Percevoir et agir : la nature sensorimotrice, multisensorielle et prédictive de la perception de la parole

Avril Treille

Communauté
UNIVERSITÉ Grenoble Alpes

THÈSE

Pour obtenir le grade de

DOCTEUR DE LA COMMUNAUTÉ UNIVERSITÉ GRENOBLE ALPES

Spécialité : **Sciences Cognitives, Psychologie Cognitive & Neurocognition**

Arrêté ministériel : 25 mai 2016

Présentée par

Avril Treille

Thèse dirigée par **Marc SATO, Coriandre VILAIN** et **Jean-Luc SCHWARTZ**

préparée au sein du **Laboratoire Grenoble Images Parole Signal & Automatique (GIPSA-Lab, UMR 5216)**
dans **l'École Doctorale Ingénierie pour la santé, la Cognition et l'Environnement**

Percevoir et agir : La nature sensorimotrice, multisensorielle et prédictive de la perception de la parole

Thèse soutenue publiquement le **24 avril 2017**
devant le jury composé de :

Madame Cécile COLIN
Professeure, Université Libre de Bruxelles, Belgique - Rapporteuse
Madame Anne-Lise GIRAUD
Professeure, Université de Genève, Suisse - Rapporteuse
Madame Sonia KANDEL
Professeure, Université Grenoble-Alpes – Examinatrice et Présidente du jury
Madame Pascale TREMBLAY
Professeure, Université Laval, Canada - Examinatrice
Monsieur Marc SATO
Chargé de recherche, Université Aix-Marseille - Directeur de thèse (invité)
Monsieur Jean-Luc SCHWARTZ
Directeur de recherche, Université Grenoble Alpes – Co-Directeur de thèse
Monsieur Coriandre VILAIN
Ingénieur de recherche, Université Grenoble-Alpes – Co-Directeur de thèse

"[...] Scientists are people who know more and more about less and less, until they know everything about nothing."

Konrad Lorenz

REMERCIEMENTS

Il est de coutume de chercher à remercier de manière exhaustive tous les gens qui ont participé, de près ou de loin, à la réussite de cette thèse. C'est mission impossible, les personnes et les évènements s'entre-mêlent au-delà de l'explicable pour que je puisse n'oublier personne.

Alors, je vais commencer par te remercier toi, lecteur, pour être sûre de ne pas t'oublier. Que tu sois un ami, un chercheur (jeune ou moins jeune) ou juste une personne curieuse, je te remercie de prendre le temps de lire ce manuscrit (même si ce n'est qu'en pointillés). J'espère que tu trouveras ce que tu es venu chercher.

La validation de ce manuscrit et de tout mon travail de recherche n'aurait pu se faire sans les précieuses personnes qui ont eu la gentillesse de composer mon jury. Ainsi, je souhaiterais remercier sincèrement Anne-Lise Giraud et Cécile Colin pour leur travail de relecture et pour leurs questionnements pertinents qui m'ont permis d'entrevoir un avenir pour mes recherches. J'ai apprécié partager cette soutenance avec vous, même si tout m'a semblé se dérouler en un claquement de doigt ! Je voudrais ensuite adresser toute mon amitié à Sonia Kandel, qui en plus d'avoir présidé le jury, a souvent pris du temps pour m'écouter, m'aider, me soutenir. Tu as été quelqu'un sur qui j'ai pu compter Sonia, et je t'en remercie. Tes mots, toujours positifs et encourageants m'ont permis de garder le cap jusqu'à la fin de toute cette aventure. Tu as toujours cru en mes travaux, peut-être parfois plus que moi-même, et c'était un soutien moral non négligeable, alors merci. Je souhaiterai terminer par Pascale Tremblay, que je ne cesserai jamais de remercier pour tout ce qu'elle a pu m'apporter, scientifiquement et humainement. Durant trois mois tu as été ma directrice de substitution. J'ai pu m'adonner à ce que j'aime par-dessus tout : l'expérimentation. J'ai expérimenté la vie au Québec en même temps que la TMS, j'ai découvert la vie d'un autre labo et d'un autre pays. C'était une expérience déroutante et tu as fait en sorte qu'elle soit, pour moi, la plus enrichissante possible. Ta disponibilité, avant, pendant et même après mon séjour a été une aubaine pour moi. Tu as été quelqu'un d'une grande écoute et tes conseils, autant scientifiques que personnels m'ont été très précieux. Je te remercie pour cette vision féminine de la recherche qui manquait quelque peu dans mon encadrement… !

Mais pour que cette aventure se termine, il a bien fallu qu'elle débute quelque part. Et bien qu'elle ait commencé timidement lors d'un stage de master, dans une chambre sourde et sombre, encadrée par deux gars d'une quarantaine d'année, passionnés et un peu bizarres (il faut se le dire), elle a pris de l'ampleur lorsqu'elle a pris officiellement le nom de thèse. On m'avait prévenu : tu commence LA grande aventure. Evidemment je ne pouvais croire mes encadrants, devenus directeurs pour l'occasion. Mais j'ai fini par en prendre conscience lorsque je me suis retrouvée à négocier avec non plus deux, mais trois individus, avec leurs petits défauts et leurs immenses qualités. Parce qu'il faut être honnête, même si ça n'a pas toujours été facile, nous avons formé une sacré bonne équipe tous les 4 quand même non ? Coriandre, la lumière de ton bureau visible depuis le couloir m'a permis de toujours trouver porte ouverte pour mes questionnements, mes angoisses et ma motivation défaillante. De façon maladroite parfois, mais bien réelle, tu as toujours fait de ton mieux pour que j'arrive à trouver ma place au sein de ce petit groupe de chercheurs aux tempéraments complexes, et je t'en remercie chaleureusement. Je retiendrai nos concours de rapidité pour mettre du gel et poser les électrodes sur la tête de nos participants, tu as toujours su les mettre à l'aise, et moi avec. Ca doit être pour ça qu'ils sont toujours revenus, avec le sourire ! Alors merci de ne pas avoir fait fuir mes amis-cobayes !

Jean-Luc, ta capacité à te souvenir de tout et tout le monde m'a toujours impressionné, tu as été capable de passer d'un étudiant à l'autre en nous donnant à chacun l'impression d'être le seul. C'est formidable. Je souhaiterai te remercier pour ta bienveillance et ta positivité, tu as su voir le meilleur et c'est bien agréable lorsque l'on doute de tout, pour rien. Tu as su trouver du temps, même en dehors des heures respectables de travail pour répondre à mes questions et corriger ce manuscrit, tu as toujours été disponible malgré tes journées chargées. Ta place dans mon encadrement n'a pas été simple à trouver, mais à force de tatonnements et de négociations, je crois qu'on a fini par trouver un équilibre pour que tout se termine - bien. J'ai particulièrement pris plaisir à donner ce cours avec toi (merci Sonia de m'avoir offert cette opportunité), même si l'enseignement n'était pas du tout mon objectif, tu as eu raison de me répéter qu'il ne faut pas se fermer des portes. J'ai découvert la pédagogie et j'en suis tombée amoureuse. Merci d'avoir participé à mon avenir.
Et finalement Marc. Il y a 6 ans maintenant, tu avais oublié notre premier rendez-vous pour un stage de M1. J'aurais pu rencontrer quelqu'un d'autre, et devenir une toute autre personne, mais cette proposition loufoque d'un travail sur la perception tactile de la parole à l'aide d'électrodes avait piqué ma curiosité, et je n'ai pu me résoudre à abandonner. Et me voilà aujourd'hui, rendant ce manuscrit de thèse basé sur des expériences complètements abracadabrantes mêlant techniques de neuro complexes et modalités atypiques. Je crois que c'est cette folie qui m'a permis de continuer cette grande aventure. Je me suis sentie comme dans mes rêves de petite fille, la blouse blanche en moins. Tu as su me donner envie de travailler d'arrache-pied pour découvrir un jour le monde de la recherche. Tu as été un encadrant formidable durant mes stages de master et tu as su me laisser une certaine liberté à mon entrée en thèse, même si ça n'a pas toujours été simple. J'ai découvert ce monde merveilleux de l'expérimentation grâce à toi et je ne pourrais jamais te remercier suffisament pour cette opportunité. Tu m'as formé pour que je puisse décider de mon avenir et je tiens très sincèrement à te dire merci pour tout ça.

J'ai vécu 6 années au Gipsa-lab, un laboratoire très accueillant, particulièrement notre nid douillet stendhalien (froid en hiver, frais en été) ! J'aimerai remercier l'équipe PCMD et ses chercheurs et apprentis chercheurs (venus nombreux à ma soutenance !) pour leur dynamique et leur volonté de travailler, d'échanger ensemble sur des sujets qui nous regroupent tous, même si toutes les tentatives n'aboutissent pas toujours ! Je voudrais remercier particulièrement Elisabetta pour sa gentillesse infinie. Tu me connais depuis que je suis arrivée à l'université, tu m'as vue évoluer petit à petit vers ce que je suis aujourd'hui et je t'ai toujours vu avec le sourire, porteuse de cette bienveillance qui te va si bien. Merci de m'avoir donné confiance en moi et de m'avoir permis d'accèder aux bonnes personnes au bon moment. Je voudrais également souligner l'arrivée de Ito-san dans notre équipe. Ce n'est pas une histoire de culture, tu es bel et bien quelqu'un de très curieux, discret et extrêmement patient et attentif. Nous n'avons échangé que peu de mots, par timidité pour moi (la barrière de la langue), mais tu as assisté à chacune de mes représentations et tu as toujours trouvé le temps de venir me saluer et me dire que tu avais compris, même en français. Je suis heureuse d'avoir fait ta connaissance Takayuki. Mais j'ai également eu l'immense chance de voyager à l'autre bout de monde pour découvrir un autre laboratoire, au Québec. Et j'aimerai remercier les membres de l'équipe qui m'ont si bien accueillie dans le laboratoire des neurosciences de la parole et de l'audition : Camille, Marie-Hélène, Chloé, Isabelle, Catherine, Pascale B., j'ai été heureuse de passer ces quelques mois en votre compagnie ! Merci pour votre accent que j'adore ;) Et bien évidemment, une mention spéciale à Melo ! Tu as essayé de me (moi l'associable) faire découvrir du pays et des traditions d'ici, des choses qui en tant que française (mais québécoise de cœur j'en suis sûre) t'ont marquées, interloquées (les queues de castors, le hockey, la neige…). Merci d'avoir pris ce temps pour moi, de mon arrivée à l'aéroport à mon départ de l'aéroport. La boucle est bouclée. Les filles (et Max), je vous attends en France de pied ferme !

J'ai testé pas loin d'une centaine de personnes, et mine de rien avec le recul, ça fait une centaine de personnes que je n'ai pas suffisamment remerciées : participants/cobayes, français et québecois, famille et amis, inconnus parfois, merci du fond du cœur pour votre dévouement et votre

participation à des expériences parfois vraiment étranges. Vous êtes le cœur de ma thèse et c'est grâce à vous si j'ai pu satisfaire ma soif d'expériences.

Mais plus honnêtement, sans des amis très compréhensifs, tout ça n'aurait jamais été possible ! J'avais besoin de mes deux mondes : amies du labo (co-bureau proches et éloignées), parce qu'elles étaient les seules à mesurer la joie que l'on peut éprouver lors de l'acceptation de son premier papier, ou le stress de partir en conf, et les seules à comprendre toutes les subtilités des potins du labo. Merci Lulu pour tous ces moments à échanger (sciences et surtout autre…) côte à côte face à nos ordis respectifs. Un peu de fraicheur et quelques moments de pauses grâce à toi ! Tes post-it insistants et tes vidéos inutiles (mais drôles) me manqueront tout autant que ton humour ! Merci également à Marjo pour beaucoup de choses, pour être restée dans le bureau après le départ de lulu (c'était presque trop studieux…), pour avoir mangé avec moi presque tous les midis cette dernière année et surtout pour avoir partagé le iai (puis le jo) avec moi. Je crois que ça nous a permis de ne pas devenir totalement folles ! Notre havre de paix (et la base d'à peu près 90% de nos discussions) ! Je voudrais remercier également Diane pour ses conseils, pour nos longs papotages, pour les soirées filles qui font un bien fou ! Promis, mes câlins seront toujours disponibles, même quand on sera toutes les deux docteures au chômage ! Et enfin je voudrais remercier Laure, notre petite gestionnaire de Stendhal qui est devenue une véritable amie ! Merci pour ta motivation sans faille pour la piscine, merci de nous avoir accueillie dans ton super appart tout neuf pour nos soirées filles, merci de continuer encore à m'aider avec l'administration du labo, même si tu n'es plus là !! Bientôt futures collocs ;-) Je voudrais également écrire un petit mot pour Marie-Lou, mon premier contact avec d'autres doctorants de mon équipe (et pas loin d'être la seule…) ! On a commencé timidement par le club anime, puis on a trouvé un équilibre agréable entre discussions scientifiques et discussions philosophiques. Je voudrais te remercier pour tous ces moments passés ensemble, pour nos goûters chocolatés et nos séances cinés. Tu m'as soutenue pendant ma rédaction (et il fallait de la motivation pour me supporter…), et comme c'est bientôt ton tour, je vais faire de mon mieux pour te divertir sans t'empêcher de travailler ! Les filles, chacunes à votre manière, vous avez fait de ma thèse une expérience presque joyeuse, et je vous en remercie !

