cercle vertueux, un meilleur retour auditif peut lui permettre d'ajuster ses gestes moteurs lorsqu'ils ne produisent pas le bon retour auditif (Levelt, 1989, 1995). Un des problèmes liés à la T21 et qui concerne la phonation est la prise d'air à l'inspiration. Pour rappel, les individus avec T21 ont des capacités pulmonaires réduites (da Silva et al., 2010) et un faible contrôle de la pression intra-orale lors de la production de sons (Rosin et al., 1988). Une étude (Casey & Emes, 2011) a ainsi montré que l'activité physique – et plus particulièrement la natation – serait un moyen d'aider les locuteurs avec T21 à réguler leurs capacités respiratoires puisqu'un entraînement de 12 semaines a montré des bénéfices en durée de phonation pour un même débit d'air.

Il est également conseillé d'utiliser dès le plus jeune âge une (ou plusieurs) méthodes de communication augmentée et alternative (CAA ; Cress & Marvin, 2003) pour venir soutenir le développement du langage et de la parole (Kumin, 2012 ; J. E. Roberts, Price, & Malkin, 2007). Ce terme désigne l'ensemble des moyens communicatifs supplémentaires visant à une meilleure communication (plus d'information sur ce site). La CAA peut être mise en œuvre à tout âge et quelle que soit la capacité langagière du locuteur. Un outil de CAA peut être de n'importe quelle nature (comportementale, technologique, etc.). On pourra citer comme exemples l'utilisation de gestes manuels et/ou de signes pour supplémenter ou remplacer la parole, l'utilisation de tablettes avec des logiciels basés sur des pictogrammes et une synthèse vocale en sortie ou encore l'utilisation de pictogrammes classés selon des règles pragmatiques adaptées à une communication orale. La CAA concerne de plus une population très variée qui a en commun une difficulté de communication par la parole mais qui peut être d'origines très variées et s'accompagner ou non d'une déficience intellectuelle.

En complément des méthodes de prise en charge et de CAA évoquées ci-dessus, d'autres portant sur l'auditeur, et non le locuteur, pourraient induire des gains d'intelligibilité pour l'auditeur tout en fournissant des perspectives intéressantes de généralisation. L'adaptation de l'interlocuteur est d'ailleurs considérée comme un type de CAA (Cress & Marvin, 2003). La communication implique en effet une coopération entre locuteur et auditeur (Pickering & Garrod, 2013) : on pourrait ainsi proposer, dans le cas de la T21, de donner à l'auditeur les moyens de pallier au moins certaines difficultés rencontrées par le locuteur pour améliorer l'efficacité de l'interaction communicative. Les voies que nous explorons ici sont celles de la perception multisensorielle et de l'apprentissage perceptif.

2.5.5 La perception multisensorielle et l'apprentissage perceptif pour compenser les limites d'intelligibilité des personnes avec T21

Dans les sections précédentes, nous avons passé en revue deux séries de paradigmes opérant sur l'auditeur plutôt que sur le locuteur pour améliorer la perception de la parole et donc, potentiellement, l'intelligibilité d'un locuteur :

l'exploitation perceptive de la multimodalité de la parole (Section 2.1.3) et l'apprentissage perceptif (Section 2.2.2). Ces méthodes pourraient-elles être utilisées pour améliorer l'intelligibilité de la parole produite par un locuteur avec T21 ?

À notre connaissance, cette question n'a pas été posée auparavant pour le cas de la T21 en particulier. En revanche, elle a été explorée pour la dysarthrie induite par d'autres causes. Concernant d'abord l'apport de la modalité visuelle dans la perception de la parole, la réponse n'est pas si évidente : la dysarthrie entravant les capacités articulatoires du locuteur concerné, l'information visuelle des gestes articulatoires pourrait elle aussi être altérée et donc, ne pas apporter le même bénéfice pour la perception que dans le cas de locuteurs tout-venant. Les études ayant exploré ce point semblent suggérer que la présence de la modalité visuelle pourrait améliorer l'intelligibilité de la parole dysarthrique. Cependant, cette amélioration ne serait effective que pour les locuteurs ayant une dysarthrie jugée légère (Hustad & Cahill, 2003 ; Keintz et al., 2007) ou modérée (Borrie, 2015 ; Hustad, 2007 ; Hustad et al., 2007) mais pas nécessairement pour un degré de dysarthrie sévère (Hustad & Cahill, 2003 ; Hustad et al., 2007 ; pas de bénéfice en AV dans Keintz et al., 2007). Bien que les scores d'intelligibilité rapportés ne reflètent pas directement les capacités de compréhension des auditeurs (Hustad, 2008), ils dénotent que l'information visuelle convoyée par certains locuteurs dysarthriques peut parfois aider leurs interlocuteurs. Particulièrement, une étude (Borrie, 2015) s'est intéressée à l'apport du visuel afin de percevoir la parole d'un locuteur avec dysarthrie. Des mesures de reconnaissance des phonèmes et de mots a mis en évidence un bénéfice de la présence d'information visuelle en plus de l'information auditive.

Comme nous l'avons vu en Section 2.3, l'apprentissage perceptif peut aussi permettre de mieux percevoir une parole différente, que celle-ci soit prononcée avec un accent étranger (voir Section 2.2.2, artificiellement distordue (Davis, Johnsrude, Hervais-Adelman, Taylor, & McGettigan, 2005 ; Hervais-Adelman, Davis, Johnsrude, & Carlyon, 2008 ; Hervais-Adelman, Davis, Johnsrude, Taylor, & Carlyon, 2011), ou synthétique (Francis, Nusbaum, & Fenn, 2007 ; Greenspan,

Nusbaum, & Pisoni, 1988 ; Schwab, Nusbaum, & Pisoni, 1985). Concernant la parole dysarthrique, plusieurs études (Borrie, McAuliffe, & Liss, 2012 ; D'Innocenzo, Tjaden, & Greenman, 2006 ; Kim, 2015 ; Liss, Spitzer, Caviness, & Adler, 2002 ; Tjaden & Liss, 1995) rapportent en effet une efficacité de l'apprentissage perceptif pour améliorer l'intelligibilité, et ce dans des délais brefs et avec différents types et quantités d'énoncés. Nous allons dans ce qui suit nous intéresser plus particulièrement à l'étude de Borrie et Schäfer (2015) qui a évalué les effets de l'apprentissage perceptif d'un locuteur américain avec une dysarthrie modérée sur sa perception. Cette étude est proche de l'expérience princeps de Adank, Hagoort et Bekkering (2010 ; cf. Section 2.4.3 pour plus de détails) comparant différentes modalités de familiarisation sur la compréhension auditive de phrases. Dans l'étude de Borrie et Schäfer, on retrouve un pré- et un post-test (perception de phrases plongées dans du bruit) avec une phase d'apprentissage perceptif entre les deux. Chacun des tests comportait 40 phrases syntaxiquement plausibles mais sémantiquement imprévisibles. Les 100 participants devaient transcrire le plus de mots contenus dans la phrase qu'ils venaient de percevoir. Les participants étaient séparés en 5 groupes équilibrés : Contrôle (C ; ne participait pas à la phase d'apprentissage perceptif, et servait donc de référence pour contrôler un potentiel phénomène d'apprentissage entre les deux tests) – Perception Auditive (A ; entendait simplement les phrases sans retour) – A + texte écrit indiquant la phrase prononcée (AW) – A + tâche d'imitation de la phrase perçue (AI) – AWI (phrase écrite qu'il fallait ensuite imiter). Lors de l'apprentissage perceptif, les participants (tous sauf le groupe C) percevaient 80 phrases sémantiquement et syntaxiquement correctes sans bruit, contrôlées en nombre de syllabes et mots. Les performances de chaque groupe lors des pré- et post-test ont ensuite été comparées (voir Figure 13). De plus, les imitations des participants ont été analysées selon deux critères acoustiques : la fréquence fondamentale et le débit de parole.

On observe sur la Figure 13 que tous les groupes se comportent de la même façon en pré-test, avec des pourcentages de mots correctement détectés équivalents. Le groupe C ne voit pas sa performance s'améliorer entre pré- et post-tests,

contrairement au groupe A. L'apprentissage perceptif auditif seul a donc permis à ce groupe de mieux percevoir les mots en post- qu'en pré-test. De même, l'information écrite (groupe AW) et l'imitation par les participants (AI) renforcent cet apprentissage par rapport à la simple tâche d'écoute des stimuli (A). Finalement, la plus grande progression entre pré- et post-test en terme de performances perceptives est obtenue lorsque toutes les informations sont disponibles (groupe AWI). De plus, il apparaît que les meilleurs imitateurs (sur les deux critères acoustiques mentionnés) sont ceux qui obtiennent les meilleures performances soutenant l'idée que le fait de s'approcher au plus près des

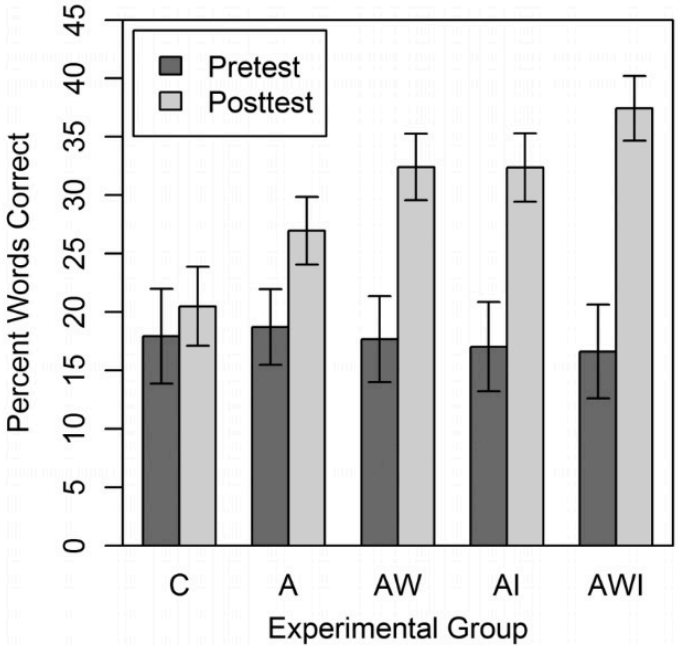


Figure 13 : Extrait de Borrie & Schäfer (2015). Pourcentage de mots corrects (*Percent Words Correct*) en ordonnée en fonction du groupe expérimental (*Experimental Group* – les associations entre groupe et lettres sont dans le texte) pour les deux Pré- et Post-tests (respectivement en gris foncé et clair).

productions du locuteur, fournirait des ajustements internes moteurs plus fins qui en retour amélioraient les performances du système perceptif. Ainsi, d'une part, les participants bénéficient d'un apprentissage perceptif accru lorsque l'information (auditive, lexicale, somatosensorielle) est la plus abondante, et d'autre part, cette information somatosensorielle est d'autant plus pertinente qu'elle correspond plus précisément à l'information sensori-motrice incidente.

Les individus avec T21 ont un ensemble de déficiences (cognitive, sensorielle, motrice) découlant d'un phénotype particulier, et menant à des capacités d'expression orale limitées. Améliorer cette parole et son intelligibilité est un enjeu essentiel pour favoriser l'insertion de cette population dans la société. Or, c'est souvent sur le locuteur avec T21 que repose toute la charge de s'adapter pour se faire comprendre. Nous avons abordé deux mécanismes améliorant l'intelligibilité qui impliquent avant tout l'auditeur, lui aussi acteur de l'échange communicatif : la multimodalité de la parole et l'apprentissage perceptif. Incidemment, ces deux mécanismes engagent les représentations motrices de l'auditeur. Ainsi, ils pourraient être perturbés lors de la communication entre locuteur avec T21 et auditeur tout-venant. ∂C'est ce que visent à étudier les deux études qui seront présentées dans le chapitre suivant.

3 Partie expérimentale

Cette partie présente les résultats expérimentaux obtenus au cours de la thèse, ayant été publiés ou soumis à publication. Chacun des deux articles présentés est centré autour d'une question dont l'intérêt théorique et les enjeux pratiques seront discutés. Le lien avec les questions théoriques abordées précédemment dans cette thèse permettra de justifier à la fois les motivations pour effectuer ces études, et ce qu'apportent les résultats obtenus.

3.1 « Does the Visual Channel Improve the Perception of Consonants Produced by Speakers of French With Down Syndrome? »

Est-ce que la modalité visuelle peut améliorer la perception des consonnes produites par des locuteurs francophones avec trisomie 21 ?

Le travail présenté dans cette partie a été effectué en collaboration avec Marion Dohen, Amélie Rochet-Capellan et Silvain Gerber. Ce travail a été publié en 2018 dans la revue *Journal of Speech, Language, and Hearing Research*.

Hennequin, A., Rochet-Capellan, A., Gerber, S., & Dohen, M. (2018). Does the visual channel improve the perception of consonants produced by speakers of French with Down syndrome? *Journal of Speech, Language, and Hearing Research*, *61*(4), 957. https://doi.org/10.1044/2017_JSLHR-H-17-0112

Nous avons discuté en détail dans la Section 2 la nature multisensorielle de la parole et notamment les processus d'intégration des modalités auditive et visuelle. Cette dernière résulte en un bénéfice pour l'auditeur, à la fois en termes de charge cognitive mais surtout d'acuité de reconnaissance de la parole perçue. De plus, nous avons vu que ces performances accrues pourraient s'appuyer sur les connaissances motrices, à la fois celles de l'auditeur et celles que celui-ci peut inférer du locuteur (section 2.2). Nous argumentons dans le présent article qu'une personne avec T21 produit une parole très différente de celle produite par un locuteur tout-venant, comme cela a été avancé plus tôt dans cette thèse (section 2.5). La littérature portant sur la perception de la parole produite par les personnes avec T21 montre un déficit en intelligibilité auditive. Elle ne traite cependant pas de la dimension visuelle de la parole produite par un individu avec T21, et du potentiel apport de celle-ci dans la perception multimodale de cette

parole spécifique. Les études perceptives se sont en effet jusqu'ici centrées sur la modalité auditive. C'est sur la base de ce constat que nous avons mis en place une étude sur la perception multisensorielle de la parole produite par des locuteurs avec T21.

48 personnes tout-venant et locuteurs natifs du français ont participé à un test de perception de parole plongée dans du bruit. Ils/elles n'avaient aucune ou peu d'expérience *a priori* avec la parole produite par un locuteur avec T21. Ils/elles avaient pour tâche de répéter les stimuli qu'ils/elles percevaient dans trois modalités différentes (auditive seule – A, visuelle seule – V, audiovisuelle – AV). Les stimuli étaient des logatomes de la forme Voyelle-Consonne-Voyelle (VCV) composés de deux voyelles /a/ combinées avec les 16 consonnes du français. Ces stimuli ont été enregistrés auprès d'un total de 8 locuteurs constituant deux groupes équilibrés en genre et âge (avec T21 et tout-venant). Chaque participant de l'étude percevait l'ensemble des stimuli énoncés par 4 parmi les 8 locuteurs (2 locuteurs de chaque groupe), et ce dans chacune des trois modalités. L'ordre de présentation de ces modalités était contre-balancé sur l'ensemble des participants. Pour chaque stimulus présenté dans une modalité, le participant donnait – dans la mesure du possible – une réponse orale caractérisant ce qu'il avait perçu.

Dans un second temps, ces réponses orales ont été retranscrites puis analysées afin de comparer les réponses de l'ensemble des participants selon trois facteurs : la modalité de présentation des stimuli, le groupe du locuteur, et le stimulus présenté. Nous avons énoncé trois questions, auxquelles les données collectées devaient tenter de répondre :

(a) D'après la littérature, les locuteurs avec T21 sont moins intelligibles que les locuteurs typiques en modalité A. Qu'en est-il en modalité V ?

(b) Est-ce que l'information en modalité V, combinée à celle de la modalité A, permet d'améliorer la perception des consonnes produites par les locuteurs avec T21 ?

(c) Quel est le profil d'erreur d'identification des consonnes, sur l'ensemble des locuteurs avec T21 et en comparaison avec le groupe de locuteurs typiques, dans chacune des modalités de présentation (A, V, et AV) ?

En ce qui concerne la question **(a)**, les résultats dans la modalité A confirment que les locuteurs avec T21 sont moins intelligibles auditivement que les locuteurs typiques. C'est aussi le cas en modalité AV, mais avec une différence entre les deux groupes moins marquée qu'en A. Par contre, les deux groupes sont aussi intelligibles l'un que l'autre en modalité V. Ce résultat n'est pas lié à un effet plancher puisque les pourcentages de réponses correctes restent

significativement supérieurs au niveau du hasard. On notera néanmoins que ces résultats sont moyennés sur l'ensemble des stimuli présentés : malgré un même pourcentage global de détection, il est possible que les réponses à chaque stimulus soient différentes d'un groupe de locuteur à l'autre. Nous reviendrons sur cette inspection plus fine des données.

Quant à la question **(b)**, nous avons utilisé une mesure du gain entre les performances en modalité A et celles en AV, expliquée en détail dans l'article (Reinisch & Holt, 2014). Les données montrent qu'il n'y a pas de différence entre les groupes de locuteurs : les deux groupes de locuteurs bénéficient donc de la même manière de l'apport du visuel.

Finalement, pour répondre à la question **(c)**, nous nous sommes intéressés aux taux de bonnes réponses des participants pour chaque stimulus individuellement (16 VCV) dans chaque modalité de présentation (A, V, AV), moyennés sur chaque groupe de locuteurs (T21 ou tout-venant). Les résultats (pages 8-11 de l'article) révèlent des différences entre groupes de locuteurs dans la capacité à percevoir certaines consonnes en particulier. En regroupant ces consonnes et en s'intéressant à la transmission de traits phonologiques individuels (lieu/mode d'articulation et voisement), nous avons observé que le principal trait que les locuteurs tout-venant arrivaient significativement mieux à véhiculer que les locuteurs avec T21 était le voisement en modalité A et AV. Ce même trait phonologique n'étant pas saillant en modalité V, l'ensemble de l'information phonologique est transmise par la modalité auditive, que le locuteur appartienne au groupe avec T21 ou tout-venant.

Pour conclure, cet article rapporte que la parole produite par un ensemble de locuteurs avec T21 est moins intelligible auditivement, mais que ce n'est pas le cas de la parole visuelle, malgré des caractéristiques anatomiques et motrices particulières. L'auditeur tout-venant bénéficie de cette information visuelle, et celle-ci pourrait être exploitée dans des évaluations et traitements orthophoniques ou tout simplement dans la communication dans la vie quotidienne avec un individu avec T21.

JSLHR

Research Article

Does the Visual Channel Improve the Perception of Consonants Produced by Speakers of French With Down Syndrome?

Alexandre Hennequin,[a] Amélie Rochet-Capellan,[a] Silvain Gerber,[a] and Marion Dohen[a]

Purpose: This work evaluates whether seeing the speaker's face could improve the speech intelligibility of adults with Down syndrome (DS). This is not straightforward because DS induces a number of anatomical and motor anomalies affecting the orofacial zone.
Method: A speech-in-noise perception test was used to evaluate the intelligibility of 16 consonants (Cs) produced in a vowel–consonant–vowel context (Vo = /a/) by 4 speakers with DS and 4 control speakers. Forty-eight naïve participants were asked to identify the stimuli in 3 modalities: auditory (A), visual (V), and auditory–visual (AV). The probability of correct responses was analyzed, as well as AV gain, confusions, and transmitted information as a function of modality and phonetic features.
Results: The probability of correct response follows the trend AV > A > V, with smaller values for the DS than the control speakers in A and AV but not in V. This trend depended on the C: the V information particularly improved the transmission of place of articulation and to a lesser extent of manner, whereas voicing remained specifically altered in DS.
Conclusions: The results suggest that the V information is intact in the speech of people with DS and improves the perception of some phonetic features in Cs in a similar way as for control speakers. This result has implications for further studies, rehabilitation protocols, and specific training of caregivers.
Supplemental Material: https://doi.org/10.23641/asha.6002267

Managing to produce intelligible speech sounds is a challenge for people with Down syndrome (DS). As a result, parents, speech therapists, and researchers in speech sciences try to provide them with appropriate help, primarily oriented toward diagnostic and improvement of acoustic intelligibility (Kent & Vorperian, 2013; Kumin, 2012; Meyer, Theodoros, & Hickson, 2017). It is however well known that, in face-to-face communication, people do not use only acoustic information to process speech but also visual (V) information (e.g., lipreading). This information is particularly useful when the acoustic signal is degraded, as is the case in noisy environments (e.g., Schwartz, Berthommier, & Savariaux, 2004), and also for speech produced by a nonnative speaker (Reisberg, McLean, & Goldfield, 1987) or for dysarthric speech (Borrie, 2015; Hustad, Dardis, & McCourt, 2007). V information could therefore also improve the intelligibility of speakers with DS.

DS, however, induces craniofacial, occlusal, and dental anomalies and weak and poorly differentiated intraoral and facial muscles (Arumugam et al., 2015; Kent & Vorperian, 2013). These specificities could affect the V and audio information conveyed during speech production in DS. Is the V information preserved in speech produced by people with DS? Are some speech sounds better perceived when listeners can see the speakers' face? In particular, what is the contribution of the V channel to the perception of consonants and the transmission of phonetic features? In this article, we address these issues by analyzing the perception, by non-familiarized participants, of vowel–consonant–vowel (VoCVo) sequences produced by young adults with DS versus control (Ctr) speakers.

What is DS?

DS is a common genetic condition related to the presence of an extra chromosome 21. It is the best-known genetic origin of intellectual deficiency (Katz & Lazcano-Ponce, 2008). DS is present worldwide, but its live births' prevalence varies depending on the country, mainly in

[a]Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, 38000 Grenoble, France
Correspondence to Marion Dohen: marion.dohen@gipsa-lab.grenoble-inp.fr
Editor-in-Chief: Frederick (Erick) Gallun
Editor: Daniel Fogerty
Received March 30, 2017
Revision received July 30, 2017
Accepted December 8, 2017
https://doi.org/10.1044/2017_JSLHR-H-17-0112

Disclosure: *The authors have declared that no competing interests existed at the time of publication.*

relation to maternal age, health care facilities, and fetal termination politics (cf. Loane et al., 2013; Parker et al., 2010). As an illustration, DS concerned ~570 newborns in France in 2012 (Haute Autorité de Santé, 2015) and ~5,657 newborns per year in the United States in 2004–2006 (Parker et al., 2010). The life expectancy of people with DS has increased from 12 years in 1940 to about 60 years nowadays (Bittles, Bower, Hussain, & Glasson, 2007). Providing adapted medical care and educational support and promoting the social integration of these persons are worldwide challenges. Improving their communication is part of it.

Intelligibility of Speech Produced by People With DS

Speech intelligibility is frequently reported as impaired for speech produced by people with DS with crucial consequences on their social participation and integration. Parental surveys revealed that, on a 10-point scale (1 corresponding to *completely unintelligible* and 10 to *completely intelligible*), intelligibility is rated on average between 4 and 5 (Kumin (2006): n = 1,620, age = 1 to 21 years, mean age = 8.2, USA; Toğram (2015): n = 319, age = 1 to 19 years, mean age = 5.3, Turkey). Only a minority of parents evaluated their child as being completely intelligible (Kumin: 1.5% of the parents; Toğram: 6%). More parents reported systematic or frequent difficulties for consonant (Kumin: 64.7%; Toğram: 45.5%) and, to a lesser extent, for vowel (Kumin: 42.4%; Toğram: 33.8%) production. In these surveys, parents evaluated the intelligibility of their child on the basis of everyday experience and communication in natural settings. V correlates of speech are thus implicitly integrated in these evaluations even if they were not specifically evaluated.

Kent and Vorperian (2013) reviewed the clinical and experimental studies on speech production in DS from 1950 to 2012. They reported that most of the examined studies on speech intelligibility were based on transcriptions of audiotaped speech (narrative, conversational, picture naming, etc.), intelligibility being quantified as the proportion of complete and intelligible utterances. The percentage of correct consonants, calculated on the basis of transcriptions by speech therapists, was also a frequent indicator because "it has been found to be correlated with speech intelligibility" and "is a good index of speech disorder severity" (Barnes et al., 2009). Based on these measures, acoustic intelligibility was found to be reduced in children and/or adolescents with DS when compared with typical speakers matched in nonverbal mental age (Rupela, Velleman, & Andrianopoulos, 2016; see Kent & Vorperian, 2013, for a review of studies before 2012). Intelligibility of children with DS also appears to improve with chronological age (Chapman & Hesketh, 2001; Rosin, Swift, Bless, & Vetter, 1988). Surprisingly, as underlined by Kent and Vorperian (2013), few studies used methods from speech production and perception research to investigate the phonetic intelligibility of people with DS.

Bunton, Leddy, and Miller (2007) audiotaped five male adult speakers with DS while they were producing lists of words chosen to evaluate 19 minimal-paired phonological contrasts (single word production). The productions were then transcribed by five experts and used as stimuli in a multichoice perception test involving 10 naïve participants. The two groups evaluated overall intelligibility consistently showing high variability between speakers with DS. A detailed analysis suggested that the largest proportions of errors were observed for initial and final clusters, which were often misperceived as singletons. The proportion of errors was also relatively high for pairs contrasting in place of articulation for both stop and fricative manners and for vowels in the front–back, high–low, and long–short dimensions. In a following X-ray study, Bunton and Leddy (2011) analyzed tongue movements during vowel production by two speakers with DS. They found a reduced F1/F2 acoustic vowel space in speakers with DS compared with Ctrs and a reduction of the articulatory space. A reduced F1/F2 space was also observed for children by Moura et al. (2008).

Based on transcriptions by trained listeners, Timmins, Hardcastle, Wood, and Cleland (2011) found that /t/ was produced correctly in average in 71.5% of the trials by children with DS (n = 26; mean age = ~13), but in 100% of the trials when produced by typically developing children matched in cognitive age. Similarly, Timmins, Cleland, Wood, Hardcastle, and Wishart (2009) reported that /ʃ/ was produced correctly in 46% of the trials by children with DS (n = 20, mean age = ~13), but in more than 90% of the trials in a Ctr group. Children with DS were reported to produce a nonsibilant fricative instead of /ʃ/ but also, to a lesser extent, a nasal, a plosive, or a liquid. Liquid and nasal simplifications were also outlined in children with DS in other studies (Crosley & Dowling, 1989; Sommers, Patterson, & Wildgen, 1988).

Most of the studies on speech sound disorders in DS (for a more complete analysis, see Kent & Vorperian, 2013) focused on transcriptions by specialists and, more rarely, on acoustical or articulatory analyses or perceptual evaluations by nonspecialists. In everyday life, the social integration of people with DS depends on their ability to be understood by nonfamiliarized listeners. Moreover, to our knowledge, there is no published work systematically analyzing the perception of consonants produced by adult speakers with DS and/or evaluating the contribution of V information to this perception.

Causes of Intelligibility Reduction in DS

Speech impairment in people with DS can be linked to various well-known types of difficulties induced by the chromosomal aberration, including breathing limits, hearing loss, malformations of speech articulators related to craniofacial anomalies, and neuromuscular issues (Kent & Vorperian, 2013). As an illustration, the size of the oral cavity was reported to be smaller in people with DS than in typical individuals (Borghi, 1990), in relation to an underdevelopment of midface bones. By contrast, pharyngeal length and volume (Xue, Kaine, & Ng, 2010) and tongue

size (Guimaraes, Donnelly, Shott, Amin, & Kalra, 2008; Macho, Andrade, Areias, Coelho, & Melo, 2014) were found to be average. Put together, these factors result in an atypical resonance cavity, a well-known relative macroglossia and occlusal/dental anomalies. Movements are also specifically impaired in people with DS. Hypotonia, low muscle tone, is a commonly reported feature that seems to affect all muscles, including facial and intraoral ones (e.g., Connaghan & Moore, 2013; Latash, Wood, & Ulrich, 2008). All these anomalies contribute to a disorder in speech sound articulation.