Je voudrais remercier des amis, et futurs docteurs de Paris, Sushi et Arthur ! Merci d'être venus jusqu'à Grenoble pour assister à ma soutenance, c'était vraiment super de vous avoir à mes côtés ! Merci à Angélique aussi, d'avoir abandonné ses jumelles pour venir m'écouter ! Ca m'a fait très plaisir de te revoir. Un grand merci également aux amis curieux du club de iai et de jo qui sont venus à ma soutenance : Anastasia et Sébastien, Clément, Fab (et Thibi et Dom pour votre intention de venir), merci ! Puis il y a les amis qui ne comprennent rien de rien à la recherche, mais qui ont toujours été là pour me changer les idées et m'écouter râler quand mes analyses ne fonctionnaient pas. Merci d'avoir fait semblant de comprendre, d'avoir essayé de vous intéresser et surtout d'avoir été là dans les moments difficiles. Magali, Flora, on s'est rencontré à la fac, puis on a pris des chemins différents, mais on a continué à partager nos vies et c'est le plus important pour moi. Et je voudrais terminer ma tirade amico-sentimentale sur ma meilleure amie, Carole-Anne. Je voudrais te remercier pour ta présence à mes côtés depuis si longtemps. Merci pour tout, tout simplement.

Un remerciement chaleureux s'impose à ma famille havraise, Titine, Camille, Alice, Pascal, Manu, et les petits plus si petits que ça (Suzanne, Clémence, Luna et Milo). Je suis venu chercher la fraicheur chez vous et surtout un endroit où m'aérer l'esprit pendant ma thèse, et ça m'a toujours fait du bien. J'ai même fait en sorte de vous apporter un bout de soleil avec moi, à chaque fois ;) Merci de m'avoir encouragée jusqu'au bout ! Merci à mon grand frère Romain pour le pot de ma soutenance en particulier, tu as gérer ça d'une main de maître ;-) Merci à mon jumeau. Je crois que parmi tous les visages bienveillants présents à ma soutenance, c'est le tien qui m'a le plus rassurée ! Et finalement, je voudrais remercier mon papa. Je suis fière d'avoir un père qui a pris le temps de lire ma thèse pour de vrai et jusqu'au bout et qui a essayé de comprendre le sens de mes recherches, un père qui fait tout pour me soutenir quelque soit mes choix, qui les acceptent (même si ce n'est pas toujours évident) et qui m'encourage à faire ce qui me rendra heureuse. Merci papa !

Un dernier énorme merci pour petit Seb qui partage ma vie. Je n'ai pas été facile à vivre cette dernière année, mais tu m'as supportée sans t'énerver, tu as essayé de me faire relativiser et m'a soutenue dans toutes mes lubies pour contrer l'effet fin de thèse (du naturel pour mes cheveux à mon envie de tout jeter par la fenêtre). Merci pour ta patience infinie, surtout aux cours de jo !

Hajime !

FINANCEMENTS

Mon InDoc au Québec a été en partie financé par la Région Rhône-Alpes, grâce à la bourse exploraDoc'.

Cette thèse a été entièrement financée par l'ERC Speech Unit(s) de Jean-Luc Schwartz, ce qui m'a permis de réaliser de très belles choses. Je voudrais donc remercier Jean-Luc de la confiance qu'il m'a accordée lorsqu'il a accepté de financer cette thèse, et m'a permis de participer à ce beau projet de recherche.

RESUME

PERCEVOIR ET AGIR : LA NATURE SENSORIMOTRICE, MULTISENSORIELLE ET PREDICTIVE DE LA PERCEPTION DE LA PAROLE

Voir les gestes articulatoires de son interlocuteur permet d'améliorer significativement le décodage et la compréhension du signal acoustique de parole émis. Un premier objectif de cette thèse était de déterminer si les interactions multimodales lors de la perception de parole, en plus d'impliquer classiquement les informations auditives et visuelles transmises par le son et le visage du locuteur, pouvaient être déclenchées par d'autres sources sensorielles moins communément utilisées dans la communication parlée, comme la perception tactile de la parole ou encore la perception visuelle des mouvements de la langue. Parallèlement, nos travaux avaient également pour but de déterminer l'implication possible du système moteur dans ces mécanismes de perception multisensorielle. Enfin, un autre enjeu de nos recherches était de déterminer plus avant le décours temporel et l'organisation neuroanatomique fonctionnelle de ces mécanismes d'intégration à l'aide de différentes techniques comme l'électro-encéphalographie, l'imagerie par résonance magnétique fonctionnelle ou encore la stimulation magnétique transcrânienne. Nos travaux ont permis d'élargir la notion de « multisensorialité de la parole » en mettant en évidence une facilitation des traitements temporels auditifs lors de la perception audio-tactile de la parole et lors de l'observation de nos propres mouvements articulatoires. D'autre part, nos études ont fourni de nouveaux arguments en faveur d'un rôle fonctionnel du système moteur lors de la perception de parole en montrant une activation plus importante des régions motrices lors de l'observation de mouvements de la langue ainsi qu'un recrutement plus bilatéral du cortex prémoteur ventral au cours du vieillissement. Pris ensemble, nos résultats renforcent l'idée d'un couplage fonctionnel, d'une co-structuration des systèmes de perception et de production de la parole. Les études présentées dans cette thèse appuient ainsi l'existence de connexions entre régions sensorielles, intégratives et motrices permettant la mise en œuvre de processus et traitements multisensoriels, sensorimoteurs et prédictifs lors de la perception et compréhension des actions de parole.

Mots-clés : Perception de la parole, interactions multisensorielles, système moteur, perception tactile, mouvements de la langue, perception de soi.

ABSTRACT

TO PERCEIVE AND TO ACT: THE SENSORIMOTOR, MULTISENSORY AND PREDICTIVE NATURE OF SPEECH PERCEPTION

Seeing the speaker's articulatory gestures significantly enhances auditory speech perception. A key issue is whether cross-modal speech interactions only depend on well-known auditory and visual inputs from the speaker's voice and face or, rather, might also be triggered by other sensory sources less common in speech communication, such as tactile information or vision of the tongue movements. Another goal of the present research was to determine the possible role of the motor system in these multisensory processes. Finally, we used electro-encephalographic, functional magnetic resonance imaging and transcranial magnetic stimulation techniques in order to better understand the time course and the functional neuroanatomical organization of these integration mechanisms. Our results extent the concept of "multisensory speech perception" by highlighting a facilitation of auditory processes during audio-haptic speech perception as well as during the observation of our own articulatory movements. They also provide new evidence in favor of a functional role of the motor system in speech perception by demonstrating an increase of motor activity during visuo-lingual speech perception and a more bilateral ventral premotor cortex recruitment during speech perception across aging. Taken together, our results reinforce the idea of a functional coupling and a co-structuring of speech perception and production systems. Our work support the existence of connections between sensory, integrative and motor regions allowing the implementation of multisensory, sensorimotor and predictive processes in the perception and understanding of speech actions.

Keywords: Speech perception, multisensory interactions, motor systems, tactile perception, lingual movements, self-speech perception.

TABLE DES MATIÈRES

TABLE DES FIGURES

AVANT-PROPOS

Ce projet de recherche porte sur la nature sensorimotrice, multisensorielle et prédictive de la perception de la parole. Initié lors de stages de master en sciences du langage, en sciences cognitives et en neuropsychologie, il a ensuite évolué naturellement vers un projet de thèse plus vaste s'intéressant aux processus de perception et d'intégration multimodale de la parole au travers des voies auditives, visuelles et motrices de notre cerveau, en combinant différentes approches issues de la phonétique, de la psychologie cognitive et des neurosciences cognitives.

Cette thèse réalisée dans le cadre du projet européen ERC « Speech Unit(e)s » se place dans un cadre théorique spécifique cherchant à développer et approfondir les connaissances des liens perceptivo-moteurs qui sont au cœur de débats très actuels en parole. Les questionnements que nous nous sommes posés tout au long de cette thèse consistaient à savoir si dans des conditions de perception inhabituelles – comme percevoir par le toucher les gestes articulatoires ou encore observer visuellement les mouvements linguaux d'un locuteur – notre cerveau était capable d'intégrer ces nouvelles informations, à l'aide de nos connaissances procédurales motrices, et si ces processus permettaient de prédire, anticiper, faciliter ou améliorer la perception de la parole.

Utilisant différentes techniques expérimentales comme l'électro-encéphalographie, l'imagerie par résonance magnétique fonctionnelle ou la stimulation magnétique transcrânienne, ces travaux de recherche sont au carrefour de plusieurs domaines, exploitant des connaissances variées, allant de la perception des actions de manière générale aux mécanismes d'intégration multisensorielle des signaux de parole, en passant par l'étude de ces processus au cours du vieillissement.

Du point de vue de l'organisation de ce manuscrit de thèse, nous avons pris le parti de rédiger ces travaux de recherche dans un format de thèse sur articles afin de présenter les différentes expériences réalisées de manière concise et rigoureuse.

Ce manuscrit est composé de trois grandes parties. Une première partie théorique regroupant quatre sous-parties aura pour but d'introduire certains concepts clés, théories et modèles principaux afin de contextualiser les questionnements de nos travaux. Nous y présenterons trois des caractéristiques principales de la perception de la parole que nous avons étudiées, à savoir sa nature sensorimotrice, multisensorielle et prédictive. Cette partie se terminera par un aperçu de la littérature relative à chacune des expériences réalisées, notamment la perception et l'intégration audio-tactile de la parole, la perception et l'intégration audio-visuo-linguale de la parole, la perception et l'intégration de nos propres actions et finalement la préservation des mécanismes d'intégration au cours du vieillissement.

Une seconde partie expérimentale, composée de quatre sous-parties, sera dédiée à la présentation sous forme d'articles publiés ou soumis des cinq expériences réalisées lors de cette thèse. Un résumé en français précèdera chaque article, suivi d'un bref rappel des

questionnements, des hypothèses, de la méthode utilisée, des résultats principaux obtenus et d'une brève conclusion.

Enfin, la dernière partie de ce manuscrit sera consacrée à une discussion générale des principaux résultats observés afin de préciser l'apport de chacune des expériences menées dans la compréhension des trois caractéristiques de la perception de la parole étudiées. Autrement dit, nous tâcherons d'expliquer en quoi chacune de nos études nous a permis d'approfondir les connaissances que nous avons de la nature multisensorielle, sensorimotrice et prédictive de la parole.

PARTIE THÉORIQUE – A

LA NATURE *SENSORIMOTRICE* DE LA PERCEPTION DE LA PAROLE

Communiquer : « faire passer », « transmettre », « partager » sont autant de synonymes qui soulignent l'importance de l'échange dans l'acte de communication. Pour que cette transmission ait lieu, il faut nécessairement l'émission et la réception d'un message linguistique. Et pour qu'il y ait compréhension de ce message, il faut un système commun de références qui permette d' « unir » ce qui est produit par le locuteur et ce qui est perçu par l'auditeur. Face à cette question, ce travail de thèse s'inscrit dans le cadre théorique d'un couplage fonctionnel entre systèmes sensoriels et moteur lors de la perception de la parole et, au travers des études présentées, il a pour objectif d'interroger la nature sensorimotrice des représentations de parole.

Dans cette première partie, nous discuterons dans un premier temps de la perception des actions, en explicitant certaines découvertes et théories majeures en faveur de ce lien perceptivo-moteur, pour enfin détailler le rôle du système moteur lors de la perception de la parole.