Disorders in articulation and in prosody, fluency, and voice are observed to various degrees in people with DS and all contribute to speech intelligibility reduction (Bunton et al., 2007). A major point is that this intelligibility reduction is not only due to intellectual deficiency but is structural as well (Cleland, Wood, Hardcastle, Wishart, & Timmins, 2010): Receptive speech skills are usually better than expressive ones in people with DS. In a recent article, Rupela et al. (2016) suggest that the motor disorder of speech production in children with DS is a complex and variable combination of symptoms of childhood apraxia, as well as childhood dysarthria and "Motor Speech Disorder–Not Otherwise Specified."

Could V Information Help Perceive Speakers With DS?

Definitions of intelligibility usually include the listener. Hence, it could be "broadly defined as the accuracy with which a listener is able to decode the acoustic signal of a speaker" (Hustad & Cahill, 2003). But, as underlined by De Gelder and Bertelson (2003), in everyday life, perceivers always combine different sensory inputs to make perceptual judgments. This is all the more true for speech: It is not only heard; it is also seen. The role of V information in speech perception has been well established: When we look at a speaker while listening to her, what we perceive is actually an integration or binding of V and auditory information (Massaro, 1987; reviews: Campbell, 2008; Peelle & Sommers, 2015). Not only does seeing the speaker help, for example, identify place of articulation for consonants (Summerfield, 1987), but it also provides temporal information on when crucial acoustic cues may occur focusing the listener's attention on these cues (Schwartz et al., 2004) and helping auditory stream segregation (Carlyon, Cusack, Foxton, & Robertson, 2001) and speech detection in noise (Grant & Seitz, 2000). V information is particularly relevant in noisy environments, when the quality of the acoustic signal is reduced (Sumby & Pollack, 1954; see review: Peelle & Sommers, 2015). Such a paradigm is used very frequently in audiovisual speech perception research in order to put forward V enhancement avoiding a ceiling effect in the auditory alone modality (e.g., Bernstein, Auer, & Takayanagi, 2004; Sumby & Pollack, 1954). Hence, typical listeners are able to extract featural information from seeing the movements of the speaker's mouth. Some phonetic features have been shown to be more prominent in the auditory channel and, others, in the V one. Summerfield (1987) reported that voicing is the most robust feature in the auditory channel, whereas the place of articulation is the most robust in the V channel. Miller and Nicely (1955) analyzed confusions between 16 English consonants perceived by five participants. The auditory signals were degraded with frequency distortion and random masking noise. The authors provide confusion matrices for five signal-to-noise ratios (−18, −12, −6, 0, +6 dB) and find that voicing and nasality are quite robust to noise unlike place. Phatak, Lovitt, and Allen (2008) also analyzed confusions between English consonants perceived by 24 participants for five signal-to-noise ratios (−12, −6, 0, +6, +12 dB; white noise). They found confusion matrices very close to those of Miller and Nicely (1955).

A few studies have investigated the contribution of V information to the perception of speech in speakers with dysarthria. Keintz, Bunton, and Hoit (2007) had 10 experts and 10 inexperienced listeners transcribe sentences produced by eight speakers with Parkinson's disease in auditory (A) and auditory–visual (AV) conditions. Results showed a better intelligibility in AV than in A but only for the less intelligible speakers. Similar observations were made by Hustad and Cahill (2003). Hustad et al. (2007) and, more recently, Borrie (2015), however, found improvement in AV compared with A only for moderate dysarthria. Results concerning speakers with severe dysarthria are inconsistent. Acknowledging the discrepancy of the results in the dysarthric population and the specific anomalies observed in DS and discussed above, it is impossible to predict from the latter results what will be observed in the specific case of DS. Also note that none of the studies described above involved a V only condition making it impossible to assess the quality of the V information in dysarthric speech. Moreover, they did not provide a specific characterization of the contribution of the V modality as a function of phonetic feature and did not make direct comparisons with typical speakers. It is also possible that listeners poorly use the V channel for less severe speech impairments in unnoisy laboratory conditions. This does not mean that they do not in everyday life, as speech is often perceived in noisy conditions and as the effect of this ecological noise might be greater for impaired speech than typical speech.

The current study analyzes the potential contribution of V information in the perception of consonants produced by adults with DS by naïve participants using a classic speech-in-noise perception paradigm. The study reported hereafter was designed to address the following questions: (a) If people with DS are less intelligible than Ctr speakers in the auditory modality, is this also the case in the V modality? (b) Does V information, when combined to auditory information, improve the perception of consonants produced by speakers with DS? (c) What are the most frequent errors made in the identification of DS speech in the auditory (A), auditory–visual (AV), and V modalities? How are phonetic features transmitted in DS speech as a function of modality and compared with typical speakers?

Method

Recording and Design of the Stimuli for the Perception Test

Speakers

The speakers, all native speakers of French, were four young adults with DS (two women and two men) and four Ctr speakers matching those with DS in age (±5 years) and gender. Speakers with DS were involved in the study in collaboration with a local association of families (Association de Recherche et d'Insertion Sociale des Trisomiques: http://www.arist.asso.fr). Ctr speakers were students recruited via advertisements at Grenoble Alpes University, France. They did not report any history of speech pathology or impairment, nor facial surgery. Table 1 summarizes the main characteristics of the speakers involved in the study. Further information about the speakers with DS is available in Supplemental Material S1 and shows that speakers with DS covered a broad range of intelligibility levels.

All the speakers gave their informed written consent to participate in the study and to be video-recorded, with restricted conditions of use of their videos. For the speakers with DS, both the person and her parent(s) signed the consent and image right forms. The purpose and conditions of the study were orally explained to the person with DS by the experimenter during a video-recorded interview in order for her to give her informed agreement to participate. All the speakers received a 15€ gift card for their participation.

Speech Sequences

The speech sequences were 16 VoCVo sequences in which Vo was always /a/ and C one consonant among /b, d, g, v, z, ʒ, p, t, k, f, s, ʃ, m, n, ʁ, l/. Table 2 summarizes the articulatory features of each consonant. Nonsense sequences were used (as in, e.g., Grant, Tufts, & Greenberg, 2007) in order to test pure phonetic intelligibility ruling out semantic and lexical influences.

Recording Procedure

The speakers were recorded in a soundproof room. They wore a head-mounted microphone (Sennheiser HSP4 EW-3) and sat in a chair in front of a loudspeaker and an HD digital camera (Panasonic HC-X920). The field of view of the camera was adjusted from above the head to shoulder level. Audio was sampled at 44100 Hz (FocusRite Scarlett 6i6 soundcard). The speakers heard the VoCVo sequences, uttered by a different speaker, through the loudspeaker and were instructed to repeat what they had heard. Repetition was chosen, rather than reading, because some speakers with DS were not able to read. Each VoCVo sequence was produced three times in random order. The audio prompts were recorded from three different female speakers. The three repetitions therefore resulted from repetition after three different speakers. When the speaker did not produce the right target, the audio prompt was played again until the two experimenters judged that the speaker had uttered her best production of the intended target. This procedure was chosen to reduce perceptual errors. The clearest production was chosen as the VoCVo exemplar for the perception test, as a trade-off between auditory and Vo quality, and on the basis of the agreement of three of the authors.

Table 1. Characteristics of the speakers with DS and their control counterparts: identifier–age–gender.

Speaker type	Characteristics
Speakers with DS	DS1 – 19 – female
	DS2 – 21 – male
	DS3 – 24 – female
	DS4 – 30 – male
Control speakers	Ctr1 – 19 – female
	Ctr2 – 22 – male
	Ctr3 – 23 – female
	Ctr4 – 25 – male

Note. DS = Down syndrome; Ctr = control.

All the acoustic stimuli were normalized at 70 dB using Praat (Boersma, 2001). A "cocktail party" noise (BDBRUIT database; Zeiliger, Serignat, Autessere, & Meunier, 1994) was then mixed with the audio stream at a signal-to-noise ratio (SNR) of −4 dB. Noise was added in order to avoid a ceiling effect, especially for the Ctr speakers. Cocktail party noise was used (rather than white noise for example) for the sake of naturalness (Alm, Behne, & Wang, 2009). The resulting sound files were mixed with the corresponding video files at a 960 × 540 pixel resolution using FFmpeg (https://www.ffmpeg.org/) to create the auditory–visual (AV) version of the stimulus. The auditory-only (A) version was obtained by replacing the video stream with a static picture of a loudspeaker and the V-only version by turning the audio stream off. This resulted in a total of 48 stimuli for each of the eight speakers (three modalities × 16 VoCVos).

Participants in the Perception Study

Forty-eight typical native speakers of French participated in the perception study (24 women and 24 men—age: M = 24.9, standard error = 3.5). All of them reported normal or corrected-to-normal vision, no auditory problems, and no speech disorder or phonological issues. Before the experiment, each participant underwent a bilateral hearing test consisting of pure-tone hearing at 30 dB for 500 Hz and 1, 2, and 4 kHz. This test confirmed that all participants had normal hearing. They all had little or no experience with people with DS and received a 15€ gift card for their participation.

Procedure

In total, 384 stimuli had to be evaluated (48 stimuli × eight speakers). In order for the duration of the perception test to be reasonable (~45 min), participants were randomly assigned to two separate subtests each consisting of the stimuli of four speakers (two with DS and two Ctrs, 192 stimuli).

Table 2. Phonetic features of the 16 consonants: voicing: unvoiced (0) and voiced (1); place of articulation: labial (L), coronal (C), and dorsal (D); manner of articulation: plosive (P), fricative (F), nasal (N), and other (O).

Feature	[b]	[d]	[g]	[v]	[z]	[ʒ]	[p]	[t]	[k]	[f]	[s]	[ʃ]	[m]	[n]	[ʁ]	[l]
Voicing	1	1	1	1	1	1	0	0	0	0	0	0	1	1	1	1
Place of articulation	L	C	D	L	C	C	L	C	D	L	C	C	L	C	D	C
Mode of articulation	P	P	P	F	F	F	P	P	P	F	F	F	N	N	O	O

Participants were seated in a quiet room, approximately 60 cm from a 24-in. screen (Dell S2415H) and wore a headset with headphones and a microphone (Audio Technica BPHS1).

The perception test was programmed using the Psychophysics Toolbox (Brainard, 1997). It was divided into three blocks, one for each modality (A, V, and AV), consisting of 64 stimuli each (16 VoCVos × four speakers). The six possible presentation orders of the blocks were balanced across participants, and the stimulus order within each block was randomized. The organization of one trial is illustrated in Figure 1. An empty gray square first appeared for 1 s. The stimulus was then played twice in a row, with a pause of 500 ms (black screen) between presentations. Participants gave their response orally when a green screen appeared, after a 1.5-s pause (red screen). They then hit a key on the keyboard to move to the next trial. Participants' responses were recorded using the microphone. Oral responses were chosen rather than written transcriptions to avoid spelling ambiguities.

Before the test phase, participants were trained to the procedure using noiseless stimuli different from those of the experiment. Two stimuli per modality were presented with the same procedure as that of the test. Participants were then informed that the stimuli in the test would be played with a background noise and were familiarized with a sample of this noise.

Instructions

Participants were informed that they would hear and/or see an audio or video or audio–video stimulus twice. They were instructed to repeat what they had perceived after the second stimulus, when the green screen appeared. They were told that the stimuli were meaningless speech sequences. No further information, such as the structure of the sequences, was provided.

Transcription of the Participants' Responses

All the responses were phonetically transcribed, and each phoneme was then assigned to one of the following five items:

$$\text{BeforeVo1-Vo1-C-Vo2-AfterVo2} \quad (1)$$

C could be either a single consonant or a cluster; Vo1 and Vo2 a vowel; BeforeVo1 and AfterVo2, anything perceived before Vo1 or after Vo2. Each item could also be empty. Table 3 provides examples of transcriptions. C was then classified into one of 17 categories: one of the 16 consonants or "other" (e.g., cluster, no consonant perceived, no response provided, ambiguous response, etc.). When it was impossible to transcribe one or several of the five items, it was annotated as "?." The first author transcribed all the responses. The last author independently transcribed half the responses. The agreement score between these two annotations was of 97.6%. Another person (speech therapy student) performed independent transcription of the other half of the responses with an agreement score of 96.8%. All transcriptions were performed blindly from stimulus and experimental condition (the transcribers did not know what the initial stimulus was nor the modality it had been

Figure 1. Organization of a trial and sequencing of trials within a block (see text for details). Colored screens and video stimuli were of the same size; the figure zooms on the stimulus screen for space reason. Color names of intermediate screens are written in brackets for gray scale printing.

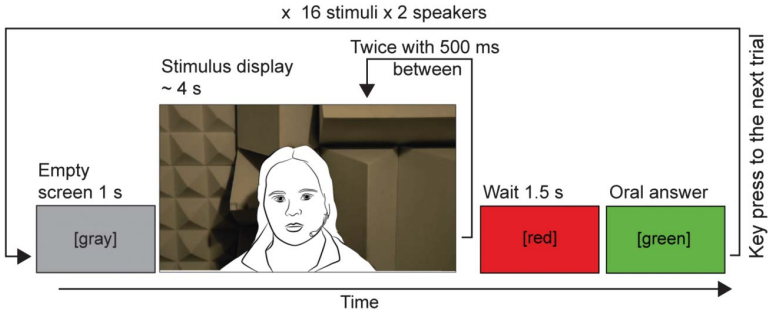


Table 3. Examples of response transcriptions for stimulus /ada/ and associated accuracy scores for the entire VoCVo (correct VoCVo) or the consonant-only (correct C); see text for details.

Response provided	Item					Score	
	BeforeVo1	Vo1	C	Vo2	AfterVo2	Correct VoCVo	Correct C
/ada/		/a/	/d/	/a/		1	1
/gʁada/	/gʁ/	/a/	/d/	/a/		0	1
/ai/		/a/		/i/		0	0
/adʁal/		/a/	/dʁ/	a	/l/	0	1
? (unintelligible)	?	?	?	?	?	0	0

Note. Vo1 and Vo2 are the first and second vowel of the VoCVo sequence, respectively. VoCVo = vowel–consonant–vowel; C = consonant.

presented in). A third person was then asked to choose between the two transcriptions for all disagreements. We kept this choice for the subsequent analyses. When this person did not agree with any of the two annotations (only 5% of the cases), the item was transcribed as "?." Note that "?" transcriptions correspond to only 1.8% of all transcriptions.

Data Analyses, Statistics, and Hypotheses

All the analyses were run using the R software (Version 3.4.2, R Development Core Team, 2008). Statistical tests were considered significant for $p < .05$. The main factors included in the analyses were *Modality* (auditory–visual [AV] vs. auditory [A] vs. V); speaker group (*Speaker_group,* DS vs. Ctr); stimulus presented (*Stimulus,* the 16 VoCVo sequences produced by the speakers); order of presentation (*Pres_order,* AV/A/V vs. AV/V/A vs. A/AV/V vs. A/V/AV vs. V/AV/A vs. V/A/AV); speaker (DS1 to DS4 and Ctr1 to Ctr4); and *participant* (48 levels). Consonants were also grouped along three phonetic features for a subpart of the analyses (cf. Table 2): voicing (voiced vs. unvoiced); place of articulation (labial vs. coronal vs. dorsal); and manner of articulation (plosive vs. fricative vs. nasal vs. other).

Analysis 1: Probability of Correct Identification of the VoCVo Sequence

We first analyzed the probability of correct responses (*Prob_correct_VoCVo*) as a function of *Modality* and *Speaker_group* to provide a global picture of VoCVo intelligibility, independently from error type. The analysis was done regardless of the presentation order of modalities because it was counterbalanced across participants but *Pres_order* effects are available in Supplemental Material S3. Based on previous work, an AV > A > V trend was expected for the Ctr group and a Ctr > DS trend in A. If the V information also plays a role in the perception of DS speech, an AV > A trend should also be observed for speakers with DS. A core question was then as follows: Does the V information benefit as much for DS than for Ctr?

The statistical analysis used was a logistic regression (in R: function *glmer* of the package *lme4,* Version 1.1.14) because response correctness is a binary variable (correct: response = stimulus—incorrect: response ≠ stimulus). *Modality, Speaker_group, Stimulus,* and their interactions were included as fixed effects and Speaker and Participant as random effects, including random slopes on the effect of *Modality, Speaker_group,* and their interaction. The predictive quality of the model was checked by computing the area under the receiver-operating characteristic (ROC) curve from the model, with values greater than 0.7 being considered as fair. Multiple comparisons were run on the model (using *glht* function of package *multcomp*; Hothorn, Bretz, & Westfall, 2008).

Paired *t* tests were used to ensure that the probabilities of correct responses were greater than 1/16 (chance) for all levels of *Modality and Speaker_group*. The corresponding Bonferroni correction was then applied to all *p* values (multiplication by the number of comparisons, i.e., six).

Analysis 2: AV Gain

A second analysis was performed to examine the effect of *Speaker_group* on AV gain relative to performance in A. AV gain was calculated for each participant as follows (Sommers, Tye-Murray, & Spehar, 2005):

$$AV\ Gain = \frac{AV - A}{100 - A} \tag{2}$$

where AV and A are the participant's scores in the respective modalities. This method was used to withdraw the impact of the participant's performance in A especially because we expect it to be *Speaker_group* dependent. AV gain provides a quantification of V enhancement relative to A-only perception (Sommers, Tye-Murray, & Spehar, 2005).

Analysis 3: Probability of Correct Identification of C

We also assessed whether the effect of *Modality* and *Speaker_group* depended on the *Stimulus* (16 levels). To do so, we considered the probability of correct identification of the consonant (C), even if Vo1 and/or Vo2 were incorrect and/or something was added before Vo1 and/or after Vo2 (cf. Table 3). The same statistical analysis as that described

for Analysis 1 was used, the only change being the observed variable (probability of correct identification of the consonant instead of *Prob_correct_VoCVo*). For space reason and overlapped conclusions with confusion matrices (see below), this analysis is provided in Supplemental Material S4.

Analysis 4: Confusion Matrices

In order to better understand perceptual errors on consonants as a function of *Speaker_group* and *Modality*, confusion matrices (*M*) were computed for each *Modality* Speaker_group* condition. They are 16 × 16 matrices in which rows correspond to the stimuli and columns to the responses. $m_{i,j}$ corresponds to the total number of responses *j* provided for stimulus *i*, all participants taken together. Note that we considered the number of observations regardless of the participant due to the small number of repetitions ($n = 2$) for each *Stimulus*Modality*Participant* condition.

Analysis 5: Transmitted Information (Entropy) of Phonetic Features

The last part of the analyses was dedicated to the "quality" of transmission of the three phonetic features (*Place, Manner,* and *Voicing*) as a function of *Modality* and *Speaker_group*. The amount of transmitted information was computed for each phonetic feature in each *Modality* and *Speaker_group*. The aim of this analysis is to examine how well a specific feature is transmitted from stimulus to response. Percentage of transmitted information (*I*) was calculated using entropy with the same method as described in Robert-Ribes, Schwartz, Lallouache, and Escudier (1998), with the following:

$$I = 100\frac{H(s,r)}{H(s)} \tag{3}$$

where $H(s,r)$ is the information shared between stimulus (s) and response (r), and $H(s)$ is the information in the stimulus. The computation is detailed in Supplemental Material S2.

Resulting *I* ranges from 0% (*no information transmitted at all from stimulus to response*) to 100% (*information systematically well transmitted*). *I_Place, I_Voicing* and *I_Manner* will further refer to the transmitted information for each phonetic feature.

On the basis of previous work (Robert-Ribes et al., 1998; Summerfield, 1987), we expected the following: for the Ctr speakers: (a) greater *I_Place* in AV than in A; (b) equivalent *I_Voicing* and *I_Manner* in AV and A. The remaining questions were as follows: Would similar trends be observed for the speakers with DS? What are the most altered features in DS speech and in which modality?

These questions were assessed using a beta regression model (function *glmmadmb* of the package *glmmADMB* 2016.0.8.3.3; Fournier et al., 2012).The complete model was used to perform the multiple comparisons, and its predictive quality was assessed by the squared correlation between the fitted and the observed values. *Modality*, *Speaker_group*, *Feature*, and their interactions were included as fixed effects and *Participant* as random effects, including random slopes on the effect of *Modality* and *Speaker_group*. Note that *I* values were transformed to be in]0; 1[to fit the requirement of the beta regression (Smithson & Verkuilen, 2006).

Results

Distribution of Response Types

A percentage of 57.4 of all the responses include at least one error (see Table 4), with more than 93.2% of these responses involving at least an error on the consonant. There are more perception errors for the speech produced by DS than Ctr speakers. We then analyzed the probability of correct response as a function of experimental condition.

Probability of Correct VoCVo Responses (Prob_Correct_VoCVo—Analysis 1)

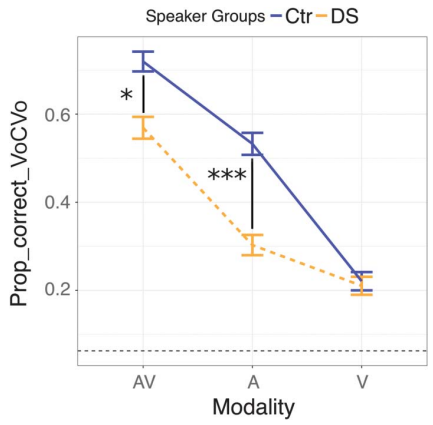
The aim of this analysis was to evaluate how accurately VoCVo sequences were perceived as a function of *Speaker_group, Modality,* and *Stimulus* (Figure 2). *Prob_correct_VoCVo* is significantly above chance in all *Speaker_group * Modality* conditions (for the six corrected *t* tests: $t(47) > 7.76$, $p < .001$). The area under the ROC curve computed from the full model (logistic regression) is 0.79 (fair

Table 4. Distribution of responses for each speaker group (Ctr vs. DS) as a function of their type (Resp. type); correct responses (Correct) and responses with an error on the following: the consonant only (Err. Cons.), the consonant and another item (Err. Cons. + Other), and another item only (Err. Other).

Resp. type	Correct	Err. Cons	Err. Cons. & Other	Err. Other	Total Err.	Conf. Err.
Ctr	2262 (49.1%)	1712 (37.2%)	469 (10.2%)	165 (3.6%)	2346	2130
DS	1662 (36.1%)	2128 (46.2%)	624 (13.5%)	194 (4.2%)	2946	2705
Total	3924 (42.6%)	3840 (41.7%)	1093 (11.9%)	359 (3.9%)	5292	4835

Note. Conf. Err. is the number of errors involving a confusion between consonants. Percentages relative to the whole data set (all responses) are provided in parentheses. Total Err. = total number of errors; Ctr = control; DS = Down syndrome.

Figure 2. Probability of correct VoCVo responses (*Prob_correct_VoCVo*) averaged across participants as a function of *Modality* and *Speaker_group*. Error bars are between-subjects 95% confidence intervals. Stars and connecting lines show significant differences (*$p < .05$, ***$p < .001$). Ctr = control; DS = Down syndrome; VoCVo = vowel–consonant–vowel; AV = auditory–visual; A = auditory; V = visual.



predictive level). Multiple comparisons were then run based on this model to analyze the effects of *Modality* as a function of *Speaker_group* and the reverse (see Tables 5 and 6). In summary, the following trends can be extracted for *Prob_correct_VoCVo*:

- AV > A > V for Ctr speakers ($p < .001$ for all comparisons) and AV > A ~ V for speakers with DS ($p < .001$ except for A–V = $p = .14$);
- Between-groups comparisons show significantly better performance for Ctr than for speakers with DS in A ($p < .001$) and AV ($p = .01$) but equivalent performances in V ($p = .97$).

AV Gain (Analysis 2)

The AV gain relative to performance in A is not significantly different for Ctr ($M = 0.39$, $SD = 0.23$) and DS ($M = 0.38$, $SD = 0.17$) speakers (paired t test: $t(47) = 0.313$, $p > .7$). V enhancement thus appears to be equivalent in both speaker groups (see Figure 3).

Probability of Correct Identification of C (Analysis 3)

We then further analyzed the probability of correctly identifying the consonant regardless of other potential errors. The analysis of the effects of *Modality* and *Speaker_group* for each consonant (cf. Supplemental Material S4) suggests that the effects depend on consonantal features. In general, labial consonants follow an AV > A ~ V trend, whereas coronals rather follow an AV ~ A > V trend and plosive dorsals an AV > A > V trend. Main differences between groups (DS < Ctr) are observed:

- in the A modality for /d/, /ʃ/, /l/, /g/ and marginally significant for /s/ (amplitude of differences: 0.32 < Ctr–DS < 0.43);
- in the AV modality only for /d/, /z/, /l/ and marginally significant for /ʒ/ (0.31 < Ctr–DS < 0.41).

In V, absolute differences between groups are always smaller than 0.18 (never significant) regardless of the consonant.

Confusion Matrices (Analysis 4)

The aim of this analysis was to examine into more details the types of errors made on consonants as a function of *Modality* and *Speaker_group*. Confusions between consonants correspond to ~98% of the errors on consonants for both groups and are detailed in the confusion matrices displayed in Figure 4. The following trends can be extracted from the analysis:

- Voicing confusions follow a trend AV ~ A < V for both Ctr speakers and speakers with DS. They are more frequent for DS than Ctr speakers in A (DS = 22%, Ctr = 9.6%) and AV (DS: 20.7%, Ctr = 7.4%) but not in V (~40% for both groups). Confusions are observed in both directions in V: voiced responses for unvoiced stimuli and the reverse. In AV and A, the tendency is to identify voiced consonants as unvoiced rather than the reverse (e.g., /b/ [resp. /d/ and /g/] identified as /p/ [resp. /t/ and /k/]).
- Manner confusions follow a trend AV < A < V for both speaker groups with no between-groups differences

Table 5. Details of the coefficient and variance estimates of the model used in Analysis 1 *(Prob_correct_VoCVo ~ Modality * Speaker_group * Stimulus + Modality * Speaker_group | Participant + Modality * Speaker_group | Speaker)*.

Fixed effects	*df*	Sum square	Mean square	*F* value
Modality	2	171.85	85.927	85.9273
Stimulus	15	589.09	39.273	39.2729
Speaker group	1	13.48	13.480	13.4801
Modality : Stimulus	30	278.46	9.282	9.2822
Modality : Speaker group	2	3.67	1.836	1.8357
Stimulus : Speaker group	29	91.72	6.115	6.1150
Modality : Stimulus : Speaker group	95	89.40	2.980	2.9801

Note. VoCVo = vowel–consonant–vowel.

Table 6. Results (estimate, standard error, *z* value, and *p* value) of multiple comparisons testing between speaker group differences as a function of modality and between modality differences as a function of speaker group.

Hypothesis		Estimate	*SE*	*z* value	*p* value	
Ctr–DS	AV	0.8647	0.2758	3.135	.013	*
	A	1.2024	0.2837	4.238	< .001	***
	V	0.2428	0.3891	0.624	.975	
	A–AV	0.3377	0.2391	1.412	.589	
A–AV	Ctr	−1.0306	0.1654	−6.232	< .001	***
	DS	−1.3683	0.1921	−7.121	< .001	***
A–V	Ctr	1.7175	0.4434	3.873	< .001	***
	DS	0.7579	0.3343	2.267	.136	
AV–V	Ctr	2.7482	0.3411	8.057	< .001	***
	DS	2.1262	0.2261	9.405	< .001	***

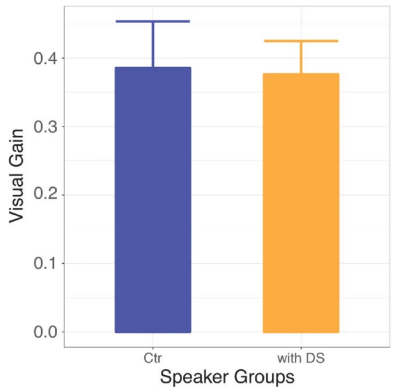
Note. These were obtained from the logistic regression model corresponding to Analysis 1 *(Prob_correct_VoCVo ~ Modality * Speaker_group * Stimulus + Modality * Speaker_group | Participant + Modality * Speaker_group | Speaker)*. Asterisks highlight significant differences. Ctr = control; DS = Down syndrome; AV = auditory–visual; A = auditory; VoCVo = vowel–consonant–vowel; V = visual.

*$p < .05$. ***$p < .001$.

(AV = 16% [DS], 11.7% [Ctr]; A = 28.8%, 24.3%; V = 36.8%, 34.2%). These confusions particularly concern nasal consonants (/m/-/n/) in all modalities.

- Place confusions follow the trend AV < V ~ A for Ctr speakers and AV < V < A for speakers with DS. They are relatively rare in AV for both speaker groups (DS = 9.2%, Ctr = 5.2%) and comparable between groups in V (DS = 17.7%, Ctr = 19.5%). The main between-groups difference is observed in A with more confusions for DS (28.3%) than Ctr (20.1%) speakers. In V, place confusions are far less frequent than in A for speakers with DS, whereas they are relatively as frequent in both modalities for Ctr speakers.
- "Other" responses are also frequently provided for some consonants (cf. /ʁ/) for both speaker groups in all modalities, the tendency being strongest in V.

Figure 3. AV gain (mean across participants) relative to performance in the A-only modality as a function of the speaker group. Error bars are between-subjects 95% confidence intervals. Ctr = control; DS = Down syndrome; AV = auditory–visual; A = auditory.



Feature Information Transmission (Entropy, Analysis 5)

We analyzed the information transmitted for each phonetic feature (see Figure 5). The pseudo-R^2 value associated with the full model (beta regression) is .79 (good predictive level).

I_Voicing – For Ctr speakers, *I_Voicing* follows a trend AV > A > V, with AV–V and A–V greater than 55% ($p < .01$), and AV–A ~ 8% ($p = .03$). For speakers with DS, the trend is the same, but the only significant difference is between AV and V (AV–V ~ 29%, $p < .01$). *I_Voicing* is also significantly greater for Ctr speakers than for speakers with DS in A and AV ($p < .04$ for both comparisons), but not in V ($p = 1$).

I_Manner – For both speaker groups, *I_Manner* follows the trend AV > A ~ V: Ctr: AV–A = 23% ($p < .01$), AV–V = 37% ($p < .01$), A–V = 15% ($p = .06$); DS: AV–A = 25% ($p < .01$), AV–V = 33% ($p < .01$), A–V = 8% ($p = .99$). *I_Manner* is comparable for both speaker groups in A and V ($p > .6$). The difference in AV is marginally significant (Ctr–DS = 8%, $p = .09$).

I_Place – For both speakers groups, *I_Place* follows the trend AV > V ~ A for Ctr speakers: AV–V = 33%, AV–A = 35% ($p < .01$ in both cases), V–A = 2.2% ($p = 1$); but less clearly for speakers with DS: AV–V = 19% ($p = .21$), AV–A = 44% ($p < .01$), V–A = 25% ($p = .6$). Differences between speaker groups are not significant (A: $p = .81$, V: $p = .99$, AV: $p = .48$). Note that, once again, it appears that whereas in V transmission of place information is equivalent between groups, it is far less efficient in A for speakers with DS than for Ctr speakers.

Discussion

The aim of this study was to characterize the quality of the V information in speech produced by people with

Figure 4. Confusion matrices for each modality and speaker group. Each cell, $m_{i,j}$, corresponds to the number of times response *j* was provided for stimulus *i*, all participants and speakers together. The number in bold on each line corresponds to the most frequent response for a given consonant. Color codes indicate the error type for the different features (see the legend below the figure). "Other" responses correspond to cases in which the response was not one of the 16 consonants (e.g., cluster, no consonant identified, no response provided, and ambiguous response). AV = auditory–visual; A = auditory; V = visual; Ctr = control; DS = Down syndrome.

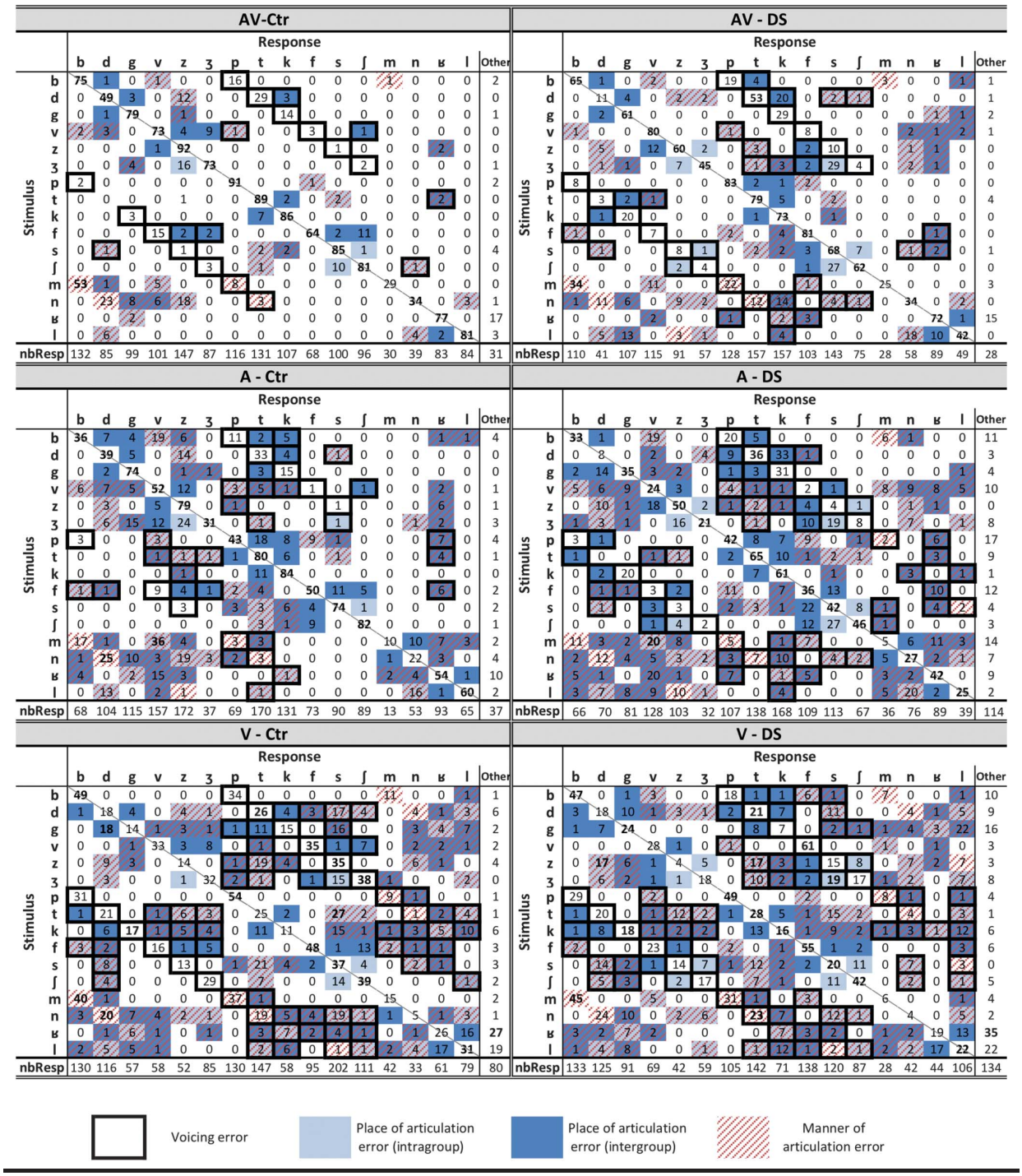
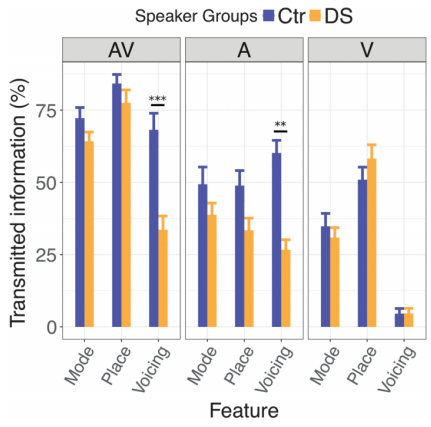


Figure 5. Percentage of transmitted information averaged across participants as a function of modality and speaker group. Significant effects between speaker groups are shown by connecting lines and asterisks (**$p < .01$, ***$p < .001$). Ctr = control; DS = Down syndrome; AV = auditory–visual; A = auditory; V = visual.



DS and its contribution to general intelligibility. In particular, it investigated the role of V information in the transmission of consonantal phonetic features by speakers with DS as compared with typical speakers matched in age and gender. To do so, a classic speech-in-noise perception test involving naïve participants was conducted in three modalities: AV (auditory–visual), A (auditory only), and V only. The results suggest that V information is relatively preserved in speech produced by people with DS despite their anatomical and motor specificities. Moreover, it improves overall intelligibility. This, however, depends on phonetic feature and consonant. The results are discussed in relation to the main questions raised in the introduction and considering previous works and methodological aspects.

Is V Information Preserved in Speech Produced by People With DS? Does It Improve Auditory Intelligibility?

Previous work extensively reported that speech intelligibility is almost always, even though to various degrees, impaired in people with DS, especially when the acoustic signal is perceived alone (e.g., Bunton et al., 2007; Kent & Vorperian, 2013; Kumin, 2006). This is confirmed by our own findings. Our first aim was to evaluate the quality of the V information in speech produced by people with DS. In this study, perception scores in V are equivalent for speakers with DS and typical speakers. This could be counterintuitive considering craniofacial, muscle, and vocal tract anomalies in DS (e.g., Kent & Vorperian, 2013; Latash et al., 2008; Macho et al., 2014). It could however be accounted for both by the interspeaker variability in the quality of the V information usually observed in typical speakers (e.g., all speakers do not provide clear V features; Mallick, Magnotti, & Beauchamp, 2015) and the listeners' ability to make use of this information (e.g., all listeners are not good at lipreading; Bernstein, Demorest, & Tucker, 2000; Mallick et al., 2015). Note that, even if they were low, the intelligibility scores in the V modality were still above chance confirming that this modality does carry information in itself. We then wanted to evaluate whether, in a situation in which processing the V information is crucial to perceive speech (speech in noise), this information could improve the intelligibility of speakers with DS. Our analyses suggest that this is the case: VoCVos produced by people with DS are globally more accurately perceived in AV than in A. Just as for typical speakers, it thus appears that seeing the speaker's face is beneficial to identify VoCVos uttered by speakers with DS. Note that V enhancement is similar in both groups showing that perceivers benefit as much of the V information for perceiving DS and Ctr speakers. V enhancement and its comparison between groups could be further investigated using a participant and a speaker-specific adaptive SNR procedure (as in Bernstein et al., 2004; Sommers et al., 2005) in order to equate performance levels in A (e.g., at 50%).

All together, the latter observations suggest that the V speech information is relatively preserved in speakers with DS, despite the anatomical and motor specificities caused by DS, and that it can be beneficially used to better perceive DS speech. Speech rehabilitation could use such findings and involve the V modality to a greater extent both in speech evaluation and rehabilitation protocols. This idea is similar to that suggested in Hustad et al. (2007) for people with dysarthria. Speech therapists could more systematically train speakers with DS to enhance their V speech cues by using, for example, systematic V feedback: simple video or ultrasound biofeedback as already used with children with different types of speech disorders (Cleland, Scobbie, & Wrench, 2015). Further work should, however, be conducted in order to extend our results to more natural speech material, such as words and sentences. It would also be interesting to compare the contribution of V information to perception by unfamiliar listeners, such as in this study, to perception by familiar listeners, such as parents, teachers, and/or professional staff. This would indeed make it possible to evaluate whether the people familiar with speakers with DS spontaneously use the V information to improve their understanding of the person or whether it would be worthwhile to train them to do so (as for example can be successfully done with patients with hearing impairment; Massaro & Light, 2004). It would also be interesting to run speaker-specific studies to assess to which extent anatomical and neuromuscular specificities in speakers with DS influence the V correlates of their speech and how these specificities interact with acoustic properties. Note that these effects could not be reliably assessed in the current study due to the small number of repetitions for each speaker imposed by experimental timing constraints. An exploratory by-speaker analysis of our data set suggests that intelligibility scores in A are similar between the speakers with DS and always smaller than for the typical speakers. By contrast, the speakers from

both groups were less distinguishable (Ctr vs. DS) in AV and even less in V.

If V information improves the perception of VoCVo sequences produced by speakers with DS, confusion and information transmission analyses clearly show that the contribution of vision depends on the consonant and, especially, on its phonetic features.

Which Phonetic Features Are Specifically Impaired in Consonants Produced by Speakers With DS? Is This Effect Modality Dependent and How?

The confusion matrix obtained for the Ctr speakers in the A modality was compared with the historical consonant confusion matrix published by Miller and Nicely (1955) for SNR = −6 dB and to that published by Phatak, Lovitt, and Allen (2008) for SNR = −6 dB. This resulted in less than 7% mean differences between their confusion matrices and ours (Miller and Nicely: M = 6.8%, SD = 10.1%; Phatak, Lovitt, and Allen = 6.4%, 11.2%). Our results are thus consistent with previous results especially considering that the language is different (English vs. French) as well as the noise type (white vs. cocktail party noise) and that the two studies used consonant–vowel sequences (vs. VoCVo in this study).

In typical speakers, labial/bilabial consonants were usually better identified in AV than in A and V (AV > A ~ V), whereas coronals rather followed the trends AV ~ A > V or AV ~ A ~ V and, plosive dorsals, the trend AV > A > V. The V information is indeed greater for labials than for coronals. Surprisingly, the perception of /k/ and /g/, however, appears to benefit from vision. Voicing was better transmitted in A than in V as classically observed (Alm et al., 2009; Summerfield, 1987). The manner of articulation followed a similar trend, even though differences were less dramatic than for voicing. Similar observations were made for speakers with DS for both place and manner of articulation. The main intergroup difference is observed for voicing; whereas it is as poorly transmitted for both groups in V, it is dramatically less well transmitted for speakers with DS than for Ctr speakers in A and AV. It therefore appears that speakers with DS have issues in producing voicing. This result is important because there are not a lot of studies reporting intelligibility of voicing in DS. Borghi (1990) had already signaled voicing errors in speech produced by people with DS (see also Bunton et al., 2007, even though results are imprecise concerning that matter). Smith and Stoel-Gammon (1983) also put forward devoicing of final stops in five children with DS. Kent and Vorperian (2013) report "increased noise in phonation" for DS speech. Also note that, whichever modality and speaker group, unvoiced responses were provided more frequently than voiced ones. This is contrary to previous findings showing that unvoiced consonants seem to be less robust to noise for typical speakers, especially babble noise (Alm et al., 2009). Interestingly, however, the latter tendency was stronger for speakers with DS than typical speakers, especially in AV and A. This, once again, puts forward the fact that voicing would be particularly affected in the speech produced by people with DS and, more specifically, that they would have a tendency to devoice voiced consonants. Because voicing is poorly transmitted in V, as for typical speakers (e.g., Binnie, Montgomery, & Jackson, 1974), adding vision cannot compensate for it. Note, however, that some researchers suggest that voicing information can partially be recovered through the V modality even though the larynx is not directly visible (Files, Tjan, Jiang, & Bernstein, 2015).

Some effects appear to be consonant specific, suggesting interactions between phonetic feature and speaker group. This is particularly the case for /d/, poorly identified for speakers with DS in all the conditions and mainly mistaken for /t/. This was also the case for typical speakers but only in V, which is trivial because voicing is not well transmitted in V. A possible explanation for this could be an effect of relative frequency of the consonants (C) in an aCa context in French words: If the /ata/ sequence is relatively more frequent in French words than the /ada/ sequence, listeners may expect more /t/ than /d/, resulting in a bias in responding /t/ rather than /d/. Note, however, that the effect is speaker group and modality dependent (e.g., for speakers with DS, /d/ responses were rare in AV whereas frequent in V), which invalidates the argument. Moreover, we found no significant correlation between the frequency of consonant occurrence in French in an aCa context (Freq_lang) and the frequency of occurrence of these consonants in the participants' responses (Freq_resp; all conditions together, R = .11). Freq_lang was computed as the frequency of occurrences of each aCa in all French words regardless of the position in the word relative to the sum of the frequencies of occurrences of all aCa (movie + book frequency, Lexique database; New, Pallier, Ferrand, & Matos, 2001). Freq_resp was computed as the number of each consonant answers divided by the total number of responses. The finding that voicing would be particularly impaired in DS could have implications for speech therapy protocols. Andrade et al. (2014) review and compare several techniques used to train voicing, some of which could easily be used with people with DS, such as the straw exercise.

It was also observed that nasal responses were less frequent than responses corresponding to other manners of articulation for both speaker groups. It may be the case that nasality was particularly affected by the type of noise used.

Finally, an intriguing finding is that, whereas place information is transmitted as efficiently in both speaker groups in V, it drops largely in A but only for speakers with DS. It could be hypothesized that, due to problems of relative macroglossia and difficulties in tongue motor control speakers with DS try to articulatorily compensate using their lips. This would compensate for place information transmission in V, resulting in no difference with Ctr speakers. This would, however, not operate anymore in A resulting in poorer place information transmission than for Ctr speakers. Similar observations, but in the reverse direction (compensation using the tongue), have

already been observed for blind speakers (Ménard, Trudeau-Fisette, Côté, & Turgeon, 2016).

Potential Influence of Methodological Limitations

One could put forward several methodological issues in this study that could influence its results. In particular, repetition tasks were used both to record the stimuli and to collect participant responses in the perceptual test.

Involving speakers with an intellectual deficiency required specific adaptations of experimental procedures. In previous work evaluating speech production in speakers with DS using a reading task and real words, repetition was required in some trials when speakers did not read correctly (cf. Bunton, Leddy, & Miller, 2007, p. 4: "If a word was mispronounced, the speaker was asked to repeat the word, if a second error occurred, the investigator read the word aloud and asked the speaker to repeat it."). On the basis of this report, we considered that repetition for all speakers was a good compromise to avoid bias between trials and speaker groups and to design a more inclusive study. However, this task may have resulted in wrong phonetic identification of the stimulus rather than pronunciation errors: Did the speakers, in particular those with DS, produce the expected stimulus? Could the lower scores observed in the perceptual test for DS be explained by wrong identification of the target utterance to be produced rather than articulation issues in achieving the target? This possible bias was first addressed in the recording procedure. As described in the methods section, each VoCVo sequence repetition was prompted by three different audio stimuli produced by three different speakers. Each stimulus was also played several times when required until the two experimenters judged that the participant did her best to produce the correct VoCVo. This reduces potential misinterpretations of what to repeat. Then, only the best repetition (e.g., the closest to the target as judged by three of the authors) was selected for the perceptual test. If this methodological approach does not exclude the bias completely (i.e., it is still possible that both typical, and even more so speakers with DS, wrongly identified the target VoCVo), it cannot account for the main results of the study (i.e., AV > A for both speaker groups and V in DS ~ V in typical). If speakers with DS produced the wrong sequence because they misinterpreted the prompt they heard, they would not be better identified in AV than in A, and identification in V would have been poorer for DS than typical speakers.

On the other hand, the repetition task used to collect participant responses during the perceptual test may have induced two problems: 1. the participant did not manage to reproduce what she had just heard; 2. this procedure requires postcoding of the responses involving interpretation by the coder. To address 1, we could have used a forced-choice task or written transcription. We did not want to choose the first option because we wanted to be sure of what the participants actually perceived (which may not be in the alternative choices, see Bunton, Leddy, & Miller, 2007, for similar issues with real words). The fact that we observed "other" responses (corresponding to none of the 16 target consonants) confirms that forced choice would not have necessarily assessed true perception. The second option was also discarded to avoid spelling ambiguities. To address Problem 2, we used multiple coding by three coders (as described in the methods section). Strong agreement between coders shows that response coding may have only had minor influence on the results.

Conclusion

The current work provides new insight relevant to the study of speech intelligibility in people with DS and to the development of speech therapy for these persons. First, it shows that, despite anatomical and motor specificities, the V speech information seems preserved in the speech of people with DS. Then, it appears that speech produced by speakers with DS can be better understood when seeing the speaker's face rather than just listening to him or her, and this V benefit is as important as for typical speakers. Part of the solution to the speech intelligibility deficit in people with DS could come from the listener himself or herself. People indeed tend not to look straight at their interlocutor with DS, often out of shyness or discomfort, but this behavior actually impairs their chances to understand what their interlocutor tries to tell them. Our results also show that the contribution of V information to the perception of consonantal features is particularly true for place of articulation and, to a lesser extent, for manner of articulation. Voicing appears to be the most altered phonetic feature in DS with a tendency toward devoicing. Vision cannot, or, at the best, barely contribute to compensate for this voicing deficit. Previous work evaluating phonetic intelligibility in adult speakers with DS mostly studied word identification using minimal pair multichoice tests or transcriptions, conducted in the auditory modality only and involving native speakers of English. This study involved native speakers of French and suggests that V information should be considered when evaluating speech intelligibility, especially in speakers with DS. Further studies involving more naturalistic speech material, investigations of speaker-specific effects, and noise effects are now required to better understand the potential role of V information in the perception of speakers with DS. Cross-linguistic studies may also help in identifying difficulties specifically related to DS.

Acknowledgment

This research has received funding from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013 Grant Agreement no.339152 "Speech Unit(e)s," awarded to PI Jean-Luc Schwartz) and from the FIRAH foundation (International Foundation of Applied Disability Research) awarded to PIs Marion Dohen and Amélie Rochet-Capellan. It was approved by the Comité d'Éthique pour les Recherches Non Interventionnelles ethics committee of Grenoble Alpes University (IRB00010290 COMUE Grenoble Alpes University IRB#1 – approval number: 2014-03-11-41) and by the ethical committee of the FIRAH. The authors thank

the Association pour la Recherche et l'Insertion Sociale des Trisomiques (Down Syndrome Research and Social Integration Association), the Établissement et Service d'Aide par le Travail—Service d'Activité de Jour (Institution and Service through Work—Day Activity Service), and the speakers who participated in this study and their families.

References

Alm, M., Behne, D. M., & Wang, Y. (2009). Audio–visual identification of place of articulation and voicing in white and babble noise. *The Journal of the Acoustical Society of America, 126*(1), 377–387. https://doi.org/10.1121/1.3129508

Andrade, P. A., Wood, G., Ratcliffe, P., Epstein, R., Pijper, A., & Svec, J. G. (2014). Electroglottographic study of seven semi-occluded exercises: LaxVox, straw, lip-trill, tongue-trill, humming, hand-over-mouth, and tongue-trill combined with hand-over-mouth. *Journal of Voice, 28*(5), 589–595. https://doi.org/10.1016/j.jvoice.2013.11.004

Arumugam, A., Raja, K., Venugopalan, M., Chandrasekaran, B., Kovanur Sampath, K., Muthusamy, H., & Shanmugam, N. (2015). Down syndrome—A narrative review with a focus on anatomical features. *Clinical Anatomy, 29*(5), 568–577. https://doi.org/10.1002/ca.22672

Barnes, E., Roberts, J., Long, S. H., Martin, G. E., Berni, M. C., Mandulak, K. C., & Sideris, J. (2009). Phonological accuracy and intelligibility in connected speech of boys with fragile X syndrome or Down syndrome. *Journal of Speech, Language, and Hearing Research, 52,* 1048–1061. https://doi.org/10.1044/1092-4388(2009/08-0001)

Bernstein, L. E., Auer, E. T., & Takayanagi, S. (2004). Auditory speech detection in noise enhanced by lipreading. *Speech Communication, 44*(1–4), 5–18. https://doi.org/10.1016/j.specom.2004.10.011

Bernstein, L. E., Demorest, M. E., & Tucker, P. E. (2000). Speech perception without hearing. *Perception & Psychophysics, 62*(2), 233–252. https://doi.org/10.3758/BF03205546

Binnie, C. A., Montgomery, A. A., & Jackson, P. L. (1974). Auditory and visual contributions to the perception of consonants. *Journal of Speech and Hearing Research, 17*(4), 619–630.

Bittles, A. H., Bower, C., Hussain, R., & Glasson, E. J. (2007). The four ages of Down syndrome. *European Journal of Public Health, 17*(2), 221–225. https://doi.org/10.1093/eurpub/ckl103

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International, 5*(9/10), 341–345.