1. Percevoir une action – Lorsque notre répertoire moteur entre en action !

Depuis de nombreuses années, les chercheurs tentent de comprendre quels sont les mécanismes permettant d'établir un lien entre le monde physique qui nous entoure et les représentations internes que nous nous en faisons. A titre d'exemple, la Théorie du Codage Evénementiel (ou *Theory of Event Coding*, TEC), développée par Prinz, Hommel et collègues (2001), est une des théories sensorimotrices de la perception des actions qui tente de poser un cadre explicatif de ce lien entre la perception et l'action.

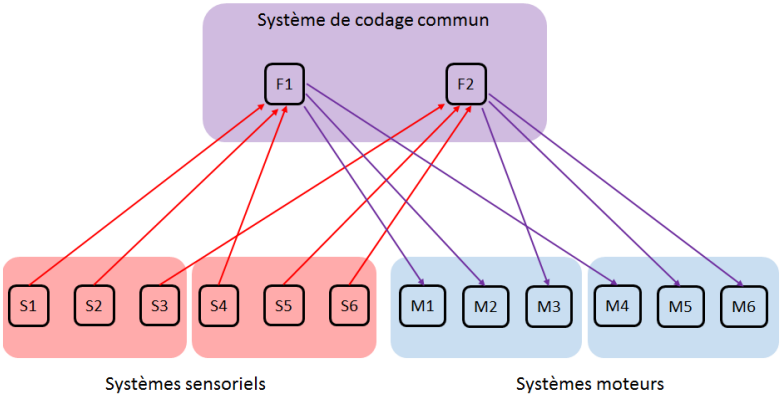

Figure 1 : Schéma illustrant la Théorie du Codage Evénementiel. En rouge sont présentés les différents systèmes sensoriels qui, lors de la perception d'une action, activeraient un certain nombre de traits (F1 ou F2 ici) abstraits relatifs à celle-ci au sein d'un cadre commun de référence (en violet). Ces traits ainsi activés propageraient l'information aux systèmes moteurs (en bleu) pour générer une réponse motrice adéquate au stimulus entrant. Ce système de codage commun permettrait ainsi d'établir un lien la perception et la réasliation d'une action (Figure tirée de Hommel et al., 2001).

D'après cette théorie, pour que nous puissions intégrer le contenu de nos expériences perceptives et les schémas d'actions relatifs à une réponse à un stimulus externe, il existerait un cadre de référence commun qui permettrait, à un niveau d'abstraction suffisant, d'intégrer ces informations de nature différente. Autrement dit, il existerait un code différent pour les informations sensorielles et les informations motrices et, afin d'intégrer ces codes dans un « langage » commun, un cadre plus abstrait serait composé de traits partagés à la fois par les systèmes sensoriels et par les systèmes moteurs (voir Figure 1). Ainsi, les informations sensorielles activeraient un certain nombre de traits abstraits présents dans ce cadre commun de référence. Cette activation se propagerait par la suite au système moteur afin de générer une réponse motrice aux stimuli entrants. Ce cadre de référence pourrait être le siège des représentations internes que nous nous faisons du monde physique extérieur permettant de faire le lien entre la perception et l'action.

Cette Théorie du Codage Evénementiel s'appuie sur un grand nombre d'études comportementales issues du domaine scientifique de la cognition incarnée (ou *embodied cognition*). Ces dernières années cependant, l'avènement des techniques de neuroimagerie a permis de déterminer plus avant l'implication possible du système moteur lors de la perception des actions. Dans cette section, nous rappellerons les principales découvertes neurobiologiques importantes qui viennent en appui de ce couplage perception-action. Ainsi, à travers une description d'un possible système miroir puis des spécificités de notre répertoire moteur, nous montrerons que nous sommes capables sinon de comprendre du moins de simuler une action observée au travers du recrutement partiel du réseau neuronal utilisé lors de la réalisation de cette action.

Le système miroir – Un des arguments neurobiologiques majeurs en faveur de ce lien perceptivo-moteur est la découverte des neurones miroirs issus du cortex préfrontal (F5) de primates non-humains (singe macaque, voir Figure 2-A). Cette région a pour fonction principale l'organisation et la planification d'un geste moteur. Mais en 1992, Rizzolatti et collègues (1996, 2001, 2004) ont découvert que certains des neurones moteurs de cette région avaient la faculté de décharger non seulement lorsque le singe produit une action dirigée vers un but précis (préhension d'un objet par la main ou par la bouche par exemple) mais également lors de l'observation d'actions similaires réalisées par un autre individu. Certains de ces neurones visuo-moteurs possèdent également la particularité d'être audio-visuo-moteurs, ceux-ci déchargeant lors de la réalisation et de l'observation visuelle mais également lors de l'écoute d'une action (Kohler et al., 2002 ; Keysers et al., 2003). Il est important de différencier ces neurones dits « miroirs » d'autres neurones moteurs dits « canoniques » qui déchargent non pas lors de l'observation d'une action mais à la simple vue d'un objet sans qu'une action particulière soit effectuée. De plus, même au sein des neurones miroirs, tous n'ont pas la même fonctionnalité. Ainsi, certains déchargent lors de l'observation d'une action transitive précise (impliquant une interaction entre un effecteur particulier et un objet) tandis que d'autres encore sont capables de généraliser le but de l'action peu importe l'effecteur utilisé en déchargeant durant l'observation d'une action (par exemple de saisie d'un objet) quel que soit l'effecteur utilisé (à l'aide de la main ou de la bouche par exemple). De plus, cette « co-activation » ne serait pas spécifique à cette espèce animale puisque les neurones miroirs chez le singe macaque s'activent aussi bien lorsque l'action est réalisée par un congénère que lorsqu'elle est réalisée par un humain.

D'autres régions étroitement liées à F5 semblent présenter des caractéristiques miroirs similaires, comme le lobe pariétal inférieur (IPL ; Fogassi et al., 2005) qui fait partie du réseau

fronto-pariétal responsable de l'organisation et exécution motrice d'une action, ou encore la partie postérieure du sillon temporal supérieur (pSTS) qui ne fait pas partie *stricto sensu* du réseau de neurones miroirs mais qui est normalement activée lors de la perception visuelle d'actions. Il n'est donc pas étonnant que cette région soit activée également lors de la réalisation d'une action visible, comme lors de la préhension d'un objet (l'action se faisant la plupart du temps dans notre champ de vision).

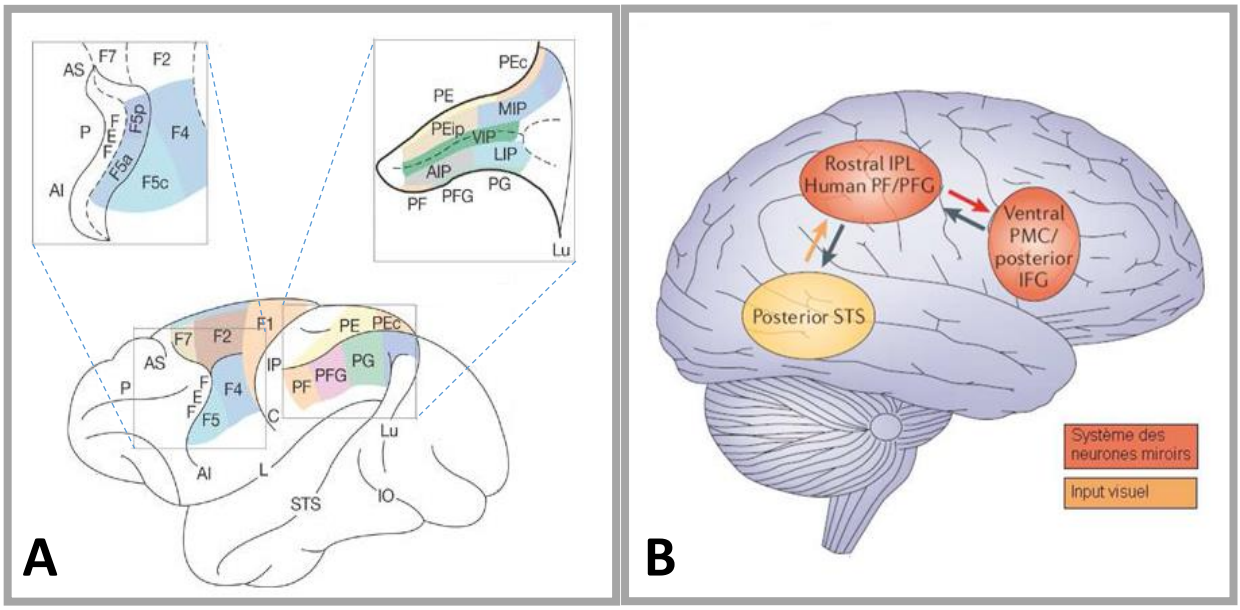

Figure 2 : ***A- Neurones miroirs chez le macaque : F5 est l'équivalent du cortex prémoteur chez le singe (figure tirée de Rizzolatti et al., 2008). B- Système miroir chez l'homme (figure tirée de Iacoboni et Wilson, 2006 ; IFG : gyrus frontal inférieur, Ventral PMC : cortex prémoteur ventral, IPL : lobe pariétal inférieur, PF/PFG : gyrus préfrontal, posterior STS : partie postérieure du sillon temporal supérieur)***

Chez l'Homme (voir Figure 2-B), il n'y a pas de preuves directes de l'existence de neurones miroirs du fait de la difficulté éthique à placer des électrodes intracérébrales. Cependant, à travers d'autres types d'investigations (comme l'imagerie par résonance magnétique fonctionnelle, ou IRMf, par exemple), l'existence d'un système miroir pariéto-frontal présentant des propriétés similaires aux neurones miroirs découverts chez le macaque, a pu être mise en évidence. En effet, une activation du cortex prémoteur ventral (possible homologue de la région F5 chez le macaque), ainsi qu'une activation de l'IPL ont été observées lorsque le sujet exécute une action et lorsqu'il observe ou entend une autre personne réaliser cette même action. D'autre part, en plus du codage d'actions transitives comme pour le singe, le système miroir de l'homme semble également s'activer lors de l'observation d'actions intransitives (sans interaction avec un objet) ou de mouvements sans finalités (voir par exemple Grèzes, et al., 1998 ; Buccino et al., 2001). L'existence d'un tel système miroir chez l'Homme suggère donc qu'observer une action recruterait au moins partiellement le circuit neuronal impliqué dans la réalisation de cette même action.

D'après la théorie du système miroir, observer des actions, c'est recueillir des informations concernant la nature de l'action (« Comment ») mais aussi le but de l'action (« Quoi »). Certaines recherches ont ainsi montré une activation différente du système miroir en fonction du but, de la finalité de l'action, alors que les mouvements permettant la réalisation de l'action étaient identiques (Iacoboni et al., 2005). Le système miroir aurait ainsi la

capacité de transformer l'information visuelle en connaissance motrice et de permettre un recodage, une transcription du but de l'action observée dans le système moteur de l'observateur. Bien que cette hypothèse soit débattue, le système miroir permettrait donc en partie de simuler et comprendre le but de l'action et par extension de prédire les mouvements à venir à partir de ceux observés, en fonction de la finalité de l'action.

Connaissances procédurales motrices – La notion de connaissances procédurales fait référence, dans la suite de ce texte, aux connaissances qu'une personne a des commandes motrices nécessaires pour exécuter tel ou tel mouvement, telle ou telle action. Ces connaissances, ou représentations motrices, seraient de fait liées aux actions que notre système moteur (dans son sens général) peut réaliser, c'est-à-dire à notre répertoire moteur.