Borghi, R. W. (1990). Consonant phoneme, and distinctive feature error patterns in speech. In D. C. Van Dyke, D. J. Lang, F. Heide, S. van Duyne, & M. J. Soucek (Eds.), *Clinical perspectives in the management of Down syndrome* (pp. 147–152). New York, NY: Springer US. https://doi.org/10.1007/978-1-4613-9644-4_12

Borrie, S. A. (2015). Visual speech information: A help or hindrance in perceptual processing of dysarthric speech. *The Journal of the Acoustical Society of America, 137*(3), 1473–1480. https://doi.org/10.1121/1.4913770

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10*(4), 433–436. https://doi.org/10.1163/156856897X00357

Bunton, K., & Leddy, M. (2011). An evaluation of articulatory working space area in vowel production of adults with Down syndrome. *Clinical Linguistics & Phonetics, 25*(4), 321–334. https://doi.org/10.3109/02699206.2010.535647

Bunton, K., Leddy, M., & Miller, J. (2007). Phonetic intelligibility testing in adults with Down syndrome. *Down Syndrome Research and Practice, 12*(1), 1–4. https://doi.org/10.3104/editorials.2034

Campbell, R. (2008). The processing of audio–visual speech: Empirical and neural bases. *Philosophical Transactions: Biological Sciences, 363*(1493), 1001–1010. https://doi.org/10.1098/rstb.2007.2155

Carlyon, R. P., Cusack, R., Foxton, J. M., & Robertson, I. H. (2001). Effects of attention and unilateral neglect on auditory stream segregation. *Journal of Experimental Psychology: Human Perception and Performance, 27*(1), 115–127. https://doi.org/10.1037//0096-1523.27.1.115

Chapman, R., & Hesketh, L. (2001). Language, cognition, and short-term memory in individuals with Down syndrome. *Down Syndrome Research and Practice, 7*(1), 1–7. https://doi.org/10.3104/reviews.108

Cleland, J., Scobbie, J. M., & Wrench, A. A. (2015). Using ultrasound visual biofeedback to treat persistent primary speech sound disorders. *Clinical Linguistics & Phonetics, 29*(8–10), 575–597. https://doi.org/10.3109/02699206.2015.1016188

Cleland, J., Wood, S., Hardcastle, W., Wishart, J., & Timmins, C. (2010). Relationship between speech, oromotor, language and cognitive abilities in children with Down's syndrome. *International Journal of Language & Communication Disorders, 45*(1), 83–95. https://doi.org/10.3109/13682820902745453

Connaghan, K. P., & Moore, C. A. (2013). Indirect estimates of jaw muscle tension in children with suspected hypertonia, children with suspected hypotonia, and matched controls. *Journal of Speech, Language, and Hearing Research, 56*(1), 123–136. https://doi.org/10.1044/1092-4388(2012/11-0161)

Crosley, P. A., & Dowling, S. (1989). The relationship between cluster and liquid simplification and sentence length, age, and IQ in Down's syndrome children. *Journal of Communication Disorders, 22*(3), 151–168. https://doi.org/10.1016/0021-9924(89)90013-0

De Gelder, B., & Bertelson, P. (2003). Multisensory integration, perception and ecological validity. *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2003.08.014

Files, B. T., Tjan, B. S., Jiang, J., & Bernstein, L. E. (2015). Visual speech discrimination and identification of natural and synthetic consonant stimuli. *Frontiers in Psychology, 6,* 878. https://doi.org/10.3389/fpsyg.2015.00878

Fournier, D. A., Skaug, H. J., Ancheta, J., Ianelli, J., Magnusson, A., Maunder, M. N., … Sibert, J. (2012). AD model builder: Using automatic differentiation for statistical inference of highly parameterized complex nonlinear models. *Optimization Methods and Software, 27*(2), 233–249. https://doi.org/10.1080/10556788.2011.597854

Grant, K. W., & Seitz, P. F. P. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. *The Journal of the Acoustical Society of America, 108*(3), 1197–1208. https://doi.org/10.1121/1.422512

Grant, K. W., Tufts, J. B., & Greenberg, S. (2007). Integration efficiency for speech perception within and across sensory modalities by normal-hearing and hearing-impaired individuals. *The Journal of the Acoustical Society of America, 121*(2), 1164–1176. https://doi.org/10.1121/1.2405859

Guimaraes, C. V. A., Donnelly, L. F., Shott, S. R., Amin, R. S., & Kalra, M. (2008). Relative rather than absolute macroglossia in patients with Down syndrome: Implications for treatment of obstructive sleep apnea. *Pediatric Radiology, 38*(10), 1062–1067. https://doi.org/10.1007/s00247-008-0941-7

Haute Autorité de Santé. (2015). *Les performances des tests ADN libre circulant pour le dépistage de la trisomie 21 fœtale* [Performances of trisomy 21 fetal screening tests by analysis of

circulating free DNA]. Retrieved from https://www.has-sante.fr/portail/upload/docs/application/pdf/2015-11/recommandation_trisomie_21.pdf

Hothorn, T., Bretz, F., & Westfall, P. (2008). Simultaneous inference in general parametric models. *Biometrical Journal, 50*(3), 346–363. https://doi.org/10.1002/bimj.200810425

Hustad, K. C., & Cahill, M. A. (2003). Effects of presentation mode and repeated familiarization on intelligibility of dysarthric speech. *American Journal of Speech-Language Pathology, 12*(2), 198–208. https://doi.org/10.1044/1058-0360(2003/066)

Hustad, K. C., Dardis, C. M., & McCourt, K. A. (2007). Effects of visual information on intelligibility of open and closed class words in predictable sentences produced by speakers with dysarthria. *Clinical Linguistics & Phonetics, 21*(5), 353–367. https://doi.org/10.1080/02699200701259150

Katz, G., & Lazcano-Ponce, E. (2008). Intellectual disability: Definition, etiological factors, classification, diagnosis, treatment and prognosis. *Salud Pública de México, 50*(2), s132–s141. https://doi.org/10.1590/S0036-36342008000800005

Keintz, C. K., Bunton, K., & Hoit, J. D. (2007). Influence of visual information on the intelligibility of dysarthric speech. *American Journal of Speech-Language Pathology, 16*(3), 222–234. https://doi.org/10.1044/1058-0360(2007/027)

Kent, R. D., & Vorperian, H. K. (2013). Speech impairment in Down syndrome: A review. *Journal of Speech, Language, and Hearing Research, 56*(1), 178–210. https://doi.org/10.1044/1092-4388(2012/12-0148)

Kumin, L. (2006). Speech intelligibility and childhood verbal apraxia in children with Down syndrome. *Down Syndrome Research and Practice, 10*(1), 10–22. https://doi.org/10.3104/reports.301

Kumin, L. (2012). *Early communication skills for children with Down syndrome: A guide for parents and professionals*. Bethesda, MD: Woodbine House.

Latash, M., Wood, L., & Ulrich, D. (2008). What is currently known about hypotonia, motor skill development, and physical activity in Down syndrome. *Down Syndrome Research and Practice (Online)*. https://doi.org/10.3104/reviews.2074

Loane, M., Morris, J. K., Addor, M.-C., Arriola, L., Budd, J., Doray, B., . . . Dolk, H. (2013). Twenty-year trends in the prevalence of Down syndrome and other trisomies in Europe: Impact of maternal age and prenatal screening. *European Journal of Human Genetics, 21*(1), 27–33. https://doi.org/10.1038/ejhg.2012.94

Macho, V., Andrade, D., Areias, C., Coelho, A., & Melo, P. (2014). Comparative study of the prevalence of occlusal anomalies in Down syndrome children and their siblings. *British Journal of Medicine and Medical Research, 4*(35), 5604–5611. https://doi.org/10.9734/BJMMR/2014/12688

Mallick, D. B., Magnotti, J. F., & Beauchamp, M. S. (2015). Variability and stability in the McGurk effect: Contributions of participants, stimuli, time, and response type. *Psychonomic Bulletin & Review, 22*(5), 1299–1307. https://doi.org/10.3758/s13423-015-0817-4

Massaro, D. W. (1987). *Speech perception by ear and eye: A paradigm for psychological inquiry*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Massaro, D. W., & Light, J. (2004). Using visible speech to train perception and production of speech for individuals with hearing loss. *Journal of Speech, Language, and Hearing Research, 47*(2), 304–320. https://doi.org/10.1044/1092-4388(2004/025)

Ménard, L., Trudeau-Fisette, P., Côté, D., & Turgeon, C. (2016). Speaking clearly for the blind: Acoustic and articulatory correlates of speaking conditions in sighted and congenitally blind speakers. *PLoS ONE, 11*(9). https://doi.org/10.1371/journal.pone.0160088

Meyer, C., Theodoros, D., & Hickson, L. (2017). Management of swallowing and communication difficulties in Down syndrome: A survey of speech-language pathologists. *International Journal of Speech-Language Pathology, 19*(1), 87–98. https://doi.org/10.1080/17549507.2016.1221454

Miller, G., & Nicely, P. (1955). An analysis of perceptual confusions among some English consonant. *The Journal of the Acoustical Society of America, 27*(2), 338–352. https://doi.org/10.1121/1.1907526

Moura, C. P., Cunha, L. M., Vilarinho, H., Cunha, M. J., Freitas, D., Palha, M., . . . Pais-Clemente, M. (2008). Voice parameters in children with Down syndrome. *Journal of Voice, 22*(1), 34–42. https://doi.org/10.1016/j.jvoice.2006.08.011

New, B., Pallier, C., Ferrand, L., & Matos, R. (2001). Une base de données lexicales du français contemporain sur internet : LEXIQUE ™ [A lexical database for contemporary French : LEXIQUE ™]. *L'année Psychologique, 101*(3), 447–462. https://doi.org/10.3406/psy.2001.1341

Parker, S. E., Mai, C. T., Canfield, M. A., Rickard, R., Wang, Y., Meyer, R. E., . . . Correa, A. (2010). Updated national birth prevalence estimates for selected birth defects in the United States, 2004–2006. *Birth Defects Research Part A—Clinical and Molecular Teratology, 88*(12), 1008–1016. https://doi.org/10.1002/bdra.20735

Peelle, J. E., & Sommers, M. S. (2015). Prediction and constraint in audiovisual speech perception. *Cortex, 68,* 169–181. https://doi.org/10.1016/j.cortex.2015.03.006

Phatak, S. A., Lovitt, A., & Allen, J. B. (2008). Consonant confusions in white noise. *The Journal of the Acoustical Society of America, 124*(2), 1220–1233. https://doi.org/10.1121/1.2913251

R Development Core Team. (2008). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

Reisberg, D., McLean, J., & Goldfield, A. (1987). Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli. In B. Dodd & R. Campbell (Eds.), *Hearing by eye: The psychology of lip-reading* (pp. 97–114). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Robert-Ribes, J., Schwartz, J. L., Lallouache, T., & Escudier, P. (1998). Complementarity and synergy in bimodal speech: Auditory, visual, and audio–visual identification of French oral vowels in noise. *The Journal of the Acoustical Society of America, 103*(6), 3677–3689. https://doi.org/10.1121/1.423069

Rosin, M. M., Swift, E., Bless, D., & Vetter, D. K. (1988). Communication profiles of adolescents with Down syndrome. *Journal of Childhood Communication Disorders, 12*(1), 49–64. https://doi.org/10.1177/152574018801200105

Rupela, V., Velleman, S. L., & Andrianopoulos, M. V. (2016). Motor speech skills in children with Down syndrome: A descriptive study. *International Journal of Speech-Language Pathology, 18*(5), 483–492. https://doi.org/10.3109/17549507.2015.1112836

Schwartz, J. L., Berthommier, F., & Savariaux, C. (2004). Seeing to hear better: Evidence for early audio–visual interactions in speech identification. *Cognition, 93*(2), B69–B78. https://doi.org/10.1016/j.cognition.2004.01.006

Smith, B. L., & Stoel-Gammon, C. (1983). A longitudinal study of the development of stop consonant production in normal and Down's syndrome children. *Journal of Speech and Hearing Disorders, 48*(2), 114–118.

Smithson, M., & Verkuilen, J. (2006). A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent

variables. *Psychological Methods, 11*(1), 54–71. https://doi.org/10.1037/1082-989X.11.1.54

Sommers, M. S., Tye-Murray, N., & Spehar, B. (2005). Auditory–visual speech perception and auditory–visual enhancement in normal-hearing younger and older adults. *Ear and Hearing, 26*(3), 263–275.

Sommers, R. K., Patterson, P., & Wildgen, P. L. (1988). Phonology of Down syndrome speakers, ages 13–22. *Journal of Childhood Communication Disorders, 12*(1), 65–91. https://doi.org/10.1177/152574018801200106

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *The Journal of the Acoustical Society of America, 26*(2), 212–215. https://doi.org/10.1121/1.1907309

Summerfield, Q. (1987). Some preliminaries to a comprehensive account of audio–visual speech perception. In B. Dodd & R. Campbell (Eds.), *Hearing by eye: The psychology of lip-reading* (pp. 3–51). Hilldale, NJ: Lawrence Erlbaum Associates, Inc. https://doi.org/10.2307/1423237

Timmins, C., Cleland, J., Wood, S. E., Hardcastle, W. J., & Wishart, J. G. (2009). A perceptual and electropalatographic study of /ʃ/ in young people with Down's syndrome. *Clinical Linguistics & Phonetics, 23*(12), 911–925. https://doi.org/10.3109/02699200903141271

Timmins, C., Hardcastle, W. J., Wood, S., & Cleland, J. (2011). An EPG analysis of /t/ in young people with Down's syndrome. *Clinical Linguistics & Phonetics, 25*(11–12), 1022–1027. https://doi.org/10.3109/02699206.2011.616981

Toğram, B. (2015). How do families of children with Down syndrome perceive speech intelligibility in Turkey? *BioMed Research International, 2015*, Article ID 707134. https://doi.org/10.1155/2015/707134

Xue, S. A., Kaine, L., & Ng, M. L. (2010). Quantification of vocal tract configuration of older children with Down syndrome: A pilot study. *International Journal of Pediatric Otorhinolaryngology, 74*(4), 378–383. https://doi.org/10.1016/j.ijporl.2010.01.007

Zeiliger, J., Serignat, J., Autessere, D., & Meunier, C. (1994). Bd_bruit, une base de données de parole de locuteurs soumis à du bruit [BDBRUIT, a database of speech produced by people in noisy environments]. *Proceedings of the 10th Journées d'Étude sur la Parole, Trégastel, France,* 287–290.

3.2 « Sensory-motor imitation benefits perceptual learning even for speech produced with an anatomical and motor disorder »

L'imitation sensori-motrice est bénéfique à l'apprentissage perceptif même pour la parole produite avec des troubles anatomiques et moteurs

L'étude présentée ici a été menée en collaboration avec Marion Dohen, Amélie Rochet-Capellan, Jean-Luc Schwartz et Silvain Gerber. Ce travail a été soumis à la revue Cognitive Science et est en cours de reprise pour être soumis à une autre revue.

Hennequin A., Rochet-Capellan, A., Schwartz J., Gerber, S., & Dohen, M. (soumis). Sensory-motor imitation benefits perceptual learning even for speech produced with an anatomical and motor disorder.

Cette étude s'articule autour de la notion d'apprentissage perceptif comme moyen permettant d'améliorer la communication (Section 2.2) et du rôle du système moteur dans ce processus (Section 2.4.3). Elle se base sur le fait que lorsqu'un auditeur perçoit la parole spécifique à un locuteur, il ajuste son système perceptif ce qui lui permet de mieux le percevoir par la suite. D'autre part, il apparaît que lorsque l'auditeur implique son système moteur dans cet apprentissage perceptif celui-ci est renforcé notamment via l'imitation des productions du locuteur. Les représentations motrices propres à l'auditeur s'aligneraient sur celles du locuteur facilitant ainsi la perception de ce dernier. Comme détaillé précédemment, la parole produite par un locuteur avec T21 est différente de celle d'un locuteur tout-venant, en raison des nombreuses spécificités anatomiques et physiologiques induites par cette pathologie (Section 2.5.2). Cette différence résulte toujours dans une réduction d'intelligibilité. L'apprentissage perceptif couplé à l'imitation pourrait représenter une méthode prometteuse pour améliorer la communication des personnes avec T21, notamment parce qu'elle repose sur une adaptation de l'auditeur. On peut cependant se poser la question de l'efficacité de cette méthode lorsqu'un auditeur tout-venant est confronté à la parole d'un locuteur avec T21. Les spécificités de ce dernier pourraient réduire l'impact des processus imitatifs mis en jeu lors de l'apprentissage. Cette étude vise donc à déterminer si l'apprentissage perceptif par un auditeur tout-venant est envisageable lorsqu'il est confronté à la parole d'un locuteur avec T21, et si l'imitation de ce dernier le renforce.

48 personnes tout-venant locuteurs natifs du français ont participé à une étude d'apprentissage perceptif. Ils/elles n'avaient aucune ou peu d'expérience *a priori* avec la parole

produite par un locuteur avec T21. Les stimuli présentés dans cette expérience étaient des mots isolés produits par une locutrice avec T21. Chaque stimulus était présenté une seule et unique fois lors de l'expérience. Le paradigme expérimental était constitué de trois parties : Pré-Test, Entraînement et Post-Test. Les phases de Pré- et Post-test étaient les mêmes pour tous les participants. Pour l'Entraînement, ces derniers étaient assignés à un groupe parmi trois possibles (Contrôle – C, Écoute – E, Imitation – I). Le groupe C n'était pas entraîné et ne passait donc que les deux Tests. Pendant la phase d'Entraînement, le groupe E écoutait des stimuli (mots isolés) sans bruit, puis lisaient le label correspondant au stimulus perçu. Enfin, le groupe I écoutait les mêmes stimuli que le groupe E, mais devaient imiter ceux-ci avant de voir le label correspondant. Les deux Tests consistaient en une tâche de discrimination de stimuli dans le bruit. Les participants devaient choisir parmi quatre propositions de mots celle qui correspondait le mieux à ce qu'ils/elles venaient d'entendre. En fonction de l'exactitude de leurs réponses, le rapport signal sur bruit (RSB) était réajusté toutes les 5 réponses. Si le participant donnait un nombre de bonnes réponses au dessus d'un certain seuil, alors le niveau de bruit était augmenté ce qui rendait la tâche plus difficile. À l'inverse, donner un nombre de bonnes réponses inférieur à ce seuil diminuait le niveau de bruit rendant la tâche plus facile. La mesure de la performance d'un participant correspondait au niveau de bruit moyen lors d'un Test auquel on soustrayait le niveau de bruit au début du Test. Cette procédure a mené à deux mesures pour chaque participant : une pour le Pré- et une pour le Post-test. Celles-ci ont été moyennées par groupe pour étudier les différences liées au type d'Entraînement. Ces résultats ont été analysés statistiquement en ayant formulé deux hypothèses :

- **(a)** Puisqu'aucun participant n'a d'expérience préalable avec le locuteur, et que le paradigme des Tests ajuste leur difficulté à chaque participant, tous les groupes devraient avoir une performance similaire lors du Pré-Test ;
- **(b)** Dans le cas où l'imitation favoriserait le processus d'apprentissage perceptif, les différences entre groupes devraient s'observer lors du Post-Test, notamment entre les groupes C et I.

L'analyse des résultats confirme l'hypothèse **(a)**. Bien qu'il existe une variabilité (attendue) en liée au niveau de l'ajustement du RSB d'un participant à l'autre, les groupes de participants présentent tous le même score de performance à l'issue du Pré-Test. Ce n'est pas le cas à l'issue du Post-Test. Après la phase d'apprentissage perceptif, seulement le groupe I a eu un

score significativement différent du groupe C, ce qui confirme l'hypothèse **(b)**. Il est cependant intéressant de noter que tous les groupes obtiennent une performance indiquant un potentiel apprentissage passif – même pour le groupe C.

Cette étude a permis de montrer que lorsque des participants tout-venant étaient exposés à la parole d'une locutrice avec T21, une tâche d'imitation de cette parole a favorisé le processus d'apprentissage perceptif bénéficiant en retour à une meilleure perception de cette parole. La parole T21 est peu intelligible et la rééducation orthophonique permettant de l'améliorer complexe et ayant des limites. Des méthodes dépendant des interlocuteurs, telles que l'apprentissage perceptif, constituent des pistes additionnelles pour améliorer la communication des personnes avec T21.

Sensory-motor imitation benefits perceptual learning even for speech produced with an anatomical and motor disorder

Alexandre Hennequin [1],Amélie Rochet-Capellan [1],Jean-Luc Schwartz [1],Silvain Gerber [1],Marion Dohen [1]

[1] Univ. Grenoble Alpes, CNRS, Grenoble INP*, GIPSA-lab,
38000 Grenoble, France
* Institute of Engineering Univ. Grenoble Alpes

To improve speech perception, listeners consistently learn from their interlocutor. Perceptual learning refers to the improvement of perception performances over time and after exposure. This process has been shown to be strengthened by overtly imitating the speaker. This advantage would result from a better access to the speaker's motor representations favoring predictions on future productions. The present study investigated whether imitation could help typical listeners adapt to atypical speech produced by a speaker with trisomy 21 (T21). This genetic disorder indeed induces a number of anatomical and motor specificities resulting in intelligibility issues. Such specificities could make it more difficult for typical speakers to imitate the speech produced by people with T21, which could result in the lack of a benefit of imitation. 48 participants performed an identification task on words produced by a speaker with T21, embedded in speech shaped noise. After a pre-test, they were then assigned to one of three groups. Participants in the control group were not trained while those of the other two groups were trained by listening to clear isolated words with ensuing written feedback. In one of these two groups, participants were additionally instructed to imitate the speaker's productions. Post-test performances showed that only the group trained via imitation significantly better identified the words uttered by the speaker with T21 than the control group. This shows that imitation crucially reinforces listeners' adaptation even to a speaker with a different morphology and using different production strategies and thus more difficult to imitate.

Introduction

We perceive someone else's speech better when we get used to it. In order to improve communication, listeners indeed constantly adapt to newly incoming speech. This adaptation may capitalize on imitative processes that can induce modifications in the way interlocutors both perceive and produce speech. Overt imitation of the speaker has been found to reinforce perceptual learning and thus comprehension in listeners (Adank, Hagoort, & Bekkering, 2010). The present study adresses whether imitation processes also serve perceptual learning when listeners are exposed to the speech produced by a speaker with a number of anatomical and physiological specificities related to trisomy 21.

Perceptual learning consists in re-adjusting perceptual processes to better grasp our environment by using information perceived from it (Goldstone, 1998). Concerning speech, perceptual learning is well documented in the litterature (for a review, see Samuel & Kraljic, 2009). Through perceptual learning, listeners increase their performances when perceiving the speech they learned from. This effect has been studied in different types of speech. Its robustness improves the learning of new non-native phonemic contrasts (Jamieson & Morosan, 1986; Logan, Lively, & Pisoni, 1991). Perceptual learning promotes adaptation to accented speech (Adank et al., 2010; Maye, Aslin, & Tanenhaus, 2008), artificially distorted speech such as time compressed speech (Sebastián-Gallés, Dupoux, Costa, & Mehler, 2000), synthetic speech (Greenspan, Nusbaum, & Pisoni, 1988; Schwab, Nusbaum, & Pisoni, 1985), noise-vocoded speech (Davis, Johnsrude, Hervais-Adelman, Taylor, & McGettigan, 2005; Hervais-Adelman, Davis, Johnsrude, & Carlyon, 2008; Hervais-Adelman, Davis, Johnsrude, Taylor, & Carlyon, 2011) and neurologically degraded speech such as dysarthric speech (for a review, see Borrie, McAuliffe, & Liss, 2012). The studies cited above also investigate the effect of factors such as presence of feedback, natural vs. synthetic speech, differences between languages, training material, learning conditions, etc.

Moreover, some studies have shown that benefits in perception resulting from perceptual learning sometimes transfer to production: Bradlow, Pisoni, Akahane-Yamada and Tohkura (1997) found that participants' perceptual learning of a new non-native phonemic contrast resulted not only in an improvement of performances in perceptual categorization of the contrast but also of its production. This observation is in line with the large set of experimental evidence for bidirectional links between perception and production of speech (for a review, Skipper, Devlin, & Lametti, 2017), and in agreement with the perceptuo-motor theoretical frameworks developed accordingly (e.g. Galantucci, Fowler, & Turvey, 2006; Schwartz, Basirat, Ménard, & Sato, 2012). The motor

system has been shown to be involved in speech perception at least in some conditions (for a review, Skipper, Devlin, & Lametti, 2017). While gathering knowledge from the speaker, listeners also adapt their own motor representations in the light of the speech they encounter. Importantly, this could also result in improving the perceptual process in return (Pickering & Garrod, 2013), in line with the experimental evidence that stimulating the motor system modulates the perceptual responses to speech inputs (e.g. D'Ausilio et al., 2009; Sato et al., 2011).

Considering this perceptuo-motor link in speech perception, Adank and colleagues (2010) proposed to explicitly sollicit the motor system by asking participants to imitate a speaker's speech to reinforce perceptual learning. Participants were exposed to spoken sentences produced with a novel unfamiliar Dutch accent by one speaker. A pre- and a post-test assessed the participants' perceptual performances in noise. In between those tests, some participants were trained on clear sentences (without noise) produced by the same speaker. Depending on the group they were assigned to, they were asked to perform different tasks: just listen, repeat, transcribe the sentence, imitate, and imitate without being able to hear their own production (they heard noise in headphones). Results show that the only groups who improved their performances from pre- to post-test significantly more than the ones who were not trained at all were those who were instructed to imitate. Only training via overt imitation of the speaker improved perception of the speaker in degraded conditions. Similar results were found by Borrie and Schäfer (2015) using a similar experimental design with sentences uterred by a dysarthric speaker. The tasks they compared were: just listen, listen then written feedback (transcription of the sentence), listen then imitate, listen then imitate then written feedback. Compared to a group without training, the largest increase in performance was observed for the participants who imitated the speaker and received written feedback. Perceptual learning benefits are thus, at least to a certain extent, positively modulated by overt imitation.

To account for this effect of overt imitation, one could put forward that it would simply result from an additional involvement of the motor system, which could be hypothesized to simply act as an attention enhancer or to provide an additional sensori-motor trace to promote memorization. Just repeating the stimuli indeed does not benefit perceptual abilities as much, if at all, as imitating it (Adank et al., 2010) and sometimes even impedes it (Baese-Berk & Samuel, 2016). What would there be so specific about imitation? The reasoning suggested by Adank et al. (2010) is that when imitating the speaker, it is possible that the listener actually infers the speaker's internal motor representations. Based on these inferences, the listeners could then make predictions on the future productions of the speaker based on their own motor knowledge of speech. This could then help them better comprehend the speech they subsequently perceive.

As a listener-oriented learning mechanism shown to improve perception of different kinds of less intelligible speech, perceptual learning could be a relevant tool to promote perception of the speech produced by speakers with trisomy T21 (T21). The speech of people with T21 is indeed impaired resulting in poor auditory intelligibility compared to typical speakers (Hennequin, Rochet-Capellan, Gerber, & Dohen, 2018). T21 is a genetic disease resulting in anatomical and physiological specificities, which impede spoken expressive skills (for a review, see Kent & Vorperian, 2013). People with T21 have been shown to have irregular speech articulators, a limited respiratory system, a smaller oral cavity resulting in a relative macroglossia, etc. All together, these characteristics limit the speech production capabilities of people with T21 and account for differences with typical speakers since infancy. The first aim of the present study was to investigate whether perceptual learning also occurs when perceiving a speaker with T21. Further, given the anatomical, physiological and motor specificities related to T21, one could hypothesize that a typical speaker could have difficulties imitating the speech of a speaker with T21. If the benefits of imitation over auditory perception alone are related to the building of a stronger motor representation of the speaker using the listener's own motor system, one could hypothesize that the latter could have more trouble reinforcing the former in the case in which the motor system of the speaker perceived is different. This could result in a lack of benefit of imitation over auditory perception alone when perceiving the speech produced by a speaker with T21. The present study will explore this possibility.

Method

The present study uses a pre-test/training/post-test experimental paradigm in a between-subject design. Four different sets of stimuli were built, each for a precise phase of the experimental procedure defined further (Initialization, Pre-test, Training and Post-test). Participants were assigned to one of three experimental conditions: Control, Listening and Imitation.

Participants

Fourty-eight participants took part in the experiment. They were all native speakers of French, and reported no history of oral or written language impairment. All confirmed having little or no experience with the speech of people with T21, read and signed a consent form and received a 15€ gift card after participating. They were randomly assigned to one of the three experimental conditions. Each group consisted of 13 females and 3 males. (Control: 23.9 years old (mean) ± 6.16 (std) – Listening: 25.1 ± 5.33 – Imitation: 23.9 ± 4.2). Participants all underwent a bilateral hearing test prior to the experiment showing that their detection thresholds of pure tones at 0.5, 1, 2 and 4kHz were typical. This research was approved by the ethics committee of Grenoble Alpes University (IRB00010290 COMUE Grenoble Alpes University IRB#1 – approval number: 2014-03-11-41).

Experimental material

The linguistic material used in this study consisted of 4 lists of single French words. Each list was associated to one experimental phase (Initialization, Pre-test, Training and Post-test). The words were selected from the Lexique database (New, Pallier, Ferrand, & Matos, 2001) based on the following procedure:
Extraction of French nouns with 1 to 3 syllables and a frequency of appearance of at least 5 every million words (database consisting of books and movie subtitles).
Among these words, only words being part of at least one minimal pair were kept. Pairs corresponding to insertion or deletion of one phoneme were also considered even if they are not strictly speaking minimal pairs. All possible phonological contrasts were considered except those involving semi-vowels (/j/, /ɥ/ and /w/) and schwa (/ə/) which most often result in complex minimal pairs (e.g., diphtongs and elisions). The contrasts many speakers of French do not perceive as phonologically relevant were also ignored (i.e. /ɔ/ was considered as equivalent to /o/, /ɛ/ to /e/ and /œ/ to /ø/, Coquillon, 2007), as well as the contrast between rounded and unrounded nasal vowels /œ̃/ vs. /ɛ̃/. This resulted in a list of 1961 words.
Target words had to be "simple" enough to be "very likely" to be prompted with a picture and/or known by "everyone". This probability was judged on a four-point scale by the first and the last author (1 corresponding to 'very likely to be known by everyone' and 4 'very unlikely'). The two judges disagreed on 369 words. The second author then selected between one of the two grades. Only the words with a final grade of 1 were kept, resulting in a list of 651 words.

Contrast group	Representative contrasts
Place (n=13)	b/d, b/g, d/g, p/t, p/k, t/k, z/ʒ, s/ʃ, f/s, f/ʃ, v/ʒ, v/z, m/n
Voicing (n=6)	b/p, d/t, g/k, f/v, s/z, ʃ/ʒ
Manner (n=8)	p/f, t/s, t/ʃ, b/v, d/z, d/ʒ, b/m, d/n
Liquids (n=1)	l/ʁ
Insertion/Deletion (n=4)	Cluster vs. isolated consonant in initial vs. final position
Height (n=8)	u/o, a/o, ɑ̃/ɔ̃, i/e, e/a, y/ø, a/ø, ɑ̃/ɛ̃
Fronting (n=2)	y/u, ø/o
Rounding (n=2)	y/i, ø/e
Nasality (n=3)	ɑ̃/a, ɛ̃/e, o/ɔ̃
Insertion/Deletion (n=4)	Presence vs. absence of a vowel in initial vs. final position

Table 1. List of all the phonological contrasts used as contrastive for the selection of the minimal pairs. Each line corresponds to a grouping of several contrasts based on distinctive features. The contrast v/z is greyed because not enough words were found to represent it.

Initialization, Pre-test and Post-test lists were built using words with at least 3 minimal pairs (n = 280) illustrating three different contrasts among the ones displayed in Table 1. For each word, 3 alternative responses were randomly chosen from the minimal pairs. Out of the 51 possible contrasts displayed in Table 1, one (/v/vs./z/) was not represented frequently enough to satisfy the above criterion and was thus discarded resulting in a total of 50 contrasts. For the Pre- and Post-test lists, all contrasts had to appear twice: once in each direction (e.g. /b/ -> /p/ and /p/ -> /b/), leading to two lists of 100 words each. The Initialization list consisted of 18 of the remaining words not chosen for the test lists.
The Training list was then created using the 239 words that had one and only one minimal pair. This criterion was used so that, for the word heard to be identified as a real French word, only one alternative was possible. To limit the potential bias toward training certain phonemes more frequently than others, the frequencies of occurrence of all phomemes had to be as balanced as possible. To satisfy this goal at best, we designed an optimization algorithm that created 140 lists of sizes from 100 to 239 with the smallest possible difference between the most and least frequent phonemes. For a given list size, the algorithm first generated a random list from the 239 words. One of the words with the most frequently represented phoneme in the list was then replaced with a word not from the list and pertaining one of the least frequent phonemes. This was done until the difference between the frequency of occurences of the most frequent phonemes and that of the least frequent ones did not evolve from an iteration to another. Out of the 139 resulting lists, we selected the one with the smallest difference between the most frequently occuring phoneme and the least frequently occuring one. This list consisted of 149 words in which the most frequently occuring phonemes (/ʁ/, /a/, /e/, /i/, /l/, /o/, /t/) appeared 29 times and the least frequently occuring ones (/ʃ/, /ʒ/, /f/, /g/, /n/, /v/, /y/, /z/) 17 times.
The words from the four lists were recorded in a sound-proof room to become stimuli for the perception test. The speaker was an 18 year-old native female with T21 who signed a consent form authorizing auditory recording of her speech. She was able to read but pictures were presented together with the written word when possible to support reading. This was done on a 24" computer screen approximately 80 cm from the speaker. Her speech was recorded using a microphone (Prodipe STC3D) placed 30 cm from her mouth at a 44.1 kHz sampling rate (Focusrite Scarlett 6i6 soundcard). The resulting recordings were resampled at 16 kHz.
A 0.5s silence was added before and after each word. A speech-shaped noise was generated using all the recorded stimuli: the acoustic signals of all the words were put one behind the other, the spectrum of the resulting acoustic signal was computed and used to filter a white noise. This noise was used in the Test and Initialization experimental phases (see section 2.3.1).

Procedure

Participants were randomly assigned to one of three experimental conditions: Control, Listening or Imitation. Fig. 1 illustrates and summarizes the experimental procedure. For the three conditions, the expriment started

with the Initialization phase followed by the Pre-test phase (see section 2.3.1 for details). After that, the participants from the Listening and Imitation groups underwent a Training phase (see section 2.3.2 for details) before the Post-test while the Control group performed the Post-test right-away.

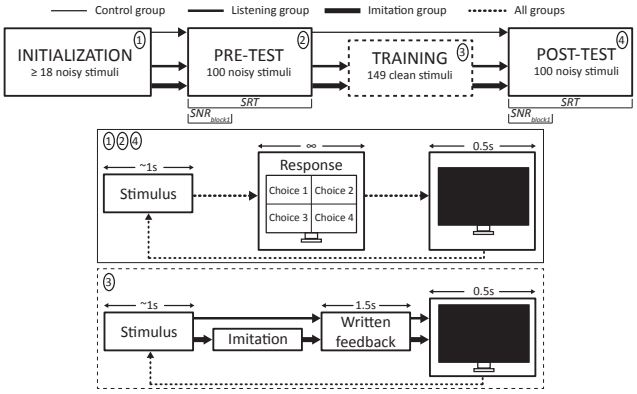


Figure 1. Description of the overall procedure and the different experimental phases: Initilization, Tests and Training.

1.1.1 Initialization & Test phases

Fig 1 illustrates these experimental phases. Participants were informed that they would hear stimuli consisting of a word mixed with noise. After each stimulus, they would have to select the closest written word to what they heard among 4 possibilities in a forced-choice task. Possible answers were presented on the screen and participants responded by clicking on the chosen word. There was no time limitation. A pause of 0.5 s with a blank screen followed their response. The next stimulus was then presented. Stimulus presentation order was randomized across participants.
For each stimulus, the four possible answers corresponded to the target word and three other words differing only by one phoneme from the target word (see section 2.2). The order of the words provided as choices were randomized across stimuli and participants.
The speech-shaped noise described above was added to all the stimuli from the Initialization and Test sets at various signal-to-noise ratios (SNR). The stimuli were presented in five-item blocks, in which the SNR was constant. At the end of each block, the SNR was updated using an up-down procedure aiming at maintaining a constant 60% correct score corresponding to 3 correct responses out of 5. If the participant responded correctly to 4 or 5 out of 5 stimuli, the SNR for the next block was decreased by one step. Conversely, if the participant scored less than 3 out of 5, the SNR for the next block was increased by one step. If the participant scored exactly 3 out of 5, the SNR remained constant for the next block. The step was 1 dB for the Test phases, and 2 dB for the Initialization phase. All stimuli were presented at a 70 dB sound pressure level (SPL).
The Initialization phase aimed at establishing the baseline SNR for each participant. Participants were shown 5-stimulus blocks using the Initialization set. For each participant, three 5-word groups were created randomly using the 18-word set (see section 2.2). At the beginning of the Initialization phase, the SNR was set at -6dB for all participants. It was adapted after each block using the procedure described above until the participant got a 60% correct score. If more than 3 blocks were required before reaching 60% (for 15% of the participants), 3 new 5-word blocks were created randomly using the Initialization set. At the end of the Pre-test, a new SNR could be computed which corresponded to the one to be used for a subsequent new block. This SNR value was used as the inital value for the Post-test.

1.1.2 Training phase

Fig. 1 illustrates this experimental phase. During the Training phase, participants listened to the stimuli without noise, in random order. Both the Listening and Imitation groups had to listen to the presented stimulus but the Listening group was instructed not to repeat it, while the Imitation group was asked to imitate the stimulus as accurately as possible. Both groups were then presented with a written version of the target word displayed at the middle of the screen for 1.5 s. They were told beforehand that this corresponded to the word that they had just heard. A 0.5 s blank screen was then displayed before presentation of the next stimulus.

Measure

The measure presented in this article is based on the Speech Reception Threshold (SRT) as described by Plomp and Mimpen (1979). The SRT corresponds to the mean SNR computed over all the blocks in the corresponding phase (see Fig. 1).
The up-down procedure used to adjust the SNR (see section 2.3.1) aims at maintaining a constant 60% correct score. The SRT is thus supposed to reflect the SNR corresponding to a 60% correct score. The problem is that the scores for each block can obviously not be constantly equal to 60%. In the present experiment, we obtained correct scores ranging from 46% to 72%. If calculated as the mean SNR, the SRT does therefore not represent a constant 60% correct score. We thus adapted the SRT measure to take this problem into account and obtain an accurate estimation of the SNR corresponding to a 60% correct score. This was done by plotting SNR values as a function of corresponding correct scores for all the blocks from each participant and Test phase (20 data points for each participant and test corresponding to the 20 5-stimulus blocks). A linear regression was then used to modelize the relationship between the two variables. Based on this, we extracted the SNR value which would truly correspond to 60% correctness (3/5 correct answers). See Fig. 2 for an illustration.

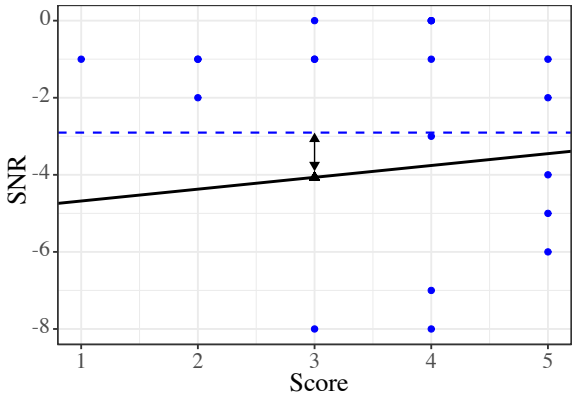


Figure 2. SNR values as a function of the number of correct responses for each block for one participant in one test (there are not 20 points because participants could obtain the same score in two different blocks resulting in overlapping points on the graph). Black full line: linear regression between the two variables. Blue dashed line: mean SNR over all the blocks (including the predicted value which would be used for a subsequent block). The corrected SRT value (triangle) corresponds to the y value of the linear regression for x=3. The arrow signals the difference between the 'traditional' SRT value and the corrected SRT.

Our aim was then to examine the evolution of performances during a single test. A differential measure was used for this and computed as:

$$Diff = SRT - SNR_{block\,1}$$

where SRT corresponds to the measure described above (see Fig. 2) and $SNR_{block\,1}$ corresponds to the SNR used for the 1st block of the test phase (see Fig. 1). This results in two values of $Diff$, one for the Pre-test and one for the Post-test, for each participant.

Since participants started off the Post-test with the same SNR value as that of the end of the Pre-test (see section 2.3.1), if training has an effect, the resulting measure should be different between the groups. The more the impact of training, the lower the value of the measure should be, suggesting that lower SNRs (more noise) were necessary to prevent the score from exceeding 60% correct.

Our predictions are that there should be no difference between groups for Pre-test. If training benefits performance, we expect lower values (see above) for the Listening and Imitation groups than for the Control Group. If imitation has an even more beneficial effect on performances (as in Adank, Hagoort, & Bekkering, 2010), we expect even lower values for the Imitation group than for the Listening group.

Results

Statistical analyses were used to compare the $Diff$ values between groups for each test phase. A linear regression analysis (*lm* function of package *stats* from R - version 3.4.2) was performed on this variable along with two factors: Phase (Pre-Test or Post-Test) and Group (Control or Listening or Imitation). Factors and their interaction were all significant in the final model. Satisfactory conditions of application (normality, homogeneity of variance) for the final model were graphically assessed. This model was used to perform corrected multiple comparisons (using *glht* function of package *multcomp*; Hothorn, Bretz, & Westfall, 2008) to compare the performances between groups of participants (factor Group) for each Test (factor Phase). Descriptive results of the $Diff$ measure as a function of these factors can be seen on Fig. 3.

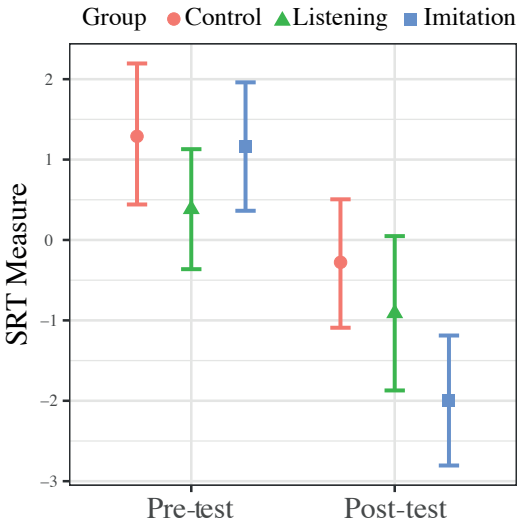


Figure 3. $Diff$ values as a function of test phase and experimental group. The star signals the only significant difference between groups (p=.015).

Comparison indicated that there was no significant difference between the groups in the Pre-Test (p-values > 0.34). Participants from all three groups (i.e. Control, Listening or Imitation) performed the same in Pre-test. Group performances ranged from 0.38 to 1.32 (in dB). Positive values correspond to an increase in SNR over the course of the Pre-test (see section 2.4) suggesting that the initial evaluation of the SNR in the Initialization phase was slightly underestimated (too much noise relative to signal): to obtain an exact level of 60% correct responses, all three groups tended toward greater SNRs over the course of the Pre-test than at its start .

In the Post-test, all $Diff$ values were negative indicating that the SNRs decreased (more noise relative to signal) over the course of the test. A decrease in SNR corresponds to correct scores exceeding 60% suggesting that the performances of all groups were better in the Post-Test than at the end of Pre-Test. Comparisons between groups on the Post-Test revealed that $Diff$ values were significantly smaller for the Imitation than for the Control group (-2dB vs. –0.29 dB, p = 0.015). The Listening group

was at an intermediate value of -0.91 dB, but no other comparison reached significance ($p > 0.24$).

Discussion

This study aimed to assess the effect of perceptual training and imitation on the perception of the speech of a speaker with trisomy 21 (T21) by typical listeners. Perception of isolated French words was tested in noise using a pre-test / post-test paradigm. Participants were separated into three groups. In the Control group, participants did not train in between pre- and post-test. In the Listening group, they were trained on a set of 149 words: they heard the word being uttered by the speaker with T21 (no noise) and then got written feedback of the word. In the Imitation group, they additionally were asked to imitate the production of the speaker as accurately as possible just after hearing it and before getting written feedback. The noise level (signal-to-noise ratio; SNR) during pre- and post-test was adapted every five stimuli using an up-down procedure aiming at maintaining a constant 60% correct response score in a forced four-level choice task. An adapted measure reflecting the evolution of the SNR during each test phase was used to compare the groups. There was no difference between groups at pre-test showing that groups were equivalent in terms of ability to perceive the speaker with T21 before training. At post-test, the measure was significantly lower for the Imitation than the Control group. No other group differences reached significance. This suggests that training, relative to none, had a positive effect on the perception of the speech produced by a speaker with T21 only when it was imitated during training. Note that, even though the Listening group also tended to perform better than the Control group, no significant difference was found between those groups.
The results of this study sustain the idea that the implication of the listener's motor system is relevant for perceptual learning of speech. They are in line with those of Adank and colleagues (2010) who found that imitation during training helped listeners to better perceive a speaker uttering sentences with an unfamiliar accent. In their study, the only significant differences between the control and experimental groups were found for the two groups in which participants were asked to imitate during training (not when they simply listened, repeated or transcribed). Borrie & Schäfer (2015) replicated this finding with a speaker with a motor speech disorder (dysarthria) and comparing no training to that with listening (with or without feedback) or with imitation (with or without feedback). Imitation would then be critical for perceptual training to be efficient.
Note that, in the Adank et al. (2010) and Borrie & Schäfer (2015) studies, the speakers did not have physiological and anatomical specificities of the vocal tract as compared to the typical listeners who trained perceiving their speech. The present study replicated the finding of the two latter studies but with a speaker with T21. This genetic disorder induces a number of physiological, anatomical and motor specificities of the vocal tract and speech organs. These specificities strongly impair their speech production capacities resulting in poor intelligibility of their speech (for a review, see Kent & Vorperian, 2013; Kumin, 1994). Apart from just being different from what listeners are used to, as in Adank et al. (2010) and Borrie & Schäfer (2015), the speech of people with T21 can be seen as a model of speech produced in a strongly different way from a motor point of view. One could then argue that typical listeners cannot truly imitate such speech or at least cannot infer the underlying motor representations of the speaker by doing so. The present study shows that this is not the case and that listeners do benefit from imitation to perceptually learn the speech of a speaker with T21. Note that typical people were also shown to benefit from seeing a speaker with T21 to better perceive her in a previous study involving two speakers with T21 different from the one in this study (Hennequin, Rochet-Capellan, Gerber, & Dohen, 2018). Benefiting from visual perception has been argued to result from the fact that seeing the speaker could help make inferences about the latter's motor representations (Rosenblum, Miller, & Sanchez, 2007; von Kriegstein et al., 2008). Put together, the results from our two studies suggest that it could be possible to build motor representations of a speaker even in the case in which they are strongly different from our own.
The findings from this study could also have important clinical implications. Although T21 is a genetic disease resulting in large inter-individual differences, overall, the population with T21 has speech difficulties resulting in reduced speech intelligibility. Increasing the intelligibility of speakers with T21 is challenging but crucial to improve their everyday life and social integration. Speech intervention mainly aims at training the speaker and helping her improve her intelligibility. Speaker-independent intervention could however also be used, in combination, to generally increase the intelligibility of a speaker with T21. Perceptual learning relies mostly on the listener and could be used alongside speech therapy to benefit intelligibility and interpersonal communication between a typical speaker and one with T21. The present study indeed shows that typical listeners are able to perceptually learn from the speech of one speaker with DS provided they are instructed to imitate the speaker during training. Other studies would be required to examine whether this also applies for other speakers or even whether perceptual learning of one speaker with T21 could generalize to other speakers with T21.

Acknowledgments

This research was funded by the European Research Council under the European 37 Community's Seventh Framework Program (FP7/2007-2013 Grant Agreement no.339152- "Speech Unit(e)s"). It was approved by the ethics committee of Grenoble-Alpes University (IRB00010290 COMUE Grenoble Alpes University IRB#1 – approval number: 2014-03-11-41). We thank the ARIST (Down Syndrome Research and Social Integration Association), the speaker who participated in this study and her family as well as all the participants who took part in the experiment.

References

Adank, P., Hagoort, P., & Bekkering, H. (2010). Imitation improves language comprehension. Psychological Science, 21(12), 1903–1909. https://doi.org/10.1177/0956797610389192

Borrie, S. A., McAuliffe, M. J., & Liss, J. M. (2012). Perceptual learning of dysarthric speech: a review of experimental studies. Journal of Speech, Language, and Hearing Research, 55(1), 290–305. https://doi.org/10.1044/1092-4388(2011/10-0349)

Borrie, S. A., & Schäfer, M. C. M. (2015). The Role of Somatosensory Information in Speech Perception: Imitation Improves Recognition of Disordered Speech. Journal of Speech, Language, and Hearing Research, 58, 1708–1716. https://doi.org/10.1044/2015

Bradlow, A. R., Pisoni, D. B., Akahane-Yamada, R., & Tohkura, Y. (1997). Training Japanese listeners to identify English /r/ and /l/: IV. Some effects of perceptual learning on speech production. The Journal of the Acoustical Society of America, 101(4), 2299–2310.

Coquillon, A. (2007). Le français parlé à Marseille : exemple d'un locuteur. Bulletin PFC (Phonologie Du Français Contemporain), 7, 145–156.

D'Ausilio, A., Pulvermüller, F., Salmas, P., Bufalari, I., Begliomini, C., & Fadiga, L. (2009). The Motor Somatotopy of Speech Perception. Current Biology, 19(5), 381–385. https://doi.org/10.1016/j.cub.2009.01.017

Davis, M. H., Johnsrude, I. S., Hervais-Adelman, A. G., Taylor, K. J., & McGettigan, C. (2005). Lexical information drives perceptual learning of distorted speech: evidence from the comprehension of noise-vocoded sentences. Journal of Experimental Psychology: General, 134(2), 222–241. https://doi.org/2005-04168-006 [pii]\n10.1037/0096-3445.134.2.222

Galantucci, B., Fowler, C. A., & Turvey, M. T. (2006). The motor theory of speech perception reviewed. Psychonomic Bulletin & Review, 13(3), 361–377. https://doi.org/10.1016/j.micinf.2011.07.011.Innate

Goldstone, R. L. (1998). Perceptual learning. Annual Review of Psychology, 49, 585–612.

Greenspan, S. L., Nusbaum, H. C., & Pisoni, D. B. (1988). Perceptual Learning of Synthetic Speech Produced by Rule. Journal of Experimental Psychology: Learning, Memory, and Cognition, 14(3), 421–433. https://doi.org/10.1037/0278-7393.14.3.421

Hennequin, A., Rochet-Capellan, A., Gerber, S., & Dohen, M. (2018). Does the Visual Channel Improve the Perception of Consonants Produced by Speakers of French With Down Syndrome? Journal of Speech Language and Hearing Research, 61(4), 957. https://doi.org/10.1044/2017_JSLHR-H-17-0112

Hervais-Adelman, A. G., Davis, M. H., Johnsrude, I. S., & Carlyon, R. P. (2008). Perceptual learning of noise vocoded words: Effects of feedback and lexicality. Journal of Experimental Psychology: Human Perception and Performance, 34(2), 460–474. https://doi.org/10.1037/0096-1523.34.2.460

Hervais-Adelman, A. G., Davis, M. H., Johnsrude, I. S., Taylor, K. J., & Carlyon, R. P. (2011). Generalization of Perceptual Learning of Vocoded Speech. Journal of Experimental Psychology: Human Perception and Performance, 37(1), 283–295. https://doi.org/10.1037/a0020772

Hothorn, T., Bretz, F., & Westfall, P. (2008). Simultaneous Inference in General Parametric Models. Biometrical Journal, 50(3), 346–363. https://doi.org/10.1002/bimj.200810425

Jamieson, D. G., & Morosan, D. E. (1986). Training non-native speech contrasts in adults : Acquisition of the English /ð/ - /θ/ contrast by francophones. Perception & Psychophysics, 40(4), 205–215. https://doi.org/10.3758/BF03211500

Kent, R. D., & Vorperian, H. K. (2013). Speech Impairment in Down Syndrome: A Review. Journal of Speech, Language & Hearing Research, 56(1), 178–210. https://doi.org/10.1044/1092-4388(2012/12-0148)

Kumin, L. (1994). Intelligibility of Speech in Children With Down Syndrome in Natural Settings: Parents' Perspective. Perceptual and Motor Skills, 78(1), 307–313. https://doi.org/10.2466/pms.1994.78.1.307

Logan, J. S., Lively, S. E., & Pisoni, D. B. (1991). Training Japanese listeners to identify English /r/ and /l/: A first report. The Journal of the Acoustical Society of America, 89(2), 874–886.

Maye, J., Aslin, R. N., & Tanenhaus, M. K. (2008). The weckud wetch of the wast: lexical adaptation to a novel accent. Cognitive Science, 32(3), 543–562. https://doi.org/10.1080/03640210802035357

New, B., Pallier, C., Ferrand, L., & Matos, R. (2001). Une base de données lexicales du français contemporain sur internet : LEXIQUE ™ // A lexical database for contemporary french : LEXIQUE ™. L'année Psychologique, 101(3), 447–462. https://doi.org/10.3406/psy.2001.1341

Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. Behavioral and Brain Sciences, 36(4), 329–347. https://doi.org/10.1017/S0140525X12001495

Plomp, R., & Mimpen, A. M. (1979). Improving the Reliability of Testing the Speech Reception Threshold for Sentences. Audiology, 18(1), 43–52.

Rosenblum, L. D., Miller, R. M., & Sanchez, K. (2007). Lip-read me now, hear me better later. Psychological Science, 18(5), 392-396.

Samuel, A. G., & Kraljic, T. (2009). Perceptual learning for speech. Attention, Perception, & Psychophysics, 71(6), 1207–1218. https://doi.org/10.3758/APP.71.6.1207

Sato, M., Grabski, K., Glenberg, A. M., Brisebois, A., Basirat, A., Ménard, L., & Cattaneo, L. (2011). Articulatory bias in speech

categorization: Evidence from use-induced motor plasticity. Cortex, 47(8), 1001–1003. https://doi.org/10.1016/j.cortex.2011.03.009

Schwab, E. C., Nusbaum, H. C., & Pisoni, D. B. (1985). Some Effects of Training on the Perception of Synthetic Speech. Human Factors, 27(4), 395–408. https://doi.org/10.1177/001872088502700404

Schwartz, J.-L., Basirat, A., Ménard, L., & Sato, M. (2012). The Perception-for-Action-Control Theory (PACT): A perceptuo-motor theory of speech perception. Journal of Neurolinguistics, 25(5), 336–354. https://doi.org/10.1016/j.jneuroling.2009.12.004

Sebastián-Gallés, N., Dupoux, E., Costa, A., & Mehler, J. (2000). Adaptation to time-compressed speech: phonological determinants. Perception & Psychophysics, 62(4), 834–842. https://doi.org/10.3758/BF03206926

Skipper, J. I., Devlin, J. T., & Lametti, D. R. (2017). The hearing ear is always found close to the speaking tongue: Review of the role of the motor system in speech perception. Brain and Language, 164, 77–105. https://doi.org/10.1016/j.bandl.2016.10.004

von Kriegstein, K., Dogan, Ö., Grüter, M., Giraud, A. L., Kell, C. A., Grüter, T., Kleinschmidt, A., & Kiebel, S. J. (2008). Simulation of talking faces in the human brain improves auditory speech recognition. Proceedings of the National Academy of Sciences. https://doi.org/10.1073/pnas.0710826105

4 Discussion générale

4.1 Résumé des principaux apports des travaux présentés

4.1.1 La perception de la parole, un processus agile et adaptable

Notre chapitre introductif a permis de rassembler plusieurs séries d'éléments bibliographiques qui posent le décor théorique de cette thèse : la nature agile et adaptable de la perception de la parole. En effet, il apparaît à travers les nombreux développements théoriques et expérimentaux de ces cinquante dernières années que la perception de la parole est un phénomène agile pouvant impliquer différentes modalités sensorielles (visuelle, auditive, tactile) mais semblant aussi solliciter le système moteur de la personne qui perçoit. Ainsi, dans des situations où un canal est perturbé, l'auditeur peut s'appuyer sur les autres canaux disponibles pour percevoir la parole. Ces compensations inter-modalités s'observent pour la perception de la parole de locuteurs tout-venant, en condition de bruit, ou quand il s'agit de percevoir un accent inhabituel, par exemple.

De plus, la perception de la parole est un phénomène adaptable, de par ses propriétés dynamiques : l'auditeur s'adapte à l'interlocuteur et peut apprendre à mieux le percevoir. Cette capacité à apprendre des spécificités de l'interlocuteur passe à la fois par des processus directs d'apprentissage statistique dans la modalité auditive, mais ils impliquent également des mécanismes de transfert intermodal dans lesquels une modalité peut aider à l'apprentissage dans une autre modalité (typiquement, de l'auditif au visuel). Et là encore, il semble que le système moteur puisse être impliqué dans l'apprentissage, à travers des mécanismes d'imitation ou de résonance qui permettent à l'auditeur de mieux « rentrer dans le corps orofacial » de son interlocuteur pour mieux en apprendre les spécificités de production.