Plusieurs études montrent une activation plus importante du cortex prémoteur et du cortex pariétal postérieur lors de la perception visuelle ou audio-visuelle d'un mouvement biologique (marcher, sauter, courir par exemple, actions faisant partie de notre répertoire moteur) par rapport à un mouvement non biologique (par exemple des vidéos d'actions biologiques sous forme de points lumineux qui ont été transformées pour ne plus respecter les lois cinématiques caractéristiques des mouvements biologiques, Saygin, 2007 – voir aussi Calvert, Campbell et Brammer, 2000 ; Howard et al., 1996). Il a également été montré qu'une partie des mécanismes de reconnaissance de l'action était assujettie aux contraintes et régularités propres aux mouvements observés. Par exemple, Johansson en 1973 a démontré pour la première fois qu'il nous était possible d'identifier des actions biologiques (comme marcher, courir ou danser) simplement sur la base de mouvements reproduits à partir de points lumineux placés préalablement sur les articulations d'un acteur (voir aussi Beardworth et Buckner, 1981 ; Loula et al., 2005). Ces résultats suggèrent que notre cerveau est capable d'inférer un lien logique entre ces différents points visuels à partir de leurs positions, du mouvement des uns par rapport aux autres et des connaissances procédurales que nous avons de notre corps et de ces mouvements afin d'en extraire des informations suffisamment complètes pour identifier l'action produite. D'autre part, nos systèmes moteur et prémoteur seraient également activés lors de la seule écoute d'un son d'action comme un coup frappé à la porte, un claquement de main ou un signal auditif plus complexe comme un morceau de piano (e.g., Aziz-Zadeh et al., 2004 ; Haueisen et Knösche, 2002 ; Lahav, Saltzman et Schlaug, 2007 ; Pizzamiglio et al., 2005). Ces résultats suggèrent ainsi un lien étroit entre représentations visuelles, auditives et motrices des actions.

De plus, reconnaître une action ne semble pas dépendre uniquement des caractéristiques de notre système moteur, mais plus véritablement de la spécificité du répertoire moteur partagé entre individus de la même espèce et relatif à leurs capacités physiques et/ou communicatives. Buccino et ses collègues (2004) ont étudié les activations motrices lors de la présentation visuelle de deux types de mouvements de la bouche, communicatifs ou non, exécutés par un homme (parler, manger), un singe (action communicative de « lip-smacking », manger) ou bien un chien (aboyer, manger). Cette étude a montré une activation plus importante du système miroir lorsque les sujets (humains) observaient des actions réalisées par leurs congénères ou compatibles avec leurs propres actions et une absence d'activation du cortex moteur lorsque l'action présentée ne faisait pas ou peu partie de leur répertoire moteur (comme aboyer par exemple). De leur côté, Tai et collègues (2004) ont mis en évidence la sensibilité aux mouvements biologiques humains dans les mécanismes de reconnaissance de l'action en observant une activation du cortex prémoteur

lors de la perception d'une action de préhension de la main réalisée par un humain, mais aucune activation lorsque cette action était réalisée par un robot.

Il semblerait que ce couplage sensorimoteur se construise petit à petit au rythme de nos apprentissages et de nos expériences. Calvo-Merino et ses collègues (2005, 2006) ont ainsi montré que l'implication des aires motrices était plus liée à l'apprentissage moteur qu'à la familiarité visuelle des actions observées. Ils ont en effet observé, en plus de régions pariétales et du cervelet, une plus forte activation du cortex prémoteur quand des danseurs professionnels de sexe masculin percevaient des mouvements issus de leur propre répertoire moteur par rapport à des mouvements réalisés par des danseuses professionnelles, mouvements qu'ils ont l'habitude de voir mais qu'ils n'ont jamais réalisés. Cette étude suggère ainsi que la reconnaissance d'une action ne dépend pas uniquement de connaissances et traitements visuels (les danseurs hommes et femmes ayant une expérience visuelle de l'ensemble des mouvements masculins et féminins) mais aussi de l'utilisation des connaissances procédurales propres au sujet (voir aussi Cross et al., 2006).

Dans cette première section, nous avons rappelé quelques découvertes clés concernant le couplage perception-action, autrement dit une utilisation partielle du réseau neuronal recruté pour réaliser une action dans le but de décoder les mouvements perçus. Ces études appuient ainsi l'idée que notre répertoire moteur, construit tout au long de nos apprentissages et fonction des contraintes et spécificités propres à tel ou tel mouvement, est impliqué dans les mécanismes de reconnaissance des actions.

2. Les différentes théories de la perception de la parole

Dans le cadre de nos travaux, la parole est pour partie considérée comme une action communicative particulière, soit une succession de gestes articulatoires ayant pour but final la production de sons pourvus de sens. Par le passé, la perception de la parole et des gestes articulatoires a été envisagée selon trois approches et axes théoriques distincts : un axe purement moteur, illustré par les théories motrices de la perception de la parole, un axe purement auditif, illustré par les théories et approches auditives de la perception de la parole, et un axe perceptivo-moteur, illustré par les théories sensorimotrices de la perception de la parole. Une question est à l'origine de ces approches théoriques très variées : comment pouvons-nous décrire de manière simple les relations existantes entre les sons et les phonèmes, c'est-à-dire entre les propriétés du signal acoustique entrant et l'interprétation de ce dernier en termes d'unités minimales de parole ? Nous allons présenter dans cette section quelques-unes des théories représentatives de chacun des axes mentionnés.

Théorie motrice de la perception de la parole – Dans les années 60, dans le cadre de travaux pionniers portant sur la synthèse de la parole, Liberman et collègues (1967, 1985) ont constaté une très grande variabilité du signal acoustique, variabilité difficilement exploitable pour déterminer des paramètres suffisamment robustes pour synthétiser la parole. Ils ont en effet observé que les relations entre les variables du signal acoustique de la parole et les phonèmes étaient extrêmement complexes et non linéaires du fait, notamment, de la présence de phénomènes tels que la coarticulation. Ce phénomène est intrinsèque au flux continu de parole. Il est dû à la juxtaposition rapide de gestes articulatoires pour former un signal acoustique continu. Cette succession ininterrompue de gestes a pour conséquence une influence du son produit sur le son à produire, rendant ainsi l'extraction d'invariants acoustiques (une portion du signal sonore correspondant à un phonème) extrêmement

difficile. Ils ont alors cherché à définir de nouveaux invariants relatifs non plus aux caractéristiques acoustiques du signal auditif, mais aux gestes du locuteur, supposés plus stables, ayant permis la production de celui-ci. Selon la théorie motrice de la perception de la parole, ce serait donc les gestes moteurs du locuteur qui seraient comparés et appariés à des représentations internes motrices des gestes de paroles de l'auditeur. Trois hypothèses spécifiques sont associées à cette théorie : 1) La perception de la parole implique le système de production de la parole : la compréhension de la parole se ferait par récupération des causes motrices qui seraient ensuite décodées, 2) la parole est perçue comme une suite de gestes du conduit vocal, non une suite de sons : les propriétés acoustiques diffèrent suivant le contexte, pas les commandes motrices, 3) la parole est traitée par un système phonétique spécialisé : la perception de la parole est innée et propre à l'Homme (« ce qui compte pour le locuteur compte pour l'auditeur », Liberman et Mattingly, 1985). Ces assertions ont été depuis revues et critiquées, mais Liberman et ses collègues ont été parmi les premiers scientifiques à avoir posé l'hypothèse d'un lien étroit entre perception et production de la parole.

Théorie directe réaliste de la perception de la parole – Fowler (1986, 1996), en désaccord avec son collègue Liberman au sujet du caractère spécifique de la parole et sur la nature de l'objet perceptif, a proposé une autre théorie qui se base cependant elle aussi sur l'aspect articulatoire de la parole. Selon cette théorie, le décodage perceptif se ferait par la reconnaissance directe de la nature de la source de perturbation perçue dans l'environnement, que ce soit pour la vision, le toucher ou l'audition et qu'il s'agisse de sons de parole ou non. Dans le cas de la perception d'un son de parole, à la différence de la théorie motrice de la perception de la parole de Liberman et collègues, les mouvements du conduit vocal seraient directement récupérés à partir du signal sonore et non inférés à partir des connaissances motrices de l'interlocuteur. Ainsi, le signal acoustique de parole serait porteur d'invariants relatifs aux gestes articulatoires produits par le locuteur et récupérés directement par l'interlocuteur lors de la perception du signal sonore.

Les théories et approches auditives de la perception de la parole – A contrario, les théories et approches auditives (Stevens et Blumstein, 1978 ; Blumstein et Stevens, 1979 ; Diehl et coll., 2004) considèrent la perception de parole comme dépendant uniquement de traitements auditifs. Les invariants phonémiques seraient extraits directement du signal acoustique et traités uniquement par le système auditif. Il n'y aurait aucun processus de décodage faisant appel aux commandes motrices de l'auditeur ou aux gestes articulatoires utilisés pour produire les sons de parole. En outre, le décodage acoustico-phonétique reposerait sur des mécanismes perceptifs plus généraux, non spécifiques à la parole et non propres à l'Homme, remettant ainsi en cause la théorie motrice de la perception de la parole de Liberman et collègues. Au travers de l'exemple de la perception catégorielle (c'est-à-dire notre capacité à discriminer deux sons appartenant à deux catégories phonologiques distinctes tandis que deux sons issus d'une même catégorie sont confondus, malgré un écart acoustique similaire entre les deux sons dans les deux cas), des travaux ont en effet montré que la perception catégorielle pouvait s'appliquer à d'autres sons non porteurs de sens (Stevens et Klatt, 1974 ; Pisoni, 1977), et être mise en évidence sur d'autres espèces animales lors de la présentation de sons de parole ou non (Kuhl et Miller, 1975, 1978). La perception catégorielle ne serait donc pas liée à un système phonétique spécialisé ni aux capacités et connaissances motrices propres à la parole.

Les théories sensorimotrices de la perception de la parole – Les théories sensorimotrices considèrent que la compréhension de la parole ne repose pas uniquement sur les traitements auditifs des propriétés acoustiques du signal de parole, ni sur des mécanismes purement moteurs. Ces théories proposent différentes façons de formaliser ce couplage entre les mécanismes auditifs et les connaissances procédurales motrices lors de la perception de la parole, un couplage sensorimoteur qui reposerait sur un répertoire commun partagé par le locuteur et l'auditeur. Selon la ***théorie de la perception pour le contrôle de l'action*** (PACT, Schwartz et coll., 2002, 2012), des mécanismes imitatifs et d'apprentissage sensorimoteur permettraient chez le jeune enfant l'établissement d'une co-structuration des systèmes sensoriels et moteurs de la parole et la mise en place de cartes sensorimotrices reliant représentations sensorielles et motrices de la parole. Même à l'âge adulte, la perception de la parole pourrait alors dépendre des connaissances procédurales motrices de l'auditeur en vue de permettre d'extraire, prédire et intégrer les événements sensoriels perçus de manière cohérente.

3. Aperçu d'une sélection de modèles neurobiologiques représentatifs des liens perceptivo-moteurs dans la perception de la parole

Avant de discuter du rôle du système moteur dans la perception de la parole et de nous positionner face à ces différentes théories, il est important de s'attarder sur les bases corticales de la perception auditive de la parole à travers plusieurs modèles neurobiologiques représentatifs des liens perceptivo-moteurs dans les mécanismes de perception de la parole (voir Figure 3).

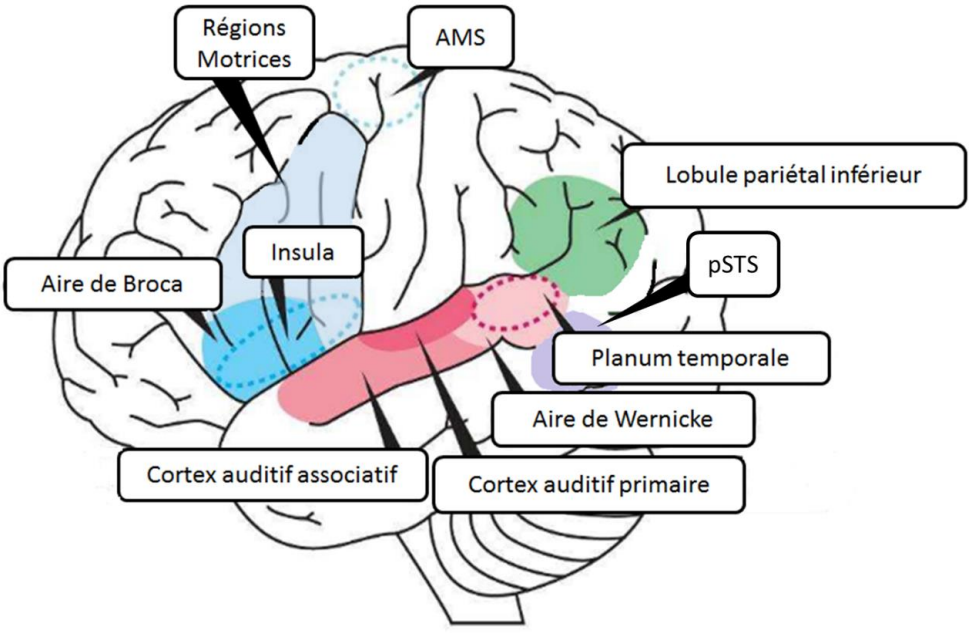

Figure 3 : Régions cérébrales impliquées dans les mécanismes de perception de la parole : Les régions auditives (en rouge ; composées du cortex auditif primaire, du cortex auditif secondaire et associatif et du planum temporale), les régions motrices (en bleu ; composées du cortex moteur primaire (M1), du cortex prémoteur ventral (PMv), de l'aire de Broca, de l'insula et de l'aire motrice supplémentaire (AMS)), le lobule pariétal inférieur (IPL ; en vert) et la partie postérieure du sillon temporal supérieur (pSTS ; en violet). Les régions en pointillés sont situées à l'intérieur des gyri, les autres sont à la surface. Figure empruntée à la thèse de Lucile Rapin (2012).

Le cortex auditif est organisé hiérarchiquement en aires primaires, secondaires et tertiaires (ou associatives) qui sont anatomiquement organisées de façon concentrique dans les parties supérieure et moyenne du lobe temporal (voir Grabski, 2012). L'aire auditive primaire est présente bilatéralement au niveau du gyrus de Heschl (bien que sa taille et sa position soient très variables entre individus). Elle permet un premier traitement, notamment fréquentiel du stimulus auditif et répond préférentiellement aux sons purs et aux variations temporelles du stimulus. Elle est entourée par les aires auditives secondaires qui répondent plutôt aux sons complexes et aux variations spectrales. Les aires auditives associatives sont situées au niveau du gyrus temporal supérieur entourant le cortex auditif primaire. Elles sont quant à elles plus spécifiques des processus de plus haut niveau comme la perception de la parole, l'attention sélective ou la mémoire auditive. La région postérieure du cortex auditif associatif (parties postérieures du planum temporale et du gyrus temporal supérieur, incluant l'aire de Wernicke) est connectée à d'autres régions sensorielles associatives (visuelles et somatosensorielles). Enfin, le sillon temporal supérieur (STS), délimitant les gyri temporaux supérieur et moyen, est également impliqué dans le traitement de la parole et de la voix ainsi que des visages pour sa partie postérieure. C'est à partir du cortex auditif que les différents modèles neurobiologiques de la perception de la parole vont tenter d'expliquer la chaîne de processus impliqués dans le décodage acoustico-phonétique du signal de parole, avec pour chacun de ces modèles une implication différentiée du système moteur.

Modèle d'analyse par synthèse de Skipper et coll. (2007 ; voir Figure 4) – Le modèle d'analyse-par-synthèse proposé par Skipper et ses collègues suppose un premier traitement du signal acoustique au niveau du cortex auditif afin de générer des hypothèses phonémiques. En cas d'ambigüités, ces hypothèses vont être envoyées au niveau du gyrus frontal inférieur pour être appariées aux buts articulatoires les plus vraisemblablement à l'origine de ces sons. Par la suite, le cortex prémoteur ventral puis le cortex moteur primaire vont simuler les commandes motrices sous-jacentes pour renvoyer les conséquences sensorielles prédites sous forme de copies d'efférence au cortex auditif. Ces copies permettront ainsi de contraindre l'interprétation des hypothèses phonémiques préalablement émises.

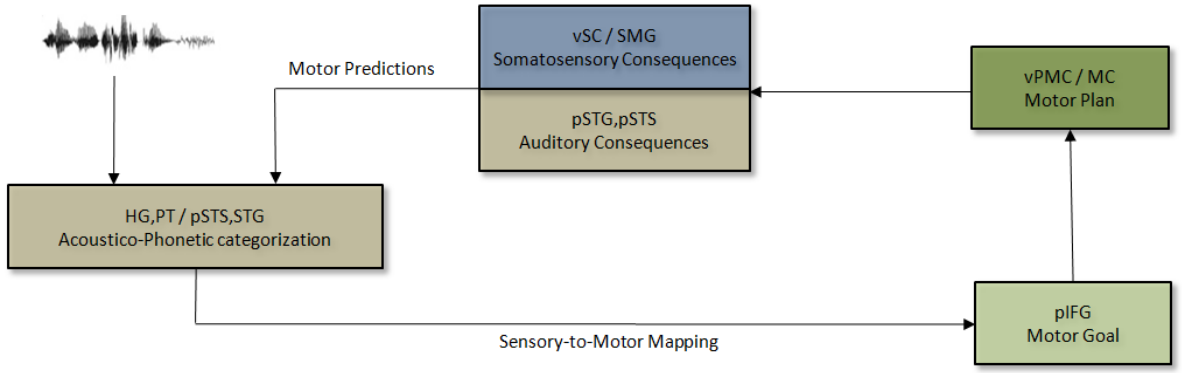

Figure 4 : Modèle d'analyse par synthèse de Skipper et collègues (2007). Figure tirée de Grabski (2012).

Modèle à double voie ventrale/dorsale de Hickok et Poeppel (2000, 2004, 2007 ; voir Figure 5) – Ce modèle s'inspire de ceux élaborés pour la perception visuelle où deux voies cortico-corticales de traitements sont différentiés. Lors de la perception d'un son, des tratiements

spectro-temporaux du signal sonore stimulerait tout d'abord le cortex auditif primaire de manière bilatérale Les sons de parole et les traitements phonologiques associés impliqueraient de plus la partie médiane des sulci et gyri temporaux supérieurs. C'est à partir de là qu'intervient la séparation en deux voies : l'une ventrale et l'autre dorsale. La voie ventrale, ou voie du « Quoi », serait localisée au niveau des parties postérieures des gyri temporaux moyen et inférieur et des parties antérieures de ces mêmes régions. Elle permettrait d'apparier les informations phonologiques avec les représentations conceptuelles lexico-sémantiques. Les régions neuronales activés par la voie dorsale sont quant à elles localisés dans les régions temporales postérieures, pariétales et frontales et apparaissent latéralisées au sein de l'hémisphère gauche. Cette voie dorsale, ou voie du « Comment », permettrait d'établir une correspondance entre les représentations auditives du cortex temporal et articulatoires du cortex moteur via une interface sensorimotrice située à la frontière pariéto-temporale (pour une revue récente, voir Schwartz et al., 2011). Ce modèle à deux voies montre ainsi qu'il y aurait un traitement en parallèle des aspects sensorimoteurs et lexico-sémantiques de la parole. De manière importante, les auteurs donnent un rôle principal à la voie dorsale uniquement durant le développement de la parole et/ou l'acquisition de nouveaux mots. En effet, la voie dorsale permettrait à l'enfant de stocker des représentations sensorielles pour pouvoir les comparer avec ses productions, lui permettant ainsi de composer son répertoire moteur en ajustant ses gestes à chaque essai de production articulatoire. Cependant les auteurs n'attribuent pas un rôle causal du système moteur dans les mécanismes de perception de la parole.

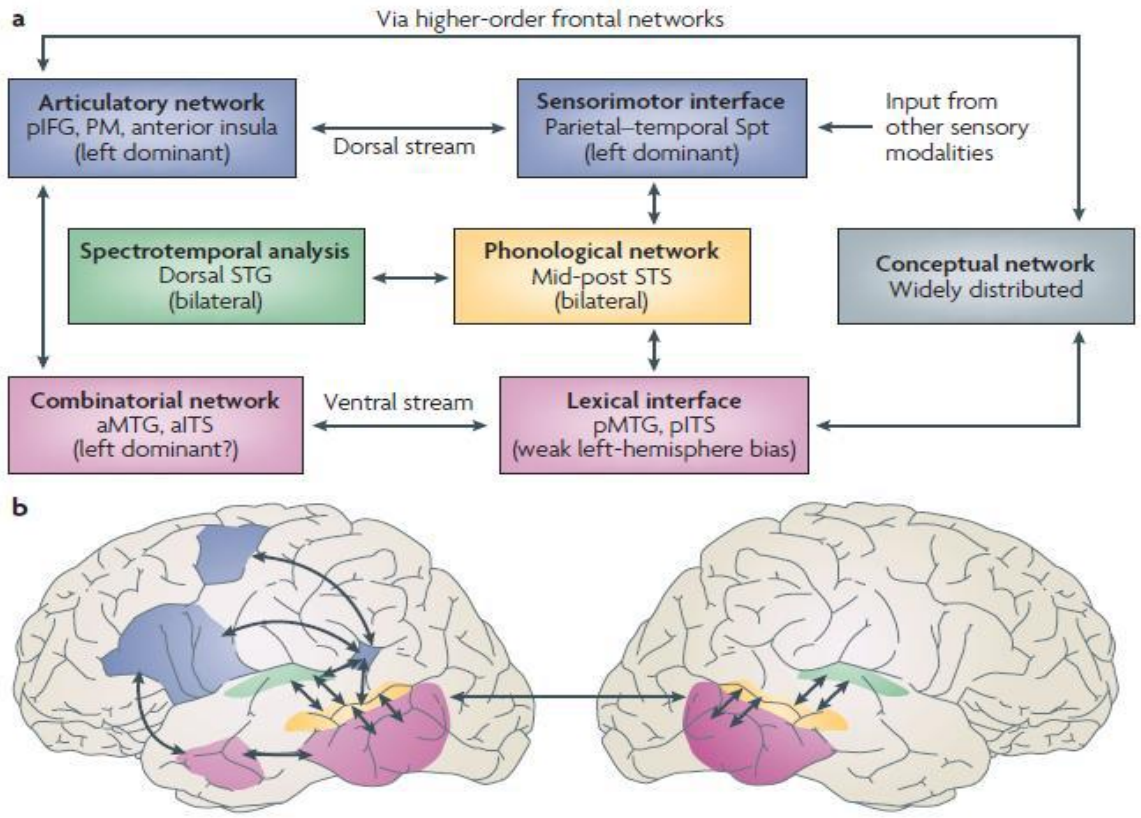

Figure 5 : Schéma du modèle à double voie de Hickok et Poeppel (2007). En rose, on peut voir la voie ventrale, ou voie du « Quoi » et en bleu, il s'agit de la voie dorsale ou voie du « Comment ».

Modèle à double voie antérieure/postérieure de Scott, Rauschecker et collègues (2003, 2009, 2011 ; voir Figure 6) – Inspiré d'un modèle de perception auditive et visuelle chez le

primate non-humain, Scott, Rauschecker et collègues ont établi un modèle de perception auditive de la parole chez l'humain. Sur un principe similaire au modèle de Hickok et Poeppel (2007), ces chercheurs ont proposé l'existence de deux voies, l'une antérieure (la voie du « Quoi ») qui jouerait un rôle dans l'identification d'objets auditifs et dans la compréhension de la parole intelligible, l'autre plus postérieure (la voie du « Où ») permettrait de traiter la question de la localisation des sources auditives. La voie du « Quoi » reflèterait un traitement acoustique de bas niveau du signal auditif qui débuterait dans le gyrus de Heschl, suivi d'un processus de décodage acoustico-phonétique au niveau de la partie antérieure du gyrus temporal supérieur ainsi que dans les régions frontales inférieures contenant a priori les représentations invariantes des différentes catégories phonétiques. De là, un appariement entre les représentations phonétiques et les représentations motrices serait effectué dans le cortex prémoteur ventral et suite à une simulation des commandes motrices, des copies d'efférence contenant les conséquences sensorielles de la simulation seraient envoyées vers le lobule pariétal inférieur (IPL). Cette région pariétale jouerait un rôle d'interface sensorimotrice, afin de permettre l'intégration des représentations auditives et motrices particulièrement lors d'une perception du signal auditif rendue difficile (bruit, sons étrangers à notre langue maternelle ou mots rares ou peu fréquents). Cette interface interviendrait en sens inverse de la voie antérieure, permettant une comparaison entre les copies d'efférences envoyées depuis les centres moteurs et les informations sensorielles réelles. Ce processus permettrait ainsi de désambiguïser le message linguistique perçu.