Ces différents processus ont une importance théorique certaine, ils permettent de mieux comprendre comment fonctionne la perception de la parole. Ils posent notamment deux questions importantes en ce qui concerne la nature des

représentations qui la sous-tendent : (1) comment s'articule l'intermodalité en terme de représentations ? (2) quelle est la spécificité des représentations qui se construisent via l'expérience avec un locuteur donné ou un groupe de locuteurs donné ? Ce travail de thèse est né à l'intersection de ces deux questions théoriques. Mais, il avait aussi pour objectif d'évaluer la possibilité d'exploiter ces propriétés de la perception de la parole pour améliorer l'intelligibilité des personnes qui rencontrent des difficultés pour s'exprimer par la parole, comme c'est le cas des personnes avec une trisomie 21 (T21). Nous avons ainsi exploré ce terrain spécifique des troubles de la production de la parole associés à la T21, à la fois comme champ d'exploration théorique pour évaluer la résilience de ces propriétés d'agilité et d'adaptabilité pour des locuteurs très spécifiques que sont les locuteurs avec T21 ; mais aussi comme espace possible de remédiation des troubles de la communication parlée associés à la T21, en explorant comment les auditeurs tout-venant pouvaient optimiser leur perception des locuteurs avec T21, dans une perspective clinique et sociale d'amélioration de la communication.

4.1.2 Deux études sur la perception de la parole de locuteurs avec T21 par des personnes tout-venant

Nous avons réalisé deux études expérimentales visant, d'une part, à explorer si et jusqu'à quel point la modalité visuelle peut améliorer la perception de la parole de locuteurs avec T21 et, d'autre part, à évaluer si une familiarisation auditive à la parole de ces locuteurs peut aider à mieux les percevoir.

D'abord, dans (Hennequin, Rochet-Capellan, Gerber, & Dohen, 2018), nous avons pu montrer que, dans le cas de la perception de séquences voyelle-consonne-voyelle, l'entrée visuelle associée aux productions de locuteurs avec T21 est intelligible, et ce, de manière frappante, autant que celle associée à la parole typique. En conséquence, en situation de perception audiovisuelle de parole, l'auditeur bénéficie de l'apport de l'information visuelle de la même manière lorsqu'il perçoit un locuteur typique ou un locuteur avec T21, et ce malgré des spécificités motrices et anatomiques particulières chez ce dernier, décrites en détail dans la Section 2.5.2. Par contre, et conformément aux données de la

littérature, nous avons mis en avant que l'intelligibilité auditive des locuteurs avec T21 est, elle, significativement plus faible que celle de locuteurs tout-venant. Plus précisément, nous avons établi des profils d'erreurs au travers de deux groupes de quatre locuteurs avec T21 ou tout-venant. En analysant ces erreurs en détails, nous avons pu mettre en évidence notamment une difficulté accrue à convoyer le voisement pour les locuteurs avec T21. Pour la parole, la modalité visuelle ne peut pas combler ce déficit. Ce trait est en effet essentiellement auditif. Par contre, la modalité visuelle semble faciliter le traitement du lieu d'articulation des consonnes, et même, jusqu'à un certain point, de leur mode d'articulation. Les résultats de cette étude suggèrent à la fois des traitements orthophoniques localisés, orientés vers les contrastes les plus dégradés acoustiquement, mais aussi la prise en compte importante de la modalité visuelle dans la communication avec des locuteurs avec T21.

Ensuite, dans Hennequin et al. (soumis) nous avons observé que l'implication du système moteur de l'auditeur lors d'une phase d'apprentissage perceptif auditif de la parole d'un locuteur avec T21 améliore la perception des mots produits par ce locuteur. Ce résultat a été obtenu grâce à une tâche impliquant un processus conscient d'imitation, alors qu'une familiarisation sur un même corpus de productions par un locuteur avec T21 mais n'impliquant qu'une écoute passive sans processus d'imitation active n'a pas produit les mêmes gains d'intelligibilité. Notons que, indépendamment de ce processus spécifique de convergence perceptuo-motrice, cette étude a montré globalement des gains de performance dans la phase d'apprentissage quelle que soit la tâche associée pour les auditeurs. Ainsi, de même que le cercle proche des personnes avec T21 semble mieux les percevoir, les auditeurs naïfs de notre expérience ont pu ajuster leur système perceptif, de manière rapide, afin de mieux percevoir un locuteur présentant des spécificités anatomiques et physiologiques. Un tel accroissement de l'intelligibilité d'un locuteur avec T21 grâce à une exposition rapide est encourageant, puisqu'il repose uniquement sur l'auditeur et pourrait donc être facilement généralisé. Par ailleurs, ce résultat illustre la notion d'apprentissage perceptif, et renforce l'idée d'un rôle du système moteur dans ce processus

cognitif, stimulé ici par le processus d'imitation lors de l'apprentissage. Ce résultat sur l'implication du système moteur est particulièrement intéressant dans le contexte de la T21, caractérisée par les spécificités anatomiques et motrices du locuteur. Il montre ainsi la capacité du système moteur à prendre en compte et s'adapter à de larges différences entre auditeur et locuteur, différences qui auraient pourtant pu gêner significativement les mécanismes imitatifs mis en œuvre dans l'expérience.

Dans l'ensemble, nos résultats suggèrent donc que (1) la modalité visuelle est dans une certaine mesure préservée chez les personnes avec T21 ou au moins que l'interlocuteur est capable d'extraire une information pertinente de cette modalité et que (2) l'implication de la motricité de l'auditeur via l'imitation dans une phase de familiarisation auditive peut améliorer la perception de la parole des personnes avec T21. Ces résultats présentent un double intérêt. D'abord clinique, puisqu'ils offrent la perspective de méthodes d'amélioration de l'intelligibilité impliquant l'auditeur, et nous verrons que cette adaptation est un des enjeux d'un domaine de recherche clinique en plein essor : la communication augmentée et alternative. Ils permettent d'autre part d'apporter de nouveaux éléments en ce qui concernent la nature de la perception de la parole. Ces deux aspects sont discutés ici en prenant en considération certaines limites méthodologiques et en évoquant des perspectives pour des travaux futurs.

4.2 Perspectives cliniques : améliorer l'intelligibilité des personnes avec trisomie 21

4.2.1 Un enjeu sociétal important, à partager entre tout-venant et locuteurs avec T21

Parmi les motivations premières des études rappelées ci-dessus viennent les préoccupations quant à la T21. Bien que la T21 soit présente dans toutes les sociétés, sa réalité est largement méconnue du grand public, et faussée par de nombreux stéréotypes et préjugés. Cette méconnaissance n'est pas sans conséquences sur la prise en charge et l'intégration de cette population dans la

société. La T21, rappelons-le, est la première cause génétique de déficience intellectuelle, et affecte l'ensemble du phénotype de l'individu. Plus particulièrement, des différences anatomiques et physiologiques ainsi que des troubles auditifs et moteurs perturbent les capacités expressives des individus avec T21. En conséquence, l'individu avec T21 accuse des retards dans son développement cognitif, ainsi qu'un ensemble de troubles liés à la communication parlée, en perception et surtout en production. Le développement des compétences langagières s'en trouve fortement perturbé et il devient difficile d'évaluer si la déficience intellectuelle est la conséquence ou la cause du trouble langagier. Longtemps, et encore trop souvent, le trouble langagier est réduit à la déficience intellectuelle alors que les compétences en compréhension sont toujours meilleures qu'en production.

La parole des personnes avec T21 est décrite comme atypique et peu intelligible au travers de la littérature, avec cependant, de grandes variabilités interindividuelles. Améliorer l'intelligibilité des personnes avec T21 bonifierait à la fois leur quotidien et promouvrait leur insertion dans la société. Or nos études montrent globalement un résultat important : les potentialités d'amélioration de la communication avec des locuteurs avec T21 ne doivent pas être recherchées seulement chez ce dernier, mais aussi chez l'auditeur tout-venant. En ce qui concerne le locuteur avec T21, l'accès à un suivi et à un traitement orthophonique pour les individus dès le plus jeune âge et tout au long de la vie est bien entendu d'une aide précieuse pour en partie combler les déficits lors du développement. Nous avons pu notamment dans notre première étude montrer le problème spécifique du contrôle du voisement, qui doit être l'objet d'une attention particulière dans la rééducation. Mais nous avons vu également le large potentiel d'adaptabilité de la perception, fourni pas ses composantes multisensorielles et dynamiques, et qui peuvent également largement faciliter la communication et l'intégration sociale des locuteurs avec T21. Ce potentiel peut être intégré dans une perspective plus globale de communication augmentée, que nous allons aborder maintenant.

4.2.2 Améliorer et augmenter la communication avec les locuteurs avec T21

Parmi les pistes d'intervention visant à améliorer les capacités expressives des locuteurs avec T21, on peut mentionner les systèmes de Communication Augmentée et Alternative (CAA ; Cress & Marvin, 2003).

Bien qu'elle offre d'incroyables possibilités aux personnes ayant des difficultés pour parler, la CAA n'est pas toujours évidente à mettre en place. En effet, le développement d'un individu est modélisé selon un schéma comportant des étapes successives. Or, lorsque le développement est atypique, comme pour l'individu avec T21, les fonctions et compétences acquises ne suivent pas le schéma classique. Ainsi, les parents et parfois les praticiens ont une vision séquentielle du développement, et craignent des interférences lors de l'apprentissage. En particulier, l'introduction précoce de la CAA (du geste manuelle, des pictogrammes, etc) est souvent ralentie par l'idée selon laquelle elle empêcherait le développement de la parole ou la retarderait (Cress & Marvin, 2003). Or, les outils de CAA cherchent avant tout à permettre de réaliser l'intention de communiquer, par n'importe quelle méthode. Tout un éventail d'outils, à adapter auprès de chaque individu, permettrait à celui-ci de communiquer, même sommairement. Le développement des compétences nécessaires pour des tâches plus complexes ou altérées se ferait sur ces bases – à la manière des individus tout-venant. Plutôt que de chercher à adapter un individu à un comportement typique, les méthodes de CAA visent à fournir les outils permettant aux individus avec des besoins spécifiques de bénéficier d'une expérience communicative nécessaire au développement.

Souvent, la prise en charge de la parole des personnes avec T21 est centrée sur le locuteur et vise à améliorer son intelligibilité via l'amélioration de l'articulation des sons de parole. Cette dernière est évaluée via les compétences perceptives des interlocuteurs potentiels. Or, bien que la production de parole chez l'individu avec T21 soit perfectible, l'adaptation de l'interlocuteur est un aspect à ne pas négliger, d'autant moins qu'elle est reconnue comme un outil de CAA (Cress

& Marvin, 2003) et que l'auditeur tout-venant a un potentiel adaptatif souvent bien supérieur à celui de la personne en situation de handicap. Ainsi, il est important d'exploiter davantage le lien auditeur-locuteur. En effet, la communication est une activité duale : si la plupart des traitements orthophoniques concernent uniquement le locuteur, il est évident que les auditeurs peuvent également contribuer à l'amélioration de la communication en cherchant à développer leur acuité perceptive à la parole des personnes avec T21. C'est d'ailleurs déjà le cas dans la vie quotidienne d'un auditeur, avec n'importe quelle parole qu'il rencontre (dysarthrique, dégradée par le canal de communication ou par le bruit environnant, etc.). Les auditeurs ont à disposition un ensemble de recours permettant de mieux percevoir la parole, et ce malgré des conditions défavorables. Ceux-ci s'alignent directement avec les préoccupations de la CAA : il convient donc de les caractériser.

Notamment, la multimodalité de la parole, et plus particulièrement sa composante visuelle, fournit un outil précieux susceptible d'améliorer la communication. Le fait de voir son interlocuteur, en plus de l'entendre, est un moyen efficace et facile à mettre en œuvre pour mieux le percevoir. Lors de la production de parole, le locuteur produit des mouvements articulatoires orofaciaux. Cependant, ceux-ci ne sont pas systématiquement visibles (par exemple lors d'une conversation téléphonique), ou bien ils ne fournissent pas le bénéfice supplémentaire attendu en situation de perception de parole audiovisuelle (par exemple dans certains cas de dysarthrie profonde ou de paralysie faciale, voir Hustad, Dardis, & McCourt, 2007). Un enjeu important est donc de s'assurer que la parole visuelle du locuteur permet effectivement d'aider à la communication.

De même, l'apprentissage perceptif peut être considéré comme un mécanisme qui peut être intégré à la panoplie des outils de CAA. Il facilite l'alignement des représentations internes de l'auditeur avec celles du locuteur. En retour, cela mène à une meilleure communication. L'adaptation de l'auditeur au locuteur est un enjeu essentiel pour favoriser la communication entre deux individus, notamment si l'un d'eux produit une parole spécifique peu intelligible.

L'apprentissage perceptif est un processus cognitif développant la communication grâce à l'exposition à la parole d'un locuteur. On peut même envisager l'effet d'un cercle vertueux, puisque l'apprentissage perceptif d'un locuteur permet davantage de communication, qui réciproquement renforce cet apprentissage. Ce phénomène implique des processus imitatifs et de résonance motrice qui pourraient toutefois être gênés lorsque l'auditeur est confronté à une parole spécifique, comme celle produite par un locuteur avec T21. Les résultats de notre seconde étude sont très encourageants en ce qui concerne ce dernier point, montrant que la gêne potentielle due à l'écart entre locuteur et auditeur n'empêche pas les phénomènes d'apprentissage dans le cadre de la T21.

4.2.3 Limitations et perspectives

Bien évidemment, l'inscription potentielle de ces deux mécanismes, associés à la multisensorialité et à l'apprentissage, dans le cadre des méthodes de CAA applicables à la T21, doit être resituée par rapport aux limitations potentielles de nos études et aux questions portant sur la généralisation des résultats obtenus pour toute la population avec T21. Outre les limites spécifiques à chaque étude et discutées dans Hennequin et al. (2018) et Hennequin et al. (soumis), nous souhaitons souligner les points suivants.

D'abord, le nombre de locuteurs étudiés est faible : 4 locuteurs pour l'étude de perception multimodale et 1 locutrice pour l'étude d'apprentissage perceptif. Ce faible nombre d'individus ne peut bien évidemment pas prétendre représenter l'ensemble de la population avec une T21, en considérant en particulier la très grande variabilité des spécificités anatomiques et physiologiques entre individus et leur impact sur la parole produite par chacun. Les résultats obtenus dans les deux études sont des indicateurs positifs de l'action des deux outils de CAA proposés ici, mais élargir la cohorte de locuteurs étudiés permettra de tenter de généraliser nos conclusions à l'entièreté de la population avec T21, et éventuellement de mieux caractériser leurs limites selon les profils de locuteurs. Une étude du transfert d'apprentissage d'un locuteur avec T21 à un autre serait

particulièrement éclairante en ce qui concerne les spécificités inter-individuelles de la parole des personnes avec T21.

Un travail plus spécifique serait nécessaire concernant les modalités relatives à l'apprentissage perceptif. D'une part, l'apprentissage a une dimension temporelle : la consolidation des connaissances apprises et la capacité de rappel sont deux points importants lors de l'apprentissage (Maas, Robin, Freedman, Wulf, & Schmidt, 2008). L'étude d'apprentissage perceptif que nous avons menée a testé les effets bénéfiques immédiatement après une phase courte d'apprentissage. La phase d'apprentissage pourrait être prolongée et même répétée sur plusieurs séances ; de même, la phase de test après apprentissage pourrait être repoussée, de façon à déterminer quel est l'empan temporel des mécanismes de plasticité impliqués. D'autre part, comme nous l'avons vu dans la Section 2.3.1, certaines modalités d'apprentissage semblent renforcer les effets bénéfiques. C'est le cas par exemple lors de la présence d'information supplémentaire au signal acoustique d'apprentissage, comme l'information lexicale (Reinisch & Holt, 2014) ou la présence de la modalité visuelle (Pilling & Thomas, 2011). Notre première étude montre en effet que cette dernière permet de désambiguïser la parole des locuteurs avec T21 ce qui pourrait ainsi faciliter l'appréhension des mouvements articulatoires produits et éventuellement renforcer l'effet de l'imitation sur l'apprentissage perceptif. Ainsi, la bonne intelligibilité visuelle des productions de locuteurs T21, mise en évidence dans notre première étude, laisse à penser que la modalité visuelle pourrait fournir des gains d'apprentissage et de transfert visuo-auditif tout à fait intéressants, qu'il conviendra d'étudier par la suite. Nous y reviendrons dans la section suivante.

4.3 Perspectives théoriques : ce que la trisomie 21 nous dit sur la perception de la parole et les processus cognitifs d'interaction multisensorielle et perceptuo-motrice

L'étude de la parole des personnes avec T21 offre une perspective théorique intéressante. D'abord, parce que percevoir la parole d'un locuteur avec T21, peu intelligible, impose le recours à des stratégies de compensation chez l'auditeur qui

sont disponibles et probablement requises lors de toute communication mais dont les effets sont particulièrement importants ici, et qui du coup permettent de fournir un éclairage particulier sur ces processus. Ensuite, les différences entre locuteur tout-venant et locuteur avec T21 mettent au défi les représentations motrices requises lors de la perception de parole : au-delà de percevoir une parole atypique, on perçoit une parole produite différemment. La trisomie 21 ouvre ainsi des perspectives intéressantes et originales sur la nature même des processus miroirs dans la communication parlée.

4.3.1 Transferts d'apprentissage entre modalités sensorielles

Parmi les processus cognitifs améliorant la communication, nous avons montré d'une part les bénéfices de la multimodalité de la parole chez les locuteurs avec T21, et d'autre part les capacités d'ajustement de l'auditeur à un locuteur inhabituel. Transversalement à ces deux idées, on peut supposer que l'ajustement au locuteur serait plus marqué encore lors de la perception de celui-ci en modalité audiovisuelle plutôt que visuelle. Ceci conduit alors vers un effet récemment mis en évidence sous le nom de « dopage perceptif » (Moradi, Lidestam, Ning Ng, Danielsson, & Rönnberg, 2019), qui est défini comme le gain obtenu en perception auditive après avoir été exposé à la parole audiovisuelle d'un locuteur. Les auteurs de cette équipe ont montré depuis plusieurs années que le transfert d'apprentissage est en effet significativement plus élevé si la situation d'apprentissage est multimodale que si elle est monomodale, et ce dans des paradigmes et sur des critères de performances variés (catégorisation, perception dans le bruit, reconnaissance de phrases) (Lidestam et al., 2014 ; Moradi et al., 2019).

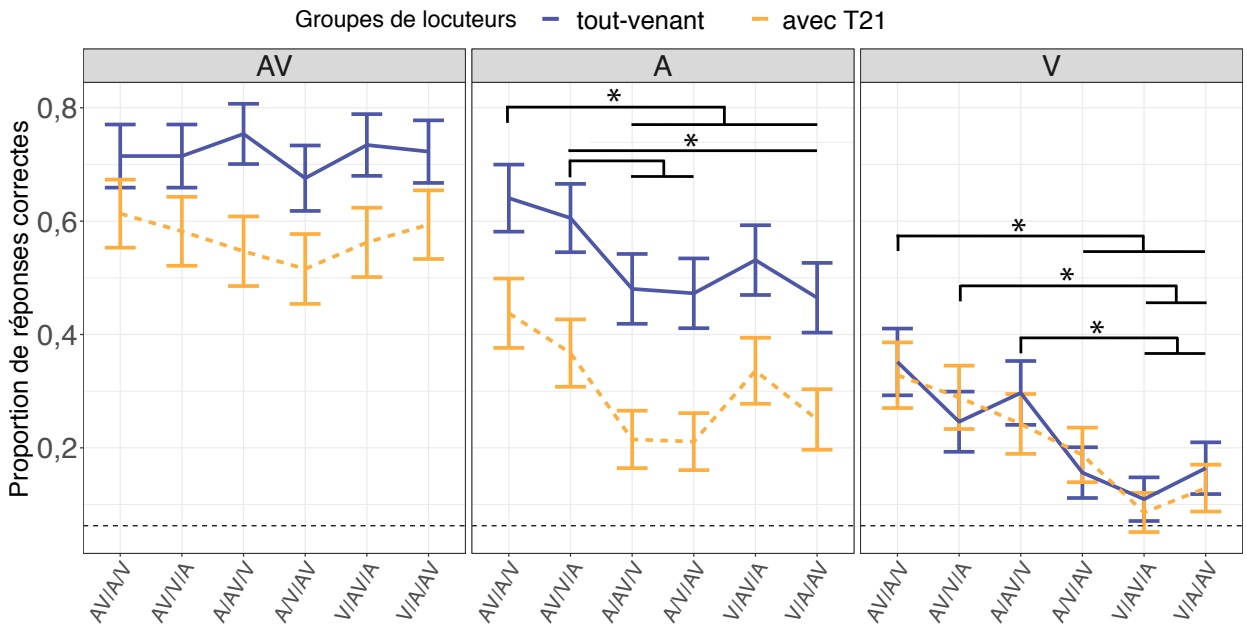


Figure 14 : Pourcentage de détection correcte des VCV en ordonnée en fonction de l'ordre de présentation en abscisse pour chaque modalité (*AV, A, V*) selon le groupe de locuteur (*tout-venant* ou *avec T21*). Nous concentrons notre attention sur les réponses en modalité A. La ligne en tirets représente le niveau de hasard pour les réponses.

Or, nous avons nous-mêmes obtenus des effets de cette nature dans notre première étude, bien qu'ils n'aient pas été analysés comme tels. En effet, dans la présentation des résultats sur la perception multimodale de la parole T21, nous n'avons pas exposé les effets d'ordre de présentation des modalités. Rappelons que les participants passaient le test perceptif dans les trois modalités, auditive (A), visuelle (V) et audiovisuelle (AV) successivement mais que l'ordre de passation était contrebalancé entre participants. Or, si on s'intéresse à l'effet de cet ordre de passation sur les performances des participants (voir Figure 14), on constate que les scores dans les modalités A et V sont nettement plus élevés lorsque celles-ci sont présentées après la modalité AV plutôt qu'avant. Cette observation est valable pour les deux types de stimuli, ceux issus des locuteurs tout-venant aussi bien que ceux des locuteurs avec T21. À l'inverse, la modalité AV ne bénéficie pas de l'apport d'une présentation unimodale préalable, et il ne semble pas non plus exister de transfert unimodal (de V vers A ou de A vers V) significatif dans ces données.

Si nous considérons le locuteur avec T21 comme représentant un modèle de locuteur atypique, on pourrait suggérer que les prédictions de l'auditeur concernant le locuteur avec T21, soutenues notamment par des mécanismes imitatifs, soient limitées lors de la perception de la parole (Pickering & Garrod,

2004). Les résultats de notre première étude montrent que la parole des personnes avec T21 est aussi intelligible visuellement que celle des locuteurs tout-venant, mais la seconde étude n'évalue l'apprentissage perceptif que dans une seule modalité (auditive). Or, nous savons que l'apprentissage de la parole d'un locuteur dans une modalité peut se transférer dans une autre (Rosenblum, Miller, & Kauyumari Sanchez, 2007 ; Sanchez, Dias, & Rosenblum, 2013 ; von Kriegstein et al., 2008). Pour un locuteur tout-venant, il semble donc que puissent s'opérer chez l'auditeur à la fois la construction d'un modèle du locuteur et l'interaction entre modalités. Cependant, les spécificités du locuteur avec T21 pourraient gêner ce phénomène notamment parce que l'auditeur pourrait avoir plus de mal à faire des inférences sur les stratégies motrices du locuteur à partir de ses propres connaissances motrices celles-ci étant trop différentes des siennes. La parole des locuteurs avec T21 offrirait donc un enjeu théorique important afin de mieux comprendre les mécanismes impliqués lors de l'apprentissage des connaissances relatives au locuteur.

Un autre enjeu important est celui du passage de l'apprentissage d'idiosyncrasies individuelles associées à des locuteurs avec T21 spécifiques, à la généralisation de ces connaissances vers une connaissance générale d'un possible ensemble de spécificités propres à la T21. En effet, notre seconde étude n'a porté sur la parole que d'un seul locuteur avec T21. Dans la section 2.2.2, nous avons mis en avant plusieurs principes susceptibles de favoriser des processus de généralisation (lexicale, phonétique) mais la généralisation d'un locuteur à l'autre ne semble pas être systématique et reste un enjeu important dans l'étude des accents étrangers (Reinisch & Holt, 2014). À l'instar de la dysarthrie, il serait nécessaire de mieux comprendre et caractériser ce qu'est la parole des locuteurs avec T21 et ce qu'elle a de spécifique avant de s'intéresser à la généralisation de l'apprentissage perceptif. Nous savons que les capacités d'expressions varient d'un individu avec T21 à l'autre, mais il reste probablement nécessaire de chercher des invariants capables de mieux représenter ce type de parole. Notre étude sur la perception multimodale, qui a permis de dresser les profils d'erreurs perceptives

sur 4 locuteurs, pourrait fournir une base intéressante en vue de mieux identifier certaines erreurs répétées d'un locuteur à l'autre.

4.3.2 Robustesse des mécanismes de résonance motrice

Le lien entre production et perception, ainsi que le rôle du système moteur, sont deux éléments majeurs pour comprendre les processus cognitifs étudiés dans cette thèse. On suppose que la perception audiovisuelle de la parole et l'apprentissage perceptif mettent en jeu les représentations motrices internes de l'auditeur, qui essaierait de les mettre en accord, en résonance, avec celles du locuteur. Cependant, la parole produite par un individu avec T21 est l'objet d'un ensemble de spécificités multiples, aussi bien anatomiques et physiologiques, que cognitives, comme nous l'avons vu dans la Section 2.5.2. Ces spécificités induisent de facto un écart cognitif avec celles d'un individu tout-venant, et on peut se demander jusqu'à quel point elles mettent en difficulté les processus de résonances et de phénomènes miroirs qui sont au cœur des relations perception-action. Elles pourraient donc, possiblement, perturber les deux processus cognitifs auxquels nous nous sommes intéressés dans cette thèse. Cependant, les deux études rapportées montrent que ces deux mécanismes ont pu jouer leur rôle et favoriser effectivement la communication. Sans que nous ayons pu tenter de préciser en détail les propriétés sous-jacentes de ces deux mécanismes et les composantes qui en assurent le fonctionnement dans cette situation perturbée, nous observons ainsi la robustesse de ces mécanismes dans le contexte d'un décalage physique et cognitif entre les systèmes moteurs du locuteur et de son auditeur.

Un aspect crucial dans l'étude de l'apprentissage perceptif est évidemment la place de l'imitation dans celui-ci. Malgré les décalages entre locuteur et auditeur que nous venons d'évoquer, notre seconde étude a montré que l'imitation a permis aux auditeurs tout-venant de mieux comprendre un locuteur avec T21. Pour rappel, les bons imitateurs (sur des critères acoustiques) sont ceux qui bénéficient le plus de l'apprentissage perceptif pour la parole dysarthrique (Borrie & Schäfer, 2015). Nous n'avons malheureusement pas eu le temps de procéder à une analyse

approfondie des performances d'imitation de la locutrice avec T21 par nos auditeurs tout-venant. Cette analyse reste à faire. Elle pourra passer, en accord avec plusieurs études de la littérature, par une évaluation acoustique des imitations, ou par un jugement perceptuel de celles-ci, pouvant englober plusieurs caractéristiques de la parole associées aux spécificités des productions de la locutrice avec T21. Ces analyses seront nécessaires afin de mieux appréhender la réussite de l'apprentissage perceptif en relation à la qualité de l'imitation. Il est à noter que lors de la perception d'action, les capacités imitatives d'un humain sont robustes et peuvent même s'appliquer à des acteurs non-humains, à la condition que les mouvements effectués soient dans le domaine du possible chez l'humain (Bisio et al., 2014). Les mouvements de parole produits par un locuteur avec T21, malgré ses spécificités anatomiques et physiologiques, seraient conformes et capables de déclencher des effets de résonance motrice chez l'auditeur.

Enfin, il faut noter que l'on peut alors tenter de séparer deux cibles du processus d'imitation : le résultat acoustique (imitation du but), ou bien les mouvements articulatoires qui l'ont produit (imitation des moyens). Concernant ces derniers, les spécificités du locuteur avec T21 devraient perturber – voire empêcher – une imitation par un locuteur tout-venant. Dans la mesure où un individu avec T21 partage un même type de spécificités physiques et cognitives avec un autre individu avec T21, on peut imaginer qu'il bénéficierait d'un meilleur apprentissage perceptif de sa parole, en comparaison à un individu tout-venant. Il en est de même pour toute pathologie présentant des caractéristiques atypiques, qui pourrait ainsi bénéficier de la même méthodologie que celle présentée dans cette thèse. On voit bien ainsi à quel point l'étude perceptuo-motrice des processus de communication entre interlocuteurs « différents » est potentiellement riche de leçons et de promesses pour l'étude générale des processus de la communication parlée.

4.4 Conclusion

La communication est une activité qui engage deux partenaires, le locuteur et l'auditeur. Elle repose sur des compétences qui se développent tout au long de la

vie de l'être humain, en perception et production de parole. Chaque acteur impliqué dans un échange communicatif met en œuvre diverses stratégies facilitant la transmission d'informations par la parole. Parmi celles-ci, cette thèse a centré son intérêt sur deux mécanismes particuliers, propres à l'auditeur : la perception multimodale et l'apprentissage perceptif de la parole. Ces deux processus cognitifs sollicitent les représentations motrices internes de l'auditeur, en regard de la parole produite par le locuteur. Les représentations de l'auditeur s'accordent avec celles du locuteur, menant de fait à une meilleure perception de celui-ci. Ces processus sont essentiels lorsque la communication est perturbée : par exemple, dans un milieu bruité ou lorsque la parole produite est dégradée – comme pour la T21. L'individu avec T21 manifeste un ensemble de spécificités qui perturbent ses capacités expressives. À ce titre, il est un candidat naturel pour bénéficier des stratégies de communication augmentatives et l'adaptation de l'interlocuteur en est une. Mais les spécificités de la production de la parole des personnes avec T21 pourraient perturber les mécanismes de résonance motrice sur lesquels cette stratégie peut s'appuyer. Or, notre travail suggère que deux voies d'adaptation de l'interlocuteur semblent relativement bien préservées pour améliorer l'intelligibilité de la parole des personnes avec T21 : la perception audiovisuelle et l'adaptation auditive de l'interlocuteur. Ainsi, nous montrons aussi que ces deux mécanismes, qui sont deux phénomènes très explorés en ce qui concerne la perception de la parole en général, et révélateurs de sa nature, sont robustes car efficaces même si l'auditeur est confronté à une parole atypique. En ce sens, ce travail fournit une base intéressante de réflexion sur deux enjeux théoriques majeurs qu'il couple à un enjeu clinique tout aussi majeur. D'abord, il nous a permis de mieux comprendre le fonctionnement des mécanismes de perception visuelle et d'adaptation en révélant qu'ils fonctionnent même quand la parole du locuteur est différente et en suggérant qu'il serait possible de mieux en tirer profit dans notre quotidien. D'autre part, il nous rappelle ce principe simple et essentiel : que la communication se déroule à deux, et qu'il est possible de construire ce succès à deux, en répartissant la « charge » adaptative sur les deux interlocuteurs et pas uniquement sur celui qui est le plus en difficulté.

Références bibliographiques

Abernethy, B., Zawi, K., & Jackson, R. C. (2008). Expertise and attunement to kinematic constraints. *Perception*, *37*(6), 931-948. https://doi.org/10.1068/p5340

Adank, P., Hagoort, P., & Bekkering, H. (2010). Imitation improves language comprehension. *Psychological Science*, *21*(12), 1903-1909. https://doi.org/10.1177/0956797610389192

Aglioti, S. M., Cesari, P., Romani, M., & Urgesi, C. (2008). Action anticipation and motor resonance in elite basketball players. *Nature Neuroscience*, *11*(9), 1109. https://doi.org/10.1038/nn.2182

Alcorn, S. (1932). The Tadoma Method. *The Volta Review*, *34*, 195-198.

Alm, M., & Behne, D. (2008). Voicing influences the saliency of place of articulation in audio-visual speech perception in babble. In *Proceedings of Interspeech 2008* (p. 2865-2868). Brisbane, Australia.

Alm, M., Behne, D. M., Wang, Y., & Eg, R. (2009). Audio-visual identification of place of articulation and voicing in white and babble noise. *The Journal of the Acoustical Society of America*, *126*(1), 377-387. https://doi.org/10.1121/1.3129508

Arnold, P., & Hill, F. (2001). Bisensory augmentation: A speechreading advantage when speech is clearly audible and intact. *British Journal of Psychology*, *92*(2), 339-355. https://doi.org/10.1348/000712601162220

Arumugam, A., Raja, K., Venugopalan, M., Chandrasekaran, B., Kovanur Sampath, K., Muthusamy, H., & Shanmugam, N. (2015). Down syndrome - A narrative review with a focus on anatomical features. *Clinical Anatomy*, *29*(5), 568-577. https://doi.org/10.1002/ca.22672

Babel, M. (2012). Evidence for phonetic and social selectivity in spontaneous phonetic imitation. *Journal of Phonetics*, *40*(1), 177-189. https://doi.org/10.1016/j.wocn.2011.09.001

Baese-Berk, M. M., & Samuel, A. G. (2016). Listeners beware: Speech production may be bad for learning speech sounds. *Journal of Memory and Language*, *89*, 23-36. https://doi.org/10.1016/j.jml.2015.10.008

Barnes, E., Roberts, J., Long, S. H., Martin, G. E., Berni, M. C., Mandulak, K. C., & Sideris, J. (2009). Phonological accuracy and intelligibility in connected speech of boys with fragile X syndrome or Down syndrome. *Journal of Speech, Language, and Hearing Research*, *52*(4), 1048-1061. https://doi.org/10.1044/1092-4388(2009/08-0001)

Béchet, M., Hirsch, F., Fauth, C., & Sock, R. (2012). Consonantal space area in children with a cleft palate: An acoustic study. In *Proceedings of Interspeech*

2012 (p. 58-61). Portland,OR, USA.

Benoit, C., Mohamadi, T., & Kandel, S. (1994). Effects of phonetic context on audio-visual intelligibility of French. *Journal of Speech, Language, and Hearing Research*, *37*(5), 1195-1203. https://doi.org/10.1044/jshr.3705.1195

Bent, T., & Bradlow, A. R. (2016). The interlanguage speech intelligibility benefit. *The Journal of the Acoustical Society of America*, *114*(3), 1600-1610. https://doi.org/10.1121/1.1603234

Bernstein, L. E., Auer, E. T., & Takayanagi, S. (2004). Auditory speech detection in noise enhanced by lipreading. *Speech Communication*, *44*(1-4), 5-18. https://doi.org/10.1016/j.specom.2004.10.011

Bernstein, L. E., Demorest, M. E., & Tucker, P. E. (2000). Speech perception without hearing. *Perception & Psychophysics*, *62*(2), 233-252. https://doi.org/10.3758/BF03205546

Bertelson, P., Vroomen, J., & De Gelder, B. (2003). Visual recalibration of auditory speech identification: A McGurk aftereffect. *Psychological Science*, *14*(6), 592-597. https://doi.org/10.1046/j.0956-7976.2003.psci_1470.x

Bicevskis, K., Derrick, D., & Gick, B. (2016). Visual-tactile integration in speech perception: Evidence for modality neutral speech primitives. *The Journal of the Acoustical Society of America*, *140*(5), 3531-3539. https://doi.org/10.1121/1.4965968

Binnie, C. A., Montgomery, A. A., & Jackson, P. L. (1974). Auditory and visual contributions to the perception of consonants. *Journal of Speech, Language, and Hearing Research*, *17*(4), 619-630. https://doi.org/10.1044/jshr.1704.619

Bisio, A., Sciutti, A., Nori, F., Metta, G., Fadiga, L., Sandini, G., & Pozzo, T. (2014). Motor contagion during human-human and human-robot interaction. *PLoS ONE*, *9*(8). https://doi.org/10.1371/journal.pone.0106172

Bittles, A. H., Bower, C., Hussain, R., & Glasson, E. J. (2007). The four ages of Down syndrome. *European Journal of Public Health*, *17*(2), 221-225. https://doi.org/10.1093/eurpub/ckl103

Blakemore, S.-J., & Decety, J. (2001). From the perception of action to the understanding of intention. *Nature Reviews Neuroscience*, *2*(8), 561-567. https://doi.org/10.1038/35086023

Blumstein, S. E., & Stevens, K. N. (1979). Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants. *The Journal of the Acoustical Society of America*, *66*(4), 1001-1017. https://doi.org/10.1121/1.383319

Borghi, R. W. (1990). Consonant phoneme, and distinctive feature error patterns in speech. In D. C. Van Dyke, D. J. Lang, F. Heide, S. van Duyne, & M. J. Soucek (Éd.), *Clinical Perspectives in the Management of Down Syndrome* (p. 147-152). New York, NY: Springer-Verlag New York Inc. https://doi.org/10.1007/978-1-

4613-9644-4_12

Borghini, G., & Hazan, V. (2018). Listening effort during sentence processing Is increased for non-native listeners: A pupillometry study. *Frontiers in Neuroscience*, *12*, 152. https://doi.org/10.3389/fnins.2018.00152

Borrie, S. A. (2015). Visual speech information: A help or hindrance in perceptual processing of dysarthric speech. *The Journal of the Acoustical Society of America*, *137*(3), 1473-1480. https://doi.org/10.1121/1.4913770

Borrie, S. A., McAuliffe, M. J., & Liss, J. M. (2012). Perceptual learning of dysarthric speech: a review of experimental studies. *Journal of speech, language, and hearing research : JSLHR*, *55*(1), 290-305. https://doi.org/10.1044/1092-4388(2011/10-0349)

Borrie, S. A., & Schäfer, M. C. M. (2015). The role of somatosensory information in speech perception: Imitation improves recognition of disordered speech. *Journal of Speech, Language, and Hearing Research*, *58*(6), 1708-1716. https://doi.org/10.1044/2015_JSLHR-S-15-0163

Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition*, *106*(2), 707-729. https://doi.org/10.1016/j.cognition.2007.04.005

Bradlow, A. R., Pisoni, D. B., Akahane-Yamada, R., & Tohkura, Y. (1997). Training Japanese listeners to identify English / r / and / l /: IV. Some effects of perceptual learning on speech production. *The Journal of the Acoustical Society of America*, *101*(4), 2299-2310. https://doi.org/10.1121/1.418276

Broersma, M., & Cutler, A. (2008). Phantom word activation in L2. *System*, *36*(1), 22-34. https://doi.org/10.1016/j.system.2007.11.003

Bruderer, A. G., Danielson, D. K., Kandhadai, P., & Werker, J. F. (2015). Sensorimotor influences on speech perception in infancy. *Psychological and Cognitive Sciences*, *112*(44), 13531-13536. https://doi.org/10.1073/pnas.1508631112

Bunton, K., & Leddy, M. (2011). An evaluation of articulatory working space area in vowel production of adults with Down syndrome. *Clinical Linguistics & Phonetics*, *25*(4), 321-334. https://doi.org/10.3109/02699206.2010.535647

Bunton, K., Leddy, M., & Miller, J. (2007). Phonetic intelligibility testing in adults with Down syndrome. *Down Syndrome Research and Practice*, *12*(1), 1-4. https://doi.org/10.3104/editorials.2034

Campbell, R., Burnham, D., Dodd, B. J., & Burnham, D. K. (1998). *Hearing Eye II : Advances in the psychology of speechreading and auditory-visual speech*. (R. Campbell, B. J. Dodd, & D. Burnham, Éd.). Hove, England: Psychology Press/erlbaum (UK) Taylor & Francis. https://doi.org/10.4324/9780203098752

Case, J., Seyfarth, S., & Levi, S. V. (2018). Does implicit voice learning improve spoken language processing? Implications for clinical practice. *Journal of Speech, Language, and Hearing Research*, *61*(5), 1251-1260.

https://doi.org/10.1044/2018_JSLHR-L-17-0298

Casey, A. F., & Emes, C. (2011). The effects of swim training on respiratory aspects of speech production in adolescents with down syndrome. *Adapted Physical Activity Quarterly*, *28*(4), 326-341. https://doi.org/10.1123/apaq.28.4.326

Centers for Disease Control and Prevention. (2006). Improved national prevalence estimates for 18 selected major birth defects--United States, 1999-2001. *MMWR. Morbidity and Mortality Weekly Report*, *54*(51).

Clarke-Davidson, C. M., Luce, P. A., & Sawusch, J. R. (2008). Does perceptual learning in speech reflect changes in phonetic category representation or decision bias? *Perception & Psychophysics*, *70*(4), 604-618. https://doi.org/10.3758/PP.70.4.604

Clarke, C. M., & Garrett, M. F. (2004). Rapid adaptation to foreign-accented English. *The Journal of the Acoustical Society of America*, *116*(6), 3647-3658. https://doi.org/10.1121/1.1815131

Cleland, J., Wood, S., Hardcastle, W., Wishart, J., & Timmins, C. (2010). Relationship between speech, oromotor, language and cognitive abilities in children with Down's syndrome. *Internation Journal of Language & Communication Disorders*, *45*(1), 83-95. https://doi.org/10.3109/13682820902745453

Connaghan, K. P., & Moore, C. A. (2013). Indirect estimates of jaw muscle tension in children with suspected hypertonia, children with suspected hypotonia, and matched controls. *Journal of Speech, Language, and Hearing Research*, *56*(1), 123-136. https://doi.org/10.1044/1092-4388(2012/11-0161)

Cooke, M., King, S., Garnier, M., & Aubanel, V. (2014). The listening talker: A review of human and algorithmic context-induced modifications of speech. *Computer Speech and Language*, *28*(2), 543-571. https://doi.org/10.1016/j.csl.2013.08.003

Creelman, C. D. (1957). Case of the unknown talker. *The Journal of the Acoustical Society of America*, *29*(5), 655. https://doi.org/10.1121/1.1909003

Cress, C. J., & Marvin, C. A. (2003). Common questions about AAC services in early intervention. *Augmentative and Alternative Communication*, *19*(4), 254-272. https://doi.org/10.1080/07434610310001598242

Crosley, P. A., & Dowling, S. (1989). The relationship between cluster and liquid simplification and sentence length, age, and IQ in Down's syndrome children. *Journal of Communication Disorders*, *22*(3), 151-168. https://doi.org/10.1016/0021-9924(89)90013-0

D'Ausilio, A., Bufalari, I., Salmas, P., & Fadiga, L. (2012). The role of the motor system in discriminating normal and degraded speech sounds. *Cortex*, *48*(7), 882-887. https://doi.org/10.1111/j.1439-0418.1989.tb00454.x

D'Ausilio, A., Pulvermüller, F., Salmas, P., Bufalari, I., Begliomini, C., & Fadiga, L. (2009). The motor somatotopy of speech perception. *Current Biology*, *19*(5),

381-385. https://doi.org/10.1016/j.cub.2009.01.017

D'Innocenzo, J., Tjaden, K., & Greenman, G. (2006). Intelligibility in dysarthria: Effects of listener familiarity and speaking condition. *Clinical Linguistics & Phonetics*, *20*(9), 659-675. https://doi.org/10.1080/02699200500224272

da Silva, V. Z. M., de França Barros, J., de Azevedo, M., de Godoy, J. R. P., Arena, R., & Cipriano, G. (2010). Bone mineral density and respiratory muscle strength in male individuals with mental retardation (with and without Down Syndrome). *Research in Developmental Disabilities*, *31*(6), 1585-1589. https://doi.org/10.1016/j.ridd.2010.05.003

Davis, C., & Kim, J. (2004). Audio–visual interactions with intact clearly audible speech. *The Quarterly Journal of Experimental Psychology Section A*, *57*(6), 1103-1121. https://doi.org/10.1080/02724980343000701

Davis, M. H., Johnsrude, I. S., Hervais-Adelman, A., Taylor, K., & McGettigan, C. (2005). Lexical information drives perceptual learning of distorted speech: Evidence from the comprehension of noise-vocoded sentences. *Journal of Experimental Psychology: General*, *134*(2), 222-241. https://doi.org/10.1037/0096-3445.134.2.222

de Graaf, G., Levine, S. P., Goldstein, R., & Skotko, B. G. (2019). Parents' perceptions of functional abilities in people with Down syndrome. *American Journal of Medical Genetics, Part A*, *179*(2), 161-176. https://doi.org/10.1002/ajmg.a.61004

DeCasper, A. J., & Spence, M. J. (1986). Prenatal maternal speech influences newborns' perception of speech sounds. *Infant Behavior and Development*, *9*(2), 133-150. https://doi.org/10.1016/0163-6383(86)90025-1

Diehl, R. L., & Kluender, K. R. (1989). On the objects of speech perception. *Ecological Psychology*, *1*(2), 121-144. https://doi.org/10.1207/s15326969eco0102_2

Diehl, R. L., Lotto, A. J., & Holt, L. L. (2004). Speech perception. *Annual Review of Psychology*, *55*(1), 149-179. https://doi.org/10.1146/annurev.psych.55.090902.142028

Dodd, B., & Campbell, R. (1988). *Hearing by eye: The psychology of lip-reading*. (B. Dodd & R. Campbell, Éd.), *The American Journal of Psychology* (Vol. 101). Lawrence Erlbaum Associates.

Eg, R., & Behne, D. (2009). Distorted visual information influences audiovisual perception of voicing. In *Proceedings of Interspeech 2009* (p. 2903-2906). Brighton, UK.

Eisner, F., & McQueen, J. M. (2005). The specificity of perceptual learning in speech processing. *Perception & Psychophysics*, *67*(2), 224-238. https://doi.org/10.3758/BF03206487

Erber, N. P. (1969). Interaction of audition and vision in the recognition of oral

speech stimuli. *Journal of Speech and Hearing Research*, *12*(2), 423-425. https://doi.org/10.1044/jshr.1202.423

Fougeron, C., Crevier-Buchman, L., Fredouille, C., Ghio, A., Meunier, C., Chevrie-Muller, C., … Vincent, C. (2010). Developing an acoustic-phonetic characterization of dysarthric speech in French. In *Proceedings of 7th International Conference on Language Ressources, Technologies and Evaluation (LREC)* (p. 2831-2838). Valetta, Malta.

Fowler, C. A. (2003). Rapid access to speech gestures in perception: Evidence from choice and simple response time tasks. *Journal of Memory and Language*, *49*(3), 396-413. https://doi.org/10.1016/S0749-596X(03)00072-X.

Francis, A. L., Baldwin, K., & Nusbaum, H. C. (2000). Effects of training on attention to acoustic cues. *Perception & Psychophysics*, *62*(8), 1668-1680. https://doi.org/10.3758/BF03212164

Francis, A. L., Nusbaum, H. C., & Fenn, K. (2007). Effects of Training on the Acoustic–Phonetic Representation of Synthetic Speech. *Journal of Speech Language and Hearing Research*, *50*(6), 1445. https://doi.org/10.1044/1092-4388(2007/100)

Galantucci, B., Fowler, C. A., & Turvey, M. T. (2006). The motor theory of speech perception reviewed. *Psychonomic Bulletin & Review*, *13*(3), 361-377. https://doi.org/10.3758/BF03193857

Gambi, C., & Pickering, M. J. (2013). Prediction and imitation in speech. *Frontiers in Psychology*, *4*, 340. https://doi.org/10.3389/fpsyg.2013.00340

Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance*, *6*(1), 110-125. https://doi.org/10.1037/0096-1523.6.1.110

Gick, B., & Derrick, D. (2009). Aero-tactile integration in speech perception. *Nature*, *462*(7272), 502-504. https://doi.org/10.1038/nature08572

Giles, H., Ogay, T., & Gallois, C. (2006). *Communication accomodation theory: A look back and a look ahead*. (W. B. Gudykunst, Éd.), *Theorizing about intercultural communication*. Thousand Oaks, CA, USA: Sage.

Goldstone, R. L. (1998). Perceptual learning. *Annual Review of Psychology*, *49*(1), 585-612. https://doi.org/10.1146/annurev.psych.49.1.585

Gonzales, K., Byers-Heinlein, K., & Lotto, A. J. (2019). How bilinguals perceive speech depends on which language they think they're hearing. *Cognition*, *182*, 318-330. https://doi.org/10.1016/j.cognition.2018.08.021

Grant, K. W., & Seitz, P.-F. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. *The Journal of the Acoustical Society of America*, *108*(3), 1197-1208. https://doi.org/10.1121/1.1288668

Greenspan, S. L., Nusbaum, H. C., & Pisoni, D. B. (1988). Perceptual learning of synthetic speech produced by rule. *Journal of Experimental Psychology:*

Learning, Memory, and Cognition, *14*(3), 421-433. https://doi.org/10.1037/0278-7393.14.3.421

Guimaraes, C. V. A., Donnelly, L. F., Shott, S. R., Amin, R. S., & Kalra, M. (2008). Relative rather than absolute macroglossia in patients with Down syndrome: implications for treatment of obstructive sleep apnea. *Pediatric Radiology*, *38*(10), 1062-1067. https://doi.org/10.1007/s00247-008-0941-7

Hashimoto, M., Igari, K., & Hanawa, S. (2014). Tongue Pressure During Swallowing in Adults with Down Syndrome and Its Relationship with Palatal Morphology. *Dysphagia*, *29*(4), 509-518. https://doi.org/10.1007/s00455-014-9538-5

Haute Autorité de Santé. (2015). *Les performances des tests de dépistage de la trisomie 21 foetale par analyse de l'ADN libre circulant*. *Rapport de recommandation en santé publique.*

Hazan, V., Sennema, A., Iba, M., & Faulkner, A. (2005). Effect of audiovisual perceptual training on the perception and production of consonants by Japanese learners of English. *Speech Communication*, *47*(3), 360-378. https://doi.org/10.1016/j.specom.2005.04.007

Hazan, V., Tuomainen, O., Kim, J., Davis, C., Sheffield, B., & Brungart, D. (2018). Clear speech adaptations in spontaneous speech produced by young and older adults. *The Journal of the Acoustical Society of America*, *144*(3), 1331-1346. https://doi.org/10.1121/1.5053218

Hennequin, A., Rochet-Capellan, A., Gerber, S., & Dohen, M. (2018). Does the Visual Channel Improve the Perception of Consonants Produced by Speakers of French With Down Syndrome? *Journal of Speech, Language, and Hearing Research*, *61*(4), 957-972. https://doi.org/10.1044/2017_JSLHR-H-17-0112

Hennequin, M., Allison, P. J., & Veyrune, J. L. (2000). Prevalence of oral health problems in a group of individuals with Down syndrome in France. *Developmental Medicine and Child Neurology*, *42*(10), 691-698. https://doi.org/10.1017/S0012162200001274

Hennequin, M., Faulks, D., Veyrune, J.-L., & Bourdiol, P. (1999). Significance of oral health in persons with Down syndrome: A literature review. *Developmental Medicine and Child Neurology*, *41*(4), 275-283.

Hervais-Adelman, A. G., Davis, M. H., Johnsrude, I. S., & Carlyon, R. P. (2008). Perceptual learning of noise vocoded words: Effects of feedback and lexicality. *Journal of Experimental Psychology: Human Perception and Performance*, *34*(2), 460-474. https://doi.org/10.1037/0096-1523.34.2.460

Hervais-Adelman, A. G., Davis, M. H., Johnsrude, I. S., Taylor, K. J., & Carlyon, R. P. (2011). Generalization of perceptual learning of vocoded speech. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(1), 283-295. https://doi.org/10.1037/a0020772

Holmes, E., Domingo, Y., & Johnsrude, I. S. (2018). Familiar voices are more intelligible, even if they are not recognized as familiar. *Psychological Science*,

29(10), 1575-1583. https://doi.org/10.1177/0956797618779083

Houde, J. F., & Jordan, M. I. (1998). Sensorimotor adaptation in speech production. *Science*, *279*(5354), 1213-1216. https://doi.org/10.1126/science.279.5354.1213

Hustad, K. C. (2007). Contribution of two sources of listener knowledge to intelligibility of speakers with cerebral palsy. *Journal of Speech, Language, and Hearing Research*, *50*(5), 1228-1240. https://doi.org/10.1044/1092-4388(2007/086)

Hustad, K. C. (2008). The relationship between listener comprehension and intelligibility scores for speakers with dysarthria. *Journal of Speech, Language, and Hearing Research*, *51*(3), 562-573. https://doi.org/10.1044/1092-4388(2008/040)

Hustad, K. C., & Cahill, M. A. (2003). Effects of presentation mode and repeated familiarization on intelligibility of dysarthric speech. *American Journal of Speech-Language Pathology*, *12*(2), 198-208. https://doi.org/10.1044/1058-0360(2003/066)

Hustad, K. C., Dardis, C. M., & McCourt, K. A. (2007). Effects of visual information on intelligibility of open and closed class words in predictable sentences produced by speakers with dysarthria. *Clinical Linguistics & Phonetics*, *21*(5), 353-367. https://doi.org/10.1080/02699200701259150

Inceoglu, S. (2016). Effects of perceptual training on second language vowel perception and production. *Applied Psycholinguistics*, *37*(5), 1175-1199. https://doi.org/10.1017/S0142716415000533

Ito, T., Tiede, M., & Ostry, D. J. (2009). Somatosensory function in speech perception. *Journal of Neurophysiology*, *107*(1), 442-447. https://doi.org/10.1152/jn.00029.2011

Jamieson, D. G., & Morosan, D. E. (1986). Training non-native speech contrasts in adults : Acquisition of the English /ð/ - /θ/ contrast by francophones. *Perception & Psychophysics*, *40*(4), 205-215. https://doi.org/10.3758/BF03211500

Jusczyk, P. W., Friederici, A. D., Wessels, J. M. I., Svenkerud, V. Y., & Jusczyk, A. M. (1993). Infants′ sensitivity to the sound patterns of native language words. *Journal of Memory and Language*, *32*(3), 402-420. https://doi.org/10.1006/jmla.1993.1022

Karmiloff-Smith, A., Al-Janabi, T., D'Souza, H., Groet, J., Massand, E., Mok, K., … Strydom, A. (2016). The importance of understanding individual differences in Down syndrome. *F1000Research*, *5*(F1000 Faculty Rev-389.). https://doi.org/10.12688/f1000research.7506.1

Katz, G., & Lazcano-Ponce, E. (2008). Intellectual disability: Definition, etiological factors, classification, diagnosis, treatment and prognosis. *Salud Pública de México*, *50*(2), s132-s141. https://doi.org/10.1590/S0036-

36342008000800005

Keintz, C. K., Bunton, K., & Hoit, J. D. (2007). Influence of visual information on the intelligibility of dysarthric speech. *American Journal of Speech-Language Pathology*, *16*(3), 222-234. https://doi.org/10.1044/1058-0360(2007/027)

Kent, R. D., & Vorperian, H. K. (2013). Speech impairment in Down syndrome: A review. *Journal of Speech, Language & Hearing Research*, *56*(1), 178-210. https://doi.org/10.1044/1092-4388(2012/12-0148)

Kim, H. (2016). Familiarization effects on consonant intelligibility in dysarthric speech. *Folia Phoniatrica et Logopaedica*, *67*(5), 245-252. https://doi.org/10.1159/000444255

Knoblich, G., & Flach, R. (2001). Predicting the effects of actions: Interactions of perception and action. *Psychological Science*, *12*(6), 467-472. https://doi.org/10.1111/1467-9280.00387

Kraljic, T., Brennan, S. E., & Samuel, A. G. (2008). Accommodating variation: Dialects, idiolects, and speech processing. *Cognition*, *107*(1), 54-81. https://doi.org/10.1016/j.cognition.2007.07.013

Kraljic, T., & Samuel, A. G. (2005). Perceptual learning for speech: Is there a return to normal? *Cognitive Psychology*, *51*(2), 141-178. https://doi.org/10.1016/j.cogpsych.2005.05.001

Kraljic, T., & Samuel, A. G. (2011). Perceptual learning evidence for contextually-specific representations. *Cognition*, *121*(3), 459-465. https://doi.org/10.1016/j.cognition.2011.08.015

Kraljic, T., Samuel, A. G., & Brennan, S. E. (2008). First impressions and last resorts: How listeners adjust to speaker variability. *Psychological Science*, *19*(4), 332-338. https://doi.org/10.1111/j.1467-9280.2008.02090.x

Kreitewolf, J., Mathias, S. R., & von Kriegstein, K. (2017). Implicit talker training improves comprehension of auditory speech in noise. *Frontiers in Psychology*, *8*, 1584. https://doi.org/10.3389/fpsyg.2017.01584

Kuhl, P. K., & Iverson, P. (1995). Linguistic experience and the « perceptual magnet effect ». In W. Strange (Éd.), *Speech perception and linguistic experience: Issues in cross-language research* (p. 121-154). Baltimore: York Press.