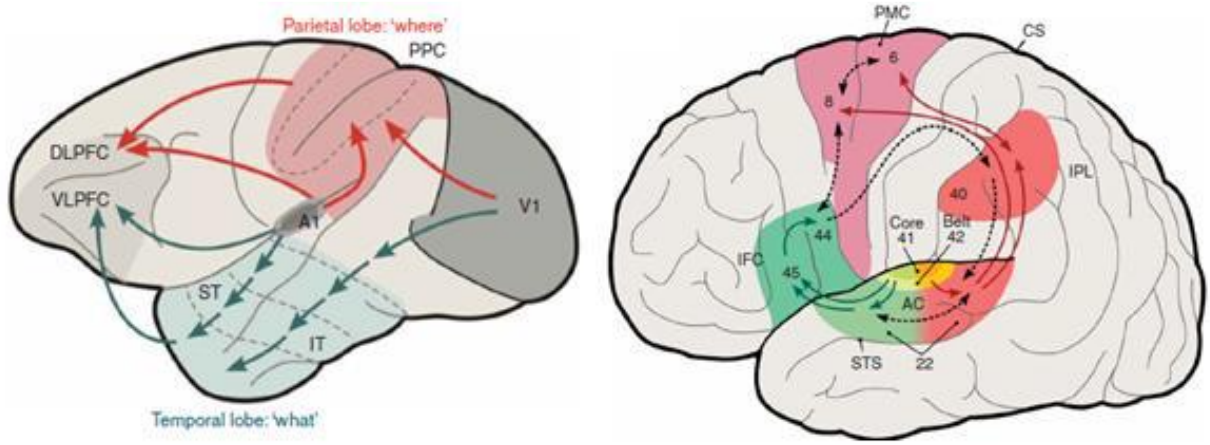

Figure 6 : ***(Gauche) Double voies visuelles et auditives chez le singe. (Droite) Modèle de double voie + de Rauscheker et Scott (2009) appliqué à la perception auditive de la parole.***

Dans leur ensemble, ces trois modèles neurobiologiques s'opposent à la fois à une conception purement auditive de la parole qui supposerait un traitement basé uniquement sur les propriétés acoustiques issues du signal auditif, mais également à une conception purement motrice qui supposerait un recodage uniquement moteur des sons perçus. D'autre part, à des niveaux différents, chacun de ces modèles questionne le possible rôle fonctionnel du système moteur dans des conditions d'écoute dites « normales ». En effet, pour Hickok et Poeppel le système moteur jouerait un rôle dans l'acquisition de la parole, d'une seconde langue ou dans le cas de tâches méta-phonologiques, mais n'aurait pas de rôle causal véritable en perception. Tandis que pour Skipper et collègues et Rauscheker et Scott, le système moteur serait plus fortement impliqué lors de la perception de stimuli ambigus ou bruités.

4. Le système moteur en perception de la parole – résultats empiriques

Malgré la différence dans le degré d'importance attribué au système moteur dans les mécanismes de perception de la parole décrits ci-dessus, les trois modèles neurobiologiques présentés précédemment s'accordent pour supposer un lien étroit entre les systèmes de perception et de production de la parole. Des études en imagerie par résonance magnétique fonctionnelle (IRMf) ont permis de confirmer l'activation des régions motrices lors de la présentation de stimuli langagiers, mais c'est à travers les études en stimulation magnétique transcrânienne (TMS) qu'on a pu étudier le rôle causal du système moteur en perception. Nous allons donner un aperçu de cette littérature dans les sections suivantes.

Activations motrices en perception – Des études en IRMf ont mis en évidence une activation des régions motrices (la partie postérieure du gyrus frontal inférieur (IFG), le cortex prémoteur ventral (PMv) et le cortex moteur primaire (PMC)) et proprioceptives (le cortex somatosensoriel) lors de la présentation auditive, visuelle ou audio-visuelle de stimuli de parole (par exemple Calvert et Campbell, 2003; Callan et al., 2003, 2004; Möttonen et al., 2004; Wilson et al., 2004; Ojanen et al., 2005; Pekkola et al., 2005; Skipper, Nusbaum et Small, 2005; Pulvermüller et al., 2006; Wilson et Iacoboni, 2006; Skipper et al., 2007; Callan et al., 2010; Tremblay et Small, 2011). De plus, un recrutement plus important du système moteur a été observé lors de tâches de perception rendues difficiles, par du bruit par exemple (Binder et al. 2004; Zekveld et al., 2006), lors de la présentation de phonèmes n'appartenant pas à la langue maternelle du participant (Callan et al., 2004; Wilson et Iacoboni, 2006) ou encore pour des stimuli audiovisuels incongruents par rapport à des stimuli congruents (Ojanen et al., 2005; Skipper et al., 2007).

Grâce à l'enregistrement des réponses musculaires (par électromyographie ou EMG) effectué sur les muscles linguaux ou labiaux lors d'une stimulation magnétique transcrânienne à impulsion unique, des chercheurs ont pu mettre en évidence un phénomène de « résonance motrice ». Lors de la stimulation du cortex moteur primaire au moment de la perception auditive ou visuelle d'un stimulus langagier, des'enregistrements EMG ont ainsi montré une augmentation des potentiels évoqués moteurs labiaux par rapport à la présentation de stimuli non langagiers (Watkins, Strafella et Paus, 2003). D'autres mesures similaires ont été réalisées sur les muscles de la langue et ont contrastés différents stimuli de parole dont la réalisation impliquait prioritairement les la langue ou les lèvres (Sundara, Namasivayam et Chen, 2001; Fadiga et al., 2002; Watkins et Paus, 2004; Roy et al., 2008; Sato et al., 2010). Fadiga et collègues (2002) ont ainsi montré pour la première fois que l'écoute de stimuli de parole produisait une activation dans les régions motrices du phonème cible. Leurs résultats montrent en effet, que suite à une impulsion magnétique envoyée au niveau du cortex moteur gauche, une augmentation des potentiels évoqués moteurs enregistrés au niveau de la langue était observée lorsque le sujet écoutait des sons de parole impliquant prioritairement des mouvements de la langue pour les produire par rapport à d'autres sons de parole (comme « rr » qui nécessite une forte implication de la langue par rapport à « ff »).

D'autre part, une organisation somatotopique du cortex moteur primaire et du cortex prémoteur ventral a été observée à travers plusieurs expériences en IRMf ou par stimulation magnétique transcrânienne à impulsion unique (Fadiga et al., 2002; Pulvermüller et al., 2006; Skipper et al., 2007; Roy et al., 2008; Sato et al., 2010). L'activité au sein de ces régions

serait fonction des effecteurs utilisés pour produire les gestes de parole perçus auditivement ou visuellement.

Rôle causal du système moteur – Dans la section précédente, nous avons établi un rôle fonctionnel des régions motrices, mais les études présentées ne permettaient pas de définir le rôle causal du système moteur dans les mécanismes de perception de la parole. Certaines études TMS ont cependant permis de répondre partiellement à cette question. Cette technique a la particularité de pouvoir agir directement sur l'activité cérébrale en inhibant ou au contraire en excitant une région grâce à l'envoi d'impulsions magnétiques à la surface du scalp. De cette façon, il est par exemple possible de créer une lésion temporaire pour vérifier l'implication de la région lésée dans telle ou telle tâche, et notamment lors de la perception de parole.

Ainsi, plusieurs études en TMS répétitives ont pu mettre en évidence une perturbation des capacités du sujet dans des tâches phonologiques « complexes », nécessitant un recrutement important des processus de mémoire de travail verbale ou de segmentation (Boatmann, 2004; Nixon et al., 2004; Romero et al., 2006; Sato et al., 2009). En revanche, lors de tâches "simples" comme l'identification syllabique par exemple, seule la reconnaissance des stimuli auditifs bruités ou ambigus a été impactée par une modulation temporaire de l'activité neuronale motrice (Meister et al., 2007; d'Ausilio et al., 2009; Möttonen et Watkins, 2009; d'Ausilio et al., 2011, 2012). D'Ausilio et collègues (2009) ont également observé une double dissociation dans les résultats obtenus, liée à la zone stimulée (région de la langue ou des lèvres dans le cortex moteur primaire) et aux stimuli présentés (impliquant la langue ou les lèvres), avec de meilleurs temps de réaction et un taux d'erreur moindre lorsque la région stimulée était relative à l'effecteur utilisé pour produire la syllabe. De façon similaire, Möttönen et Watkins (2009) ont montré qu'une inhibition de la région relative aux mouvements des lèvres biaisait la perception catégorielle de stimuli présentés sous forme d'un continuum (/ba/-/da/ ou /pa/-/ta/) vers la syllabe n'impliquant pas de mouvements labiaux (tandis que cette perturbation labiale ne générait au contraire aucune perturbation sur les continua /ka/-/ga/ ou /da/-/ga/ n'impliquant pas d'action des lèvres).

Pris ensemble, ces résultats sont en accord avec les modèles neurobiologiques présentés dans la section précédente. Bien qu'aucune étude n'ait montré de résultats probants quant à son recrutement en conditions réellement écologiques, ils démontrent une implication fonctionnelle et causale du système moteur en fonction des articulateurs utilisées lors de tâches complexes ou lors de la présentation de stimuli ambigus ou bruités. Cela suggère peut-être plutôt un rôle de soutien qu'un rôle *majeur* du système moteur dans les mécanismes de perception de la parole.

PARTIE THÉORIQUE - B

LA NATURE *AUDIOVISUELLE* DE LA PERCEPTION DE LA PAROLE

La partie précédente traitait de la nature sensorimotrice de la parole. A travers quelques découvertes et théories clés, nous avons discuté d'un possible couplage perception-action lors de la perception de la parole. Nous avons ensuite illustré, à l'aide de différents modèles neurobiologiques, les interactions entre les entrées sensorielles et les connaissances procédurales motrices au niveau neuroanatomique. Finalement, nous avons montré un certain nombre de résultats empiriques en faveur d'un rôle fonctionnel et causal de notre système moteur dans les mécanismes de perception de la parole. Nous allons dans cette seconde partie discuter de la nature audio-visuelle de la perception de la parole en nous intéressant dans un premier temps aux indices visuels exploitables pour percevoir un signal de parole, puis nous verrons que les informations visuelles et auditives sont complémentaires et que l'ajout de la modalité visuelle au signal acoustique de parole vient influencer le traitement auditif, en le facilitant ou au contraire en le perturbant. Puis nous aborderons les mécanismes d'intégration multisensorielle lors de la perception de la parole à travers la description des régions cérébrales impliquées.

1. La parole visuelle

Dans une situation bruyante les informations auditives peuvent ne plus être suffisantes pour comprendre le message que veut nous transmettre notre interlocuteur. Nous devons alors utiliser d'autres informations disponibles pour désambiguïser le signal perçu, par exemple en ayant recours à la lecture labiale. Mais afin d'expliquer plus précisément ce que nous pouvons réellement « lire sur les lèvres », il convient de rappeler l'origine du son que nous tentons de décoder. Nous verrons que les indices visuels peuvent aussi être porteurs d'ambiguïtés, que les lèvres ne sont pas les seuls indices visuels disponibles sur notre visage pour nous aider à déchiffrer le message linguistique et, finalement, que dans certains cas particuliers, lors d'une perte sévère de l'audition par exemple, une aide manuelle peut être ajoutée au signal visuel de parole pour améliorer la compréhension.

Bref aperçu des articulateurs et résonateurs utilisés dans la production de la parole – Pour comprendre ce que nous percevons, nous devons d'abord comprendre ce que nous produisons et surtout comment. Pour générer un son de parole, trois éléments sont nécessaires : une soufflerie (origine du flux d'air, les poumons), un générateur de source sonore (principalement les plis vocaux situés au sein du larynx) et des résonateurs (cavités buccale et nasale) dont les géométries sont contrôlées par des articulateurs (langue, mâchoire, lèvres, voile du palais). Dans la production de parole, les poumons permettent de générer le flux d'air nécessaire notamment à la mise en vibration des cordes vocales. Le larynx contient les plis vocaux qui, sous certaines conditions de tension musculaire, de degré d'ouverture et de pression sous-glottique, vont entrer en vibration et permettre la production de sons modulables en intensité, en hauteur et en timbre. Cependant, tous les sons de parole ne requièrent pas forcément la vibration des plis vocaux, d'autres types de

sons peuvent être émis comme les bruits de plosion (ouverture rapide du conduit vocal qui se traduit par une modification importante mais brève de la pression orale, donnant lieu à un son au contenu fréquentiel large mais très bref) ou de frication (passage de l'air dans une constriction très marquée du conduit vocal donnant naissance à un écoulement turbulent générant un bruit aléatoire au contenu fréquentiel large et étalé dans le temps). C'est notamment pour cela que nous distinguons les sons voisés (comme les voyelles ou les plosives voisées comme /b/ ou /d/, provoqués par une mise en vibration des cordes vocales suivie d'un bruit de plosion) des sons non voisés (comme les consonnes /p/ ou /t/ par exemple, produites uniquement par un bruit de plosion sans vibration des cordes vocales). Le terme de « conduit vocal » désigne communément l'ensemble des résonateurs et articulateurs qui permettent de moduler le contenu spectral des sons émis. En modifiant la position des articulateurs de manière contrôlée, nous modifions la forme du conduit vocal, ce qui permet la production d'une multitude de sons de parole pour communiquer (pour plus de détails sur l'anatomie du conduit vocal voir Perrier et Schwartz, 2016 ; Vilain, 2002).