Kuhl, P. K., & Meltzoff, A. N. (1982). The bimodal perception of speech in infancy. *Science*, *218*(4577), 1138-1141. https://doi.org/10.1126/science.7146899

Kumin, L. (1994). Intelligibility of speech in children with Down syndrome in natural settings: Parents' perspective. *Perceptual and Motor Skills*, *78*(1), 307-313. https://doi.org/10.2466/pms.1994.78.1.307

Kumin, L. (2012). *Early communication skills for children with Down syndrome: A guide for parents and professionals. Woodbine House* (Third Edit). USA: Woodbine House Inc.

Lametti, D. R., Krol, S. A., Shiller, D. M., & Ostry, D. J. (2014). Brief periods of auditory perceptual training can determine the sensory targets of speech motor learning. *Psychological Science*, *25*(7), 1325-1336. https://doi.org/10.1177/0956797614529978

Lametti, D. R., Rochet-Capellan, A., Neufeld, E., Shiller, D. M., & Ostry, D. J. (2014). Plasticity in the human speech motor system drives changes in speech perception. *The Journal of Neuroscience*, *34*(31), 10339-10346. https://doi.org/10.1523/JNEUROSCI.0108-14.2014

Latash, M., Wood, L., & Ulrich, D. (2008). What is currently known about hypotonia, motor skill development, and physical activity in Down syndrome. *Down Syndrome Research and Practice (Online)*. https://doi.org/doi:10.3104/reviews.2074

Levelt, W. J. M. (1989). *Speaking: From intentions to spoken words.* Cambridge, MA, USA: The MIT Press.

Levelt, W. J. M. (1995). The ability to speak: from intentions to spoken words. *European Review*, *3*(01), 13-23. https://doi.org/10.1017/S1062798700001290

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, *74*(6), 431-461. https://doi.org/10.1037/h0020279

Liberman, A. M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, *54*(5), 358-368. https://doi.org/10.1037/h0044417

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition*, *21*(1), 1-36. https://doi.org/10.1016/0010-0277(85)90021-6

Lidestam, B., Moradi, S., Pettersson, R., & Ricklefs, T. (2014). Audiovisual training is better than auditory-only training for auditory-only speech-in-noise identification. *The Journal of the Acoustical Society of America*, *136*(2), EL142-EL147. https://doi.org/10.1121/1.4890200

Liss, J. M., Spitzer, S. M., Caviness, J. N., & Adler, C. (2002). The effects of familiarization on intelligibility and lexical segmentation in hypokinetic and ataxic dysarthria. *The Journal of the Acoustical Society of America*, *112*(6), 3022-3030. https://doi.org/10.1121/1.1515793

Liss, J., Spitzer, S., Caviness, J., Adler, C., & Edwards, B. (2000). Lexical boundary error analysis in hypokinetic and ataxic dysarthria. *The Journal of the Acoustical Society of America*, *107*(6), 3415-3424. https://doi.org/10.1121/1.429412

Lively, S. E., Pisoni, D. B., Yamada, R. A., Tohkura, Y., & Yamada, T. (1994). Training Japanese listeners to identify English /r/ and /l/. III. Long-term retention of

new phonetic categories. *The Journal of the Acoustical Society of America*, *96*(4), 2076-2087. https://doi.org/10.1121/1.410149

Loane, M., Morris, J. K., Addor, M.-C., Arriola, L., Budd, J., Doray, B., … Dolk, H. (2013). Twenty-year trends in the prevalence of Down syndrome and other trisomies in Europe: Impact of maternal age and prenatal screening. *European Journal of Human Genetics*, *21*(1), 27-33. https://doi.org/10.1038/ejhg.2012.94

Logan, J. S., Lively, S. E., & Pisoni, D. B. (1991). Training Japanese listeners to identify English /r/ and /l/: A first report. *The Journal of the Acoustical Society of America*, *89*(2), 874-886. https://doi.org/10.1121/1.1894649

Maas, E., Robin, D. A., Freedman, S. E., Wulf, G., & Schmidt, R. A. (2008). Principles of Motor Learning in Treatment of Motor Speech Disorders. *American Journal of Speech-Language Pathology*, *17*(3), 277-298.

Määttä, T., Tervo-Määttä, T., Taanila, A., Kaski, M., & Iivanainen, M. (2007). Mental health, behaviour and intellectual abilities of people with Down syndrome. *Down Syndrome Research and Practice*, *11*(1), 37-43. https://doi.org/10.3104/reports.313

Macho, V., Andrade, D., Areias, C., Coelho, A., & Melo, P. (2014). Comparative Study of the Prevalence of Occlusal Anomalies in Down Syndrome Children and Their Siblings. *British Journal of Medicine and Medical Research*, *4*(35), 5604-5611. https://doi.org/10.9734/BJMMR/2014/12688

Mahler, L. A., & Jones, H. N. (2012). Intensive treatment of dysarthria in two adults with Down syndrome. *Developmental Neurorehabilitation*, *15*(1), 44-53. https://doi.org/10.3109/17518423.2011.632784

Maye, J., Aslin, R. N., & Tanenhaus, M. K. (2008). The weckud wetch of the wast: lexical adaptation to a novel accent. *Cognitive Science*, *32*(3), 543-562. https://doi.org/10.1080/03640210802035357

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*(5588), 746-748. https://doi.org/10.1038/264746a0

Miller, G. A., & Nicely, P. E. (1955). An analysis of perceptual confusions among some English consonants. *The Journal of the Acoustical Society of America*, *27*(2), 338-352. https://doi.org/10.1121/1.1907526

Mitchel, A. D., Gerfen, C., & Weiss, D. J. (2016). Audiovisual perceptual learning with multiple speakers. *Journal of Phonetics*, *56*, 66-74. https://doi.org/10.1016/j.wocn.2016.02.003

Mitterer, H., & McQueen, J. M. (2009). Foreign subtitles help but native-language subtitles harm foreign speech perception. *PLoS ONE*, *4*(11), e7785. https://doi.org/10.1371/journal.pone.0007785

Moradi, S., Lidestam, B., Ning Ng, E. H., Danielsson, H., & Rönnberg, J. (2019). Perceptual doping: An audiovisual facilitation effect on auditory speech

processing, from phonetic feature extraction to sentence identification in noise. *Ear and Hearing*, *40*(2), 312-327. https://doi.org/10.1097/AUD.0000000000000616

Morris, A. F., Vaughan, S. E., & Vaccaro, P. (1982). Measurements of neuromuscular tone and strength in Down's syndrome children. *Journal of Intellectual Disability Research*, *26*(1), 41-46. https://doi.org/10.1111/j.1365-2788.1982.tb00127.x

Moura, C. P., Cunha, L. M., Vilarinho, H., Cunha, M. J., Freitas, D., Palha, M., … Pais-Clemente, M. (2008). Voice parameters in children with Down syndrome. *Journal of Voice*, *22*(1), 34-42. https://doi.org/10.1016/j.jvoice.2006.08.011

Mullennix, J. W., Pisoni, D. B., & Martin, C. S. (1989). Some effects of talker variability on spoken word recognition. *The Journal of the Acoustical Society of America*, *85*(1), 365-378. https://doi.org/10.1121/1.397688

Nasir, S. M., & Ostry, D. J. (2009). Auditory plasticity and speech motor learning. *Proceedings of the National Academy of Sciences*, *106*(48), 20470-20475. https://doi.org/10.1073/pnas.0907032106

Niedzielski, N. (1999). The effect of social information on the perception of sociolinguistic variables. *Journal of Language and Social Psychology*, *18*(1), 62-85. https://doi.org/10.1177/0261927X99018001005

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology*, *47*(2), 204-238. https://doi.org/10.1016/S0010-0285(03)00006-9

Nygaard, L. C., & Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Perception & Psychophysics*, *60*(3), 355–376. https://doi.org/10.3758/BF03206860

Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1994). Speech perception as a talker-contingent process. *Psychological Science*, *5*(1), 42-46. https://doi.org/10.1111/j.1467-9280.1994.tb00612.x

Olmstead, A. J., Viswanathan, N., Aivar, M. P., Manuel, S., Kim, M., & Alexander, J. (2013). Comparison of native and non-native phone imitation by English and Spanish speakers. *Frontiers in Psychology*, *4*(475), 1-7. https://doi.org/10.3389/fpsyg.2013.00475

Pardo, J. S. (2012). Reflections on phonetic convergence: Speech perception does not mirror speech production. *Language and Linguistics Compass*, *6*(12), 753-767. https://doi.org/10.1002/lnc3.367

Pardo, J. S. (2013). Measuring phonetic convergence in speech production. *Frontiers in Psychology*, *4*, 559. https://doi.org/10.3389/fpsyg.2013.00559

Parker, S. E., Mai, C. T., Canfield, M. A., Rickard, R., Wang, Y., Meyer, R. E., … Correa, A. (2010). Updated national birth prevalence estimates for selected birth defects in the United States, 2004-2006. *Birth Defects Research Part A: Clinical*

and Molecular Teratology, *88*(12), 1008-1016. https://doi.org/10.1002/bdra.20735

Peelle, J. E., & Sommers, M. S. (2015). Prediction and constraint in audiovisual speech perception. *Cortex*, *68*, 169-181. https://doi.org/10.1016/j.cortex.2015.03.006

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *The Journal of the Acoustical Society of America*, *24*(2), 175-184. https://doi.org/10.1121/1.1906875

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, *27*(2), 169-190. https://doi.org/10.1017/S0140525X04000056

Pilling, M., & Thomas, S. (2011). Audiovisual cues and perceptual learning of spectrally distorted speech. *Language and Speech*, *54*(4), 487-497. https://doi.org/10.1177/0023830911404958

Pisoni, D. B. (1975). Auditory short-term memory and vowel perception. *Memory & Cognition*, *3*(1), 7-18. https://doi.org/10.3758/BF03198202

Plomp, R., & Mimpen, A. M. (1979). Improving the Reliability of Testing the Speech Reception Threshold for Sentences. *Audiology*, *18*(1), 43-52.

Prinz, W. (1997). Perception and action planning. *European Journal of Cognitive Psychology*, *9*(2), 129-154. https://doi.org/10.1080/713752551

Pulvermüller, F., Huss, M., Kherif, F., del Prado Martin, F. M., Hauk, O., & Shtyrov, Y. (2006). Motor cortex maps articulatory features of speech sounds. *Proceedings of the National Academy of Sciences*, *103*(20), 7865-7870. https://doi.org/10.1073/pnas.0509989103

Reinisch, E., & Holt, L. L. (2014). Lexically guided phonetic retuning of foreign-accented speech and its generalization. *Journal of Experimental Psychology Human Perception & Performance*, *40*(2), 539-555.

Reisberg, D., McLean, J., & Goldfield, A. (1987). Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli. In B. Dodd & R. Campbell (Éd.), *Hearing by eye: The psychology of lip-reading* (p. 97–114). Hillsdale, NJ, USA: Lawrence Erlbaum Associates, Inc.

Repp, B. H., & Knoblich, G. (2004). Perceiving action identity: How pianists recognize their own performances. *Psychological Science*, *15*(9), 604-609. https://doi.org/10.1111/j.0956-7976.2004.00727.x

Rizzolatti, G., & Craighero, L. (2004). The mirror-neuron system. *Annual Review of Neuroscience*, *27*(1), 169-192. https://doi.org/10.1146/annurev.neuro.27.070203.144230

Rizzolatti, G., Fogassi, L., & Gallese, V. (2001). Neurophysiological mechanisms underlying the understanding and imitation of action. *Nature Reviews Neuroscience*, *2*(9), 661-670. https://doi.org/10.1038/35090060

Roberts, J. E., Price, J., & Malkin, C. (2007). Language and communication development in down syndrome. *Mental Retardation and Developmental Disabilities Research Reviews*, *13*(1), 26-35. https://doi.org/10.1002/mrdd.20136

Roberts, J., Long, S. H., Malkin, C., Barnes, E., Skinner, M., Hennon, E. A., & Anderson, K. (2005). A comparison of phonological skills of boys with fragile X syndrome and Down syndrome. *Journal of Speech, Language & Hearing Research*, *48*(5), 980-995. https://doi.org/10.1044/1092-4388(2005/067)

Rochet-Capellan, A., & Dohen, M. (2015). Acoustic characterisation of vowel production by young adults with Down syndrome. In *Proceedings of ICPhS 2015* (p. 5). Glasgow, Scotland.

Rosen, S. M., Fourcin, A. J., & Moore, B. C. J. (1981). Voice pitch as an aid to lipreading. *Nature*, *291*(5811), 150-152.

Rosenblum, L. D. (2005). Primacy of multimodal speech perception. In D. B. Pisoni & R. E. Remez (Éd.), *The handbook of speech perception* (p. 51-78). Malden, MA: Blackwell. https://doi.org/10.1002/9780470757024.ch3

Rosenblum, L. D. (2008). Speech perception as a multimodal phenomenon. *Current Directions in Psychological Science*, *17*(6), 405-409. https://doi.org/10.1111/j.1467-8721.2008.00615.x

Rosenblum, L. D., Miller, R. M., & Sanchez, K. (2007). Lip-read me now, hear me better later. *Psychological Science*, *18*(5), 392-396. https://doi.org/10.1111/j.1467-9280.2007.01911.x

Rosin, M. M., Swift, E., Bless, D., & Kluppel Vetter, D. (1988). Communication profiles of adolescents with Down syndrome. *Journal of Childhood Communication Disorders*, *12*(1), 49-64. https://doi.org/10.1177/152574018801200105

Ross, L. A., Saint-Amour, D., Leavitt, V. M., Javitt, D. C., & Foxe, J. J. (2007). Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. *Cerebral Cortex*, *17*(5), 1147-1153. https://doi.org/10.1093/cercor/bhl024

Rubeis, G., & Steger, F. (2019). A burden from birth? Non-invasive prenatal testing and the stigmatization of people with disabilities. *Bioethics*, *33*(1), 91-97. https://doi.org/10.1111/bioe.12518

Rupela, V., Velleman, S. L., & Andrianopoulos, M. V. (2016). Motor speech skills in children with Down syndrome: A descriptive study. *International Journal of Speech-Language Pathology*, *18*(5), 483-492. https://doi.org/10.3109/17549507.2015.1112836

Samuel, A. G. (2011). Speech perception. *Annual Review of Psychology*, *62*, 49-72. https://doi.org/10.1146/annurev.psych.121208.131643

Samuel, A. G., & Kraljic, T. (2009). Perceptual learning for speech. *Attention,*

Perception, & Psychophysics, *71*(6), 1207-1218. https://doi.org/10.3758/APP.71.6.1207

Samuel, A. G., & Larraza, S. (2015). Does listening to non-native speech impair speech perception ? *Journal of Memory and Language*, *81*, 51-71. https://doi.org/10.1016/j.jml.2015.01.003

Sanchez, K., Dias, J. W., & Rosenblum, L. D. (2013). Experience with a talker can transfer across modalities to facilitate lipreading. *Attention, Perception, & Psychophysics*, *75*(7), 1359-1365. https://doi.org/10.3758/s13414-013-0534-x

Sato, M., Buccino, G., Gentilucci, M., & Cattaneo, L. (2010). On the tip of the tongue: Modulation of the primary motor cortex during audiovisual speech perception. *Speech Communication*, *52*(6), 533-541. https://doi.org/10.1016/j.specom.2009.12.004

Sato, M., Grabski, K., Glenberg, A. M., Brisebois, A., Basirat, A., Ménard, L., & Cattaneo, L. (2011). Articulatory bias in speech categorization: Evidence from use-induced motor plasticity. *Cortex*, *47*(8), 1001-1003. https://doi.org/10.1016/j.cortex.2011.03.009

Savariaux, C., Perrier, P., & Orliaguet, J. P. (1995). Compensation strategies for the perturbation of the rounded vowel [u] using a lip tube: A study of the control space in speech production. *The Journal of the Acoustical Society of America*, *98*(5), 2428-2442. https://doi.org/10.1121/1.413277

Scarbel, L., Beautemps, D., Schwartz, J.-L., & Sato, M. (2014). The shadow of a doubt? Evidence for perceptuo-motor linkage during auditory and audiovisual close-shadowing. *Frontiers in Psychology*, *5*(568), 1-10. https://doi.org/10.3389/fpsyg.2014.00568

Schütz-Bosbach, S., & Prinz, W. (2007). Perceptual resonance: action-induced modulation of perception. *Trends in Cognitive Sciences*, *11*(8), 349-355. https://doi.org/10.1016/j.tics.2007.06.005

Schwab, E. C., Nusbaum, H. C., & Pisoni, D. B. (1985). Some Effects of Training on the Perception of Synthetic Speech. *Human Factors*, *27*(4), 395-408. https://doi.org/10.1177/001872088502700404

Schwartz, J.-L., Basirat, A., Ménard, L., & Sato, M. (2012). The Perception-for-Action-Control Theory (PACT): A perceptuo-motor theory of speech perception. *Journal of Neurolinguistics*, *25*(5), 336-354. https://doi.org/10.1016/j.jneuroling.2009.12.004

Schwartz, J.-L., Berthommier, F., & Savariaux, C. (2004). Seeing to hear better: Evidence for early audio-visual interactions in speech identification. *Cognition*, *93*(2), B69-B78. https://doi.org/10.1016/j.cognition.2004.01.006

Schwartz, J.-L., Robert-Ribes, J., & Escudier, P. (1998). *Hearing Eye II*. *Hearing by eye II: Advances in the psychology of speechreading and auditory-visual speech* (1st Editio). Routledge. https://doi.org/10.4324/9780203098752

Sebastián-Gallés, N., Dupoux, E., Costa, A., & Mehler, J. (2000). Adaptation to time-compressed speech: Phonological determinants. *Perception & Psychophysics*, *62*(4), 834-842. https://doi.org/10.3758/BF03206926

Sherwin, J., & Sajda, P. (2013). Musical experts recruit action-related neural structures in harmonic anomaly detection: Evidence for embodied cognition in expertise. *Brain and Cognition*, *83*(2), 190-202. https://doi.org/10.1016/j.bandc.2013.07.002

Shiller, D. M., Sato, M., Gracco, V. L., & Baum, S. R. (2009). Perceptual recalibration of speech sounds following speech motor learning. *The Journal of the Acoustical Society of America*, *125*(2), 1103-1113. https://doi.org/10.1121/1.3058638

Silverman, W. (2007). Down syndrome: Cognitive phenotype. *Mental Retardation and Developmental Disabilities Research Reviews*, *13*(3), 228-236. https://doi.org/10.1002/mrdd.20156

Skipper, J. I., Devlin, J. T., & Lametti, D. R. (2017). The hearing ear is always found close to the speaking tongue: Review of the role of the motor system in speech perception. *Brain and Language*, *164*, 77-105. https://doi.org/10.1016/j.bandl.2016.10.004

Skotko, B. G., Capone, G. T., & Kishnani, P. S. (2009). Postnatal diagnosis of Down syndrome: Synthesis of the evidence on how best to deliver the news. *Pediatrics*, *124*(4), e751-e758. https://doi.org/10.1542/peds.2009-0480

Skotko, B. G., Kishnani, P. S., & Capone, G. T. (2009). Prenatal diagnosis of Down syndrome: How best to deliver the news. *American Journal of Medical Genetics Part A*, *149*(11), 2361-2367. https://doi.org/10.1002/ajmg.a.33082

Skotko, B. G., Levine, S. P., & Goldstein, R. (2011). Self-perceptions from people with Down syndrome. *American Journal of Medical Genetics Part A*, *155*(10), 2360-2369. https://doi.org/10.1002/ajmg.a.34235

Smith, B. L., & Stoel-Gammon, C. (1983). A longitudinal study of the development of stop consonant production in normal and Down's syndrome children. *Journal of Speech and Hearing Disorders*, *48*(2), 114-118. https://doi.org/10.1044/jshd.4802.114

Smith, E., Gray, S., Verdolini, K., & Lemke, J. (1995). Effects of voice disorders on quality of life. *Otolaryngology–Head and Neck Surgery*, *113*(2), P121-P121. https://doi.org/10.1016/S0194-5998(05)80764-8

Sommers, R. K., Patterson, P., & Wildgen, P. L. (1988). Phonology of Down syndrome speakers, ages 13-22. *Journal of Childhood Communication Disorders*, *12*(1), 65-91. https://doi.org/10.1177/152574018801200106

Souza, P., Gehani, N., Wright, R., & McCloy, D. (2013). The advantage of knowing the talker. *Journal of the American Academy of Audiology*, *24*(8), 689-700. https://doi.org/10.3766/jaaa.24.8.6

Spender, Q., Dennis, J., Stein, A., Cave, D., & Percy, E. (1995). Impaired oral-motor function in children with Down's syndrome: A study of three twin pairs. *European Journal of Disorders of Communication*, *30*(1), 77-87.

Stevens, K. N. (1989). On the quantal nature of speech. *Journal of Phonetics*, *17*, 3-45.

Stevens, K. N., & Blumstein, S. E. (1978). Invariant cues for place of articulation in stop consonants. *The Journal of the Acoustical Society of America*, *64*(5), 1358-1368. https://doi.org/10.1121/1.382102

Stoel-Gammon, C. (1997). Phonological development in Down syndrome. *Mental Retardation and Developmental Disabilities Research Reviews*, *3*(4), 300-306. https://doi.org/10.1002/(SICI)1098-2779(1997)3:4<300::AID-MRDD4>3.0.CO;2-R

Strand, J. F., Brown, V. A., & Barbour, D. L. (2018). Talking points: A modulating circle reduces listening effort without improving speech recognition. *Psychonomic Bulletin & Review*, *26*(1), 291-297. https://doi.org/10.3758/s13423-018-1489-7

Strange, W. (1995). *Speech perception and linguistic experience: Issues in cross-language research*. (W. Strange, Éd.). Baltimore: York Press.

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *The Journal of the Acoustical Society of America*, *26*(2), 212-215. https://doi.org/10.1121/1.1907309

Summerfield, Q. A. (1979). Use of visual information for phonetic perception. *Phonetica*, *36*(4-5), 314-331. https://doi.org/10.1159/000259969

Timmins, C., Hardcastle, W. J., Wood, S., & Cleland, J. (2011). An EPG analysis of /t/ in young people with Down's syndrome. *Clinical linguistics & phonetics*, *25*(11-12), 1022-1027. https://doi.org/10.3109/02699206.2011.616981

Tjaden, K. K., & Liss, J. M. (1995). The role of listener familiarity in the perception of dysarthric speech. *Clinical Linguistics & Phonetics*, *9*(2), 139-154. https://doi.org/10.3109/02699209508985329

Turner, A. C., McIntosh, D. N., & Moody, E. J. (2015). Don't listen with your mouth full: The role of facial motor action in visual speech perception. *Language and Speech*, *58*(2), 267-278. https://doi.org/10.1177/0023830914542305

Tye-Murray, N., Spehar, B. P., Myerson, J., Hale, S., & Sommers, M. S. (2013). Reading your own lips: Common-coding theory and visual speech perception. *Psychonomic Bulletin & Review*, *20*(1), 115-119. https://doi.org/10.3758/s13423-012-0328-5

Vilain, A., Dole, M., Lœvenbruck, H., Pascalis, O., & Schwartz, J.-L. (2019). The role of production abilities in the perception of consonant category in infants. *Developmental Science*, e12830. https://doi.org/10.1111/desc.12830

von Kriegstein, K., Dogan, Ö., Grüter, M., Giraud, A.-L., Kell, C. A., Grüter, T., … Kiebel,

S. J. (2008). Simulation of talking faces in the human brain improves auditory speech recognition. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(18), 6747-6752. https://doi.org/10.1073/pnas.0710826105

Wade, T., Jongman, A., & Sereno, J. (2007). Effects of acoustic variability in the perceptual learning of non-native-accented speech sounds. *Phonetica*, *64*(2-3), 122-144. https://doi.org/10.1159/000107913

Werker, J. F., Gilbert, J. H. V, Humphrey, K., & Tees, R. C. (1981). Developmental Aspects of Cross-Language Speech Perception. *Child Development*, *52*(1), 349-355. https://doi.org/10.2307/1129249

Werker, J. F., & Tees, R. C. (1984). Cross-language speech perception: Evidence for perceputal reorganization during the first year of life. *Infant Behavior and Development*. https://doi.org/10.1016/S0163-6383(84)80022-3

Werker, J. F., & Yeung, H. H. (2005). Infant speech perception bootstraps word learning. *Trends in Cognitive Sciences*, *9*(11), 519-527. https://doi.org/10.1016/j.tics.2005.09.003

Wilson, M., & Knoblich, G. (2005). The case for motor involvement in perceiving conspecifics. *Psychological Bulletin*, *131*(3), 460-473. https://doi.org/10.1037/0033-2909.131.3.460

Xue, S. A., Kaine, L., & Ng, M. L. (2010). Quantification of vocal tract configuration of older children with Down syndrome: A pilot study. *International Journal of Pediatric Otorhinolaryngology*, *74*(4), 378-383. https://doi.org/10.1016/j.ijporl.2010.01.007

Yeung, H. H., & Werker, J. F. (2013). Lip movements affect infants' audiovisual speech perception. *Psychological Science*, *24*(5), 603-612. https://doi.org/10.1177/0956797612458802

Zhang, X., & Samuel, A. G. (2014). Perceptual learning of speech under optimal and adverse conditions. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(1), 200-217. https://doi.org/10.1037/a0033182