La perception que nous avons de ces gestes – Si nous regardons rapidement l'alphabet phonétique international (ou API), près de 200 sons ont été répertoriés, chacun faisant référence à un phonème présent dans au moins une des 7000 langues ou dialectes parlés dans le monde. Mais lors de la réalisation de ces sons de parole, l'ensemble du conduit vocal n'est pas toujours visible et seuls les lèvres, l'apex de la langue, une partie des dents et la partie inférieure de la mâchoire peuvent être observables. Ce sont ces mouvements visibles qui nous permettent de « lire » sur les lèvres. Ce procédé, aussi appelé « lecture labiale », est employé par tout un chacun même si nous ne sommes pas tous dotés des mêmes capacités de lecture. De plus, tous les phonèmes de notre langue ne peuvent pas se distinguer à la seule vue des mouvements labiaux (Fisher, 1968; Summerfield, 1987, 1991), ce sont les homonymes labiaux. Certains par exemple n'ont de différent que le trait de nasalité (expulsion de l'air par le nez et/ou la bouche comme pour /p/ vs. /m/), ou bien le voisement (i.e., la mise en vibration ou non des plis vocaux comme pour /p/ vs. /b/), tandis que d'autres encore sont produits trop à l'arrière du conduit vocal pour être facilement identifiés visuellement comme /k/. Ainsi, d'après Bernstein et al. (2000), seuls 40% à 60% des phonèmes d'une langue peuvent être discriminés sur la base de l'information visuelle seule, et 10 à 20% des mots. Plusieurs études ont été consacrées à l'établissement d'un classement des phonèmes sur la base de leur intelligibilité visuelle. Fisher (1968) par exemple a été le premier à introduire le terme de « visème » (contraction de « visuel » et « phonème », autrement dit un équivalent visuel du phonème) pour désigner une unité minimale perçue distinctement d'une autre sur la base de l'information visuelle seule. On parle alors de « saillance perceptive » (Summerfield, 1987) pour distinguer un phonème très bien perçu d'un phonème confondu. Ainsi, d'après Gentil (1981), les phonèmes consonantiques les mieux perçus en position initiale sont ceux articulés à l'avant du conduit vocal comme /p/-/b/-/m/ (« visème bilabial », nous regroupons ainsi trois consonnes non distinguables en termes de mode d'articulation, voisé, non voisé, nasal, mais partageant le même lieu d'articulation), ou, /f/-/v/ (« visème labiodental ») ou ayant un trait visible au niveau des lèvres comme /ʃ/-/ʒ/. A l'inverse, les phonèmes les plus confondus sont ceux articulés à l'arrière du conduit vocal comme [/s/, /z/, /t/, /d/, /n/, /k/, /g/, /ŋ/, /ʁ/].

Ce que le reste du visage nous apprend – Nous avons jusqu'ici mentionné exclusivement les lèvres comme source d'information visuelle nécessaire pour décrypter le message transmis. On emploie l'expression « lire sur les lèvres » et on désigne souvent la perception visuelle

d'un signal de parole comme de « la lecture labiale », or les dents (McGrath, 1985; cité par Summerfield, 1991; Thomas et Jordan, 2004), la mâchoire (Guiard-Marigny, et al., 1996; Vatikiotis-Bateson, et al., 1998) et la langue (Badin, et al., 2010) présentent également une part d'information non négligeable pour comprendre visuellement un message. De plus, dans une situation conversationnelle, il est extrêmement rare de ne voir que la bouche de notre locuteur, c'est son visage en entier (ou partiellement caché parfois, voir Jordan et Thomas, 2011) qui nous fait face. Ainsi, Thomas et Jordan en 2004 ont montré que les performances en lecture labiale étaient meilleures lorsque le visage entier du locuteur était présenté par rapport à la région buccale seule. En effet, les joues (Preminger, et al., 1998), le haut de la tête (Munhall et Vatikiotis-Bateson, 1998 ; Cvejic et al., 2010) et plus particulièrement les yeux (Vatikiokis-Bateson et al., 1998) joueraient également un rôle dans la perception visuelle de la parole.

Une aide à la lecture labiale : le LPC – Mais tous ces indices précédemment cités ne sont pas toujours suffisants pour décoder le signal de parole. Comme nous l'avons mentionné précédemment, certains homonymes labiaux demeurent. Pour les personnes malentendantes ou sourdes qui sont confrontées chaque jour aux problèmes que pose la lecture labiale lors de la compréhension d'un message oral, il existe le Langage Parlé Complété (LPC ou "Cued Speech")). Il s'agit d'un code complémentaire à la lecture labiale, développé par le docteur Orin Cornett (1994), qui permet d'ôter les ambiguïtés dues aux homonymes labiaux. Le principe consiste à associer à chaque phonème un geste effectué par la main près du visage afin de compléter le message. La combinaison de la forme des doigts (en français, huit configurations différentes pour les consonnes) et du placement de la main par rapport au visage (en français, cinq positions différentes pour les voyelles) permet une représentation complète de la langue parlée. Pour reconnaître un phonème sans difficulté, il faut donc associer l'image labiale et la clé manuelle. L'unité de base du LPC est la syllabe CV (Consonne-Voyelle), la parole est donc codée syllabe par syllabe (pour de plus amples informations, voir la thèse d'Attina, 2005).

Dans cette partie, nous avons ainsi pu constater qu'une quantité importante d'informations peut être extraite à partir du signal visuel, du fait de la visibilité de certains articulateurs et de la présence de mouvements supplémentaires jusque sur la partie haute de la tête, faisant ainsi du visage entier une source d'indices pour décoder le signal de parole lorsque le son est absent. Cependant, malgré notre étonnante capacité à lire sur les lèvres, certains homonymes labiaux nous empêchent de percevoir la totalité des phonèmes de notre langue. Une information visuelle supplémentaire peut alors être ajoutée sous forme de code manuel afin de désambiguïser l'information. Mais le signal sonore, ajouté au visuel, reste la modalité perceptive la plus efficace. Nous allons voir dans une seconde partie la complémentarité, voire la compétition qui existe entre ces deux modalités, notamment dans des situations complexes et/ou ambigües.

2. Voir pour mieux entendre

Dans notre quotidien, il est rare de pouvoir tenir une conversation sans qu'aucun bruit ne vienne parasiter l'échange (bruit des voitures, bruit du vent dans les arbres ou de la pluie sur le velux, ronronnement d'un ordinateur, murmures des élèves, radio, télévision…). Nous pouvons tendre l'oreille, mais il s'agirait plutôt ici d'ouvrir les yeux pour avoir accès à toutes les informations disponibles qui pourraient venir compléter le signal sonore manquant par endroit. Nous allons ainsi détailler l'apport du visuel en milieu bruité, mais nous verrons que

les difficultés rencontrées lors de la perception auditive ne sont pas forcément inhérentes à un bruit extérieur, mais bien souvent au signal de parole lui-même. Nous montrerons également que le signal visuel n'est pas qu'une aide en cas de difficulté, mais qu'il apporte des informations complémentaires qui facilitent la perception auditive même lors de bonnes conditions d'écoute. Finalement, à travers l'exemple de l'effet McGurk, nous verrons que ces deux flux sensoriels peuvent entrer en compétition et induire des illusions perceptives.

La contribution de l'entrée visuelle en milieu bruité – Cotton, en 1935, a été le premier à mettre en évidence la contribution de la modalité visuelle lors de la perception de la parole, conforté, par la suite, par les travaux de Sumby et Pollack en 1954, qui ont proposé de quantifier cet avantage perceptif audiovisuel. Comme vu précédemment, notre capacité à lire sur les lèvres est certainement ce qui explique que nous soyons capables de comprendre en partie un message même lorsque le signal auditif est masqué par du bruit. Cependant, l'apport des informations visuelles n'est pas illimité, puisqu'il existe un niveau de bruit maximal au-delà duquel l'aide visuelle n'est plus suffisante pour désambiguïser le message. Il existe ainsi un niveau de bruit intermédiaire pour lequel le gain audio-visuel est maximum. En effet, Ross et collègues (2007) ont testé la reconnaissance de mots monosyllabiques en condition auditive seule, audio-visuelle et visuelle seule en présence de plusieurs niveaux de bruit différents. Ici, le bénéfice lié à la présence du signal visuel est maximum lorsque le rapport signal sur bruit ("Signal-to-Noise Ratio" ou SNR) est intermédiaire, soit équivalent dans cette étude à -12dB. A ce niveau de bruit, l'ajout des informations visuelles permet d'améliorer de 45% les scores de reconnaissance en modalité auditive seule (environ 20%). Lorsque le SNR est plus réduit (jusqu'à -24dB), c'est-à-dire que le niveau de bruit est plus important, le gain lié au signal visuel diminue. Seuls 19% des mots sont reconnus en perception audiovisuelle, contre 0% en perception auditive seule en présence d'un SNR égal à -24dB. De manière très intéressante, leur étude suggère également que l'information linguistique extraite par le participant en condition audio-visuelle est supérieure à la somme des indices que nous sommes capables de récupérer dans les conditions auditive seule et visuelle seule (voir également Gagné et al., 2002 ; Schwartz et al., 2004). Cela montre la grande efficacité des mécanismes d'interaction audio-visuelle en perception de la parole.

La complémentarité des deux flux sensoriels auditif et visuel – Il faut également préciser le rôle complémentaire des indices visuels et auditifs. En effet, bien que le signal sonore seul soit perçu et compris dans sa totalité alors que seuls 10 à 20% des mots sont perçus en lecture labiale seule, les traits présents dans le signal visuel ne sont pas redondants avec les informations présentes dans le flux auditif de parole (Summerfield, 1987). Certaines caractéristiques, comme le lieu d'articulation du phonème, seront très rapidement masquées par du bruit tandis qu'elles resteront très saillantes visuellement (/p/ vs. /t/ par exemple). A l'inverse, le voisement, dû à la mise en vibration d'un organe non visible, le larynx, est auditivement saillant et robuste au bruit (/p/ vs. /b/ par exemple). Des résultats similaires ont été montrés (voir Benoît et al., 1994) pour des phonèmes vocaliques : une hiérarchie perceptive en condition auditive seule est observée : /a/ > /i/ > /y/ du fait d'une intensité propre à chacune des voyelles (/a/ a une intensité supérieure à /i/ qui a elle-même une intensité plus forte que celle de /y/) et cette hiérarchie est différente en perception audio-visuelle très bruitée (SNR de -24dB), /y/ > /a/ > /i/ du fait d'une information labiale plus stable pour la voyelle /y/ (arrondissement des lèvres moins impacté par l'environnement consonantique que la voyelle /a/ et plus encore la voyelle /i/, plus

dépendante des effets de coarticulation). Ces résultats supportent l'idée d'une complémentarité des informations auditives et visuelles.

Lorsque la difficulté provient du signal sonore lui-même – Mais le bruit extérieur n'est pas la seule difficulté rencontrée lors de la réception du message. Certaines caractéristiques propres au locuteur ou au contenu du message peuvent mettre à mal notre compréhension. Reisberg et collègues (1987) ont ainsi montré qu'en condition auditive seule, la compréhension d'un contenu sémantique complexe pouvait s'avérer très moyenne. Mais grâce aux indices visuels, la compréhension s'en trouvait nettement améliorée. En 2001, Arnold et Hill ont également testé ces différentes situations : la perception d'un discours sémantiquement et syntaxiquement complexe en condition auditive seule et audio-visuelle, la perception d'un texte français par des participants de langue maternelle anglaise apprenant le français, et finalement la perception d'un texte anglais prononcé par un locuteur anglais ayant un accent différent des participants. Pour chacune de ces situations un bénéfice lié à la présence de l'information visuelle a pu être observé, montrant ainsi qu'en l'absence de bruit, la lecture labiale peut aider à désambiguïser un signal auditif complexe même si celui-ci n'est pas détérioré physiquement. Des travaux complémentaires ont été réalisés par Navarra et Soto-faraco (2005) sur une possible amélioration de la perception d'une langue seconde étrangère lors de l'ajout des informations visuelles. Des contrastes phonétiques en catalan présentés à un groupe bilingue catalan-espagnol dominant ont effectivement été mieux perçus lorsque les stimuli étaient présentés audio-visuellement.

L'illusion McGurk – Bien que l'apport de la modalité visuelle soit plus flagrant en milieu bruité, des études suggèrent que la modalité visuelle est systématiquement utilisée pour la perception de la parole. Cela suppose l'existence de mécanismes d'intégration des flux auditif et visuel. Le caractère spontané et irrépressible de la lecture labiale, même dans un contexte d'écoute confortable, a été mis en évidence lors de la présentation simultanée d'un signal auditif et d'un signal visuel incongruents. En effet, la présentation d'un message parfaitement audible en même temps que des mouvements articulatoires correspondant à un message différent conduit souvent l'auditeur à percevoir un percept qui ne correspond pas à l'information auditive et qui intègre des traits de l'information visuelle sans pour autant correspondre non plus complètement au signal visuel de parole. Ainsi, plusieurs types de percepts ont été obtenus, résultant soit d'une fusion (/ba/ auditif + /ga/ visuel = /da/), soit d'une combinaison (/ga/ auditif + /ba/ visuel = /bga/) des deux signaux. Cette perception présente une grande variabilité entre les participants (voir Sekiyama et al., 1991, 2003 pour une expérience sur des participants japonais visiblement moins sensibles à cet effet). Cette illusion perceptive, appelée « effet McGurk » (McGurk et MacDonald, 1976), démontre une utilisation de l'information visuelle présente même lorsque le signal auditif est clair et non ambigu et suggère l'utilisation de mécanismes d'intégration audio-visuelle (nous les décrirons plus en détails dans la section suivante). Ce phénomène a été largement étudié dans la littérature (pour une revue, voir Colin et Radeau, 2003) et plusieurs expériences ont montré qu'il était possible d'influencer cet effet, en l'amplifiant grâce à l'ajout de bruit par exemple (Sekiyama et al., 2003) ou au contraire en le diminuant en perturbant l'attention des participants (par exemple, Alsius et al., 2005, 2007).

A travers ces exemples, nous avons montré le rôle crucial de la lecture labiale dans la perception d'un signal sonore. Bien que seule elle ne permette pas de comprendre la totalité du message, ajoutée à un signal auditif bruité, ambigu ou complexe, elle apporte des

informations complémentaires pour reconstruire et ainsi comprendre le message linguistique transmis. Elle permet également d'améliorer la compréhension alors que la situation conversationnelle est confortable et dans certains cas, elle peut même nous jouer des tours tant elle est spontanée et irrépressible. Cela conforte l'idée d'une parole non plus simplement auditive, mais bien audio-visuelle. Dans la prochaine section, nous allons nous intéresser aux mécanismes cérébraux qui permettent à ces deux flux d'informations sensorielles différents d'être intégrés pour former un percept unifié.

3. Régions cérébrales impliquées dans la perception et l'intégration audiovisuelle de la parole

Plusieurs patrons d'intégration ont été proposés dans la littérature afin de déterminer le module de traitement au niveau duquel s'effectue l'intégration des informations auditives et visuelles en un percept unifié. D'après Schwartz et collaborateurs (1998), il existerait quatre modèles envisageables de fusion audio-visuelle de la parole (voir Figure 7) :

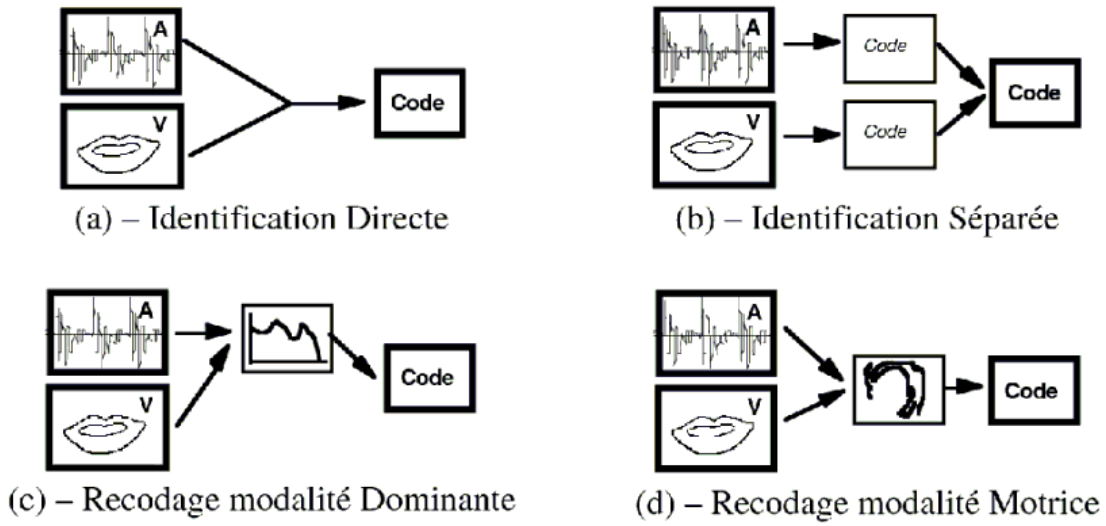

Figure 7 : Quatre principaux modèles possibles d'intégration audio-visuelle pour la perception de la parole d'après Robert-Ribes, Schwartz et Escudier (1995)

un premier modèle dans lequel la fusion audio-visuelle et l'identification phonétique se fait directement sans traitement préalable des informations unimodales ('identification directe'), un second modèle dans lequel la classification phonétique se fait séparément dans les deux modalités, la fusion s'opérant après la classification ('identification séparée'), un troisième modèle dans lequel l'entrée visuelle est recodée de manière précoce dans la modalité auditive, dite modalité dominante ('recodage dans la modalité dominante') et enfin un quatrième modèle dans lequel les deux modalités sont recodées vers la modalité motrice, l'identification phonétique s'effectuant sur la base des caractéristiques articulatoires de la représentation commune obtenue ('recodage commun des deux entrées sensorielles vers la modalité motrice').

Face à ces quatre modèles théoriques d'intégration audio-visuelle, les études de neuroimagerie ont depuis permis de préciser la localisation des régions cérébrales impliquées dans ces mécanismes et le décours temporel de l'intégration des différentes informations sensorielles, c'est-à-dire à quel moment se fait le traitement et l'unification des

deux flux. L'aspect temporel sera traité dans la partie C-2 de ce manuscrit. Dans cette section nous n'aborderons que la question des régions cérébrales impliquées dans ces traitements. Des études en imagerie par résonance magnétique fonctionnelle (functional magnetic resonance imaging, fMRI) ont, en effet, pu mettre en évidence l'existence de régions spécifiquement impliquées dans les processus d'intégration audio-visuelle de stimuli de parole

Le pSTS/pSTG – La partie postérieure du STS/STG, à la jonction entre les régions auditives et visuelles, semble être le "berceau" de l'intégration audio-visuelle de stimuli de parole. En effet, une activation a été observée, principalement dans l'hémisphère gauche, lors de la perception de parole aussi bien auditive et visuelle qu'audio-visuelle (Calvert et al., 1997, 2000; Callan et al., 2003, 2004; Calvert et Campbell, 2003; Jones et Callan, 2003; Skipper et al., 2005, 2007). D'autre part, et en accord avec les recherches effectuées sur la région homologue du cerveau de primates non-humains, une parcellisation fonctionnelle de cette région a été démontrée chez l'homme en réponse à des stimuli auditifs, visuels et tactiles (Beauchamps et al., 2004a, 2004b , 2005).

De plus, Calvert et collaborateurs (2000) ont montré que le pSTS était plus activé lors de la présentation de stimuli bimodaux, par rapport à la somme des activités observées lors de la présentation de stimuli unimodaux auditifs et visuels (réponse supra-additive). Inversement, ils ont également observé une diminution de l'activité de cette région dans le cas de stimuli audio-visuels incongruents (réponse sous-additive ; Calvert et al., 1997, 2000). Ces phénomènes de supra- et sous-additivité démontrent que les signaux auditifs et visuels ne sont pas traités séparément puis unifiés, puisqu'on aurait alors une activité similaire entre l'activité liée à la présentation AV et la somme des activités lié à A + V. Or une différence d'activation démontre bien qu'un traitement spécifique intervient dès lors que les deux percepts sont présentés simultanément. Les travaux de Beauchamp et collègues en TMS sont venus conforter l'hypothèse du rôle crucial du pSTS dans les mécanismes de fusion des flux audio-visuels de parole. En effet, ils ont montré que l'envoi d'une impulsion unique au niveau du pSTS interférait avec la perception de stimuli incongruents du type McGurk, sans pour autant entraver la perception des stimuli audio-visuels congruents (Beauchamp et al., 2010).

Cette modulation des réponses neuronales serait en partie due à des mécanismes de projection puis de rétropropagation de l'information entre les aires primaires, secondaires et associatives au sein des différentes régions auditives et visuelles (Möttönen, 2004; Hertrich et al., 2007). Selon Senkowski et collègues (2008, voir Figure 8), un des scénarios possibles pour expliquer ces modulations serait l'existence d'interactions ascendantes et descendantes entre les régions uni-sensorielles (auditives et visuelles) et intégratives (régions pariétales et temporales supérieures) pour parvenir à une cohérence/synchronisation neuronale nécessaire à l'intégration des différentes sources sensorielles.

Ainsi, l'implication du pSTS/STG en réponse à des stimuli unimodaux et audio-visuels et les phénomènes de supra- et sous-additivité lors de la perception de parole audio-visuelle laissent à penser que le pSTS serait le lieu principal d'intégration des signaux auditifs et visuels de parole.

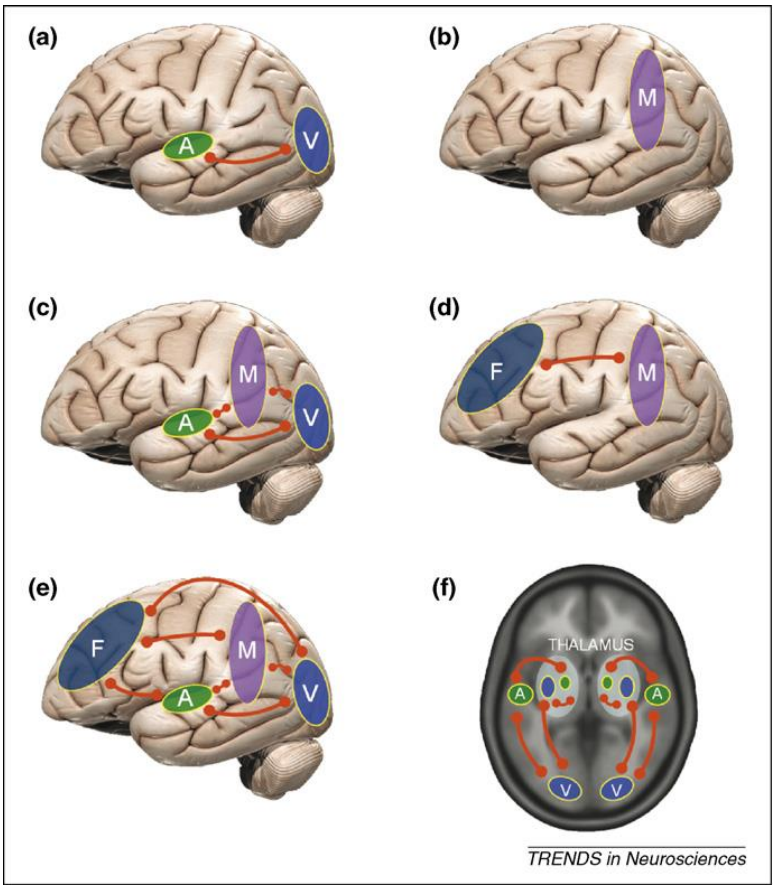

Figure 8 : ***Scénarios possibles pour le liage multimodal à travers la cohérence neuronale. (a) le scénario le plus simple prédit une synchronisation neuronale entre les régions sensorielles de bas niveaux (le cortex auditif (A) et le cortex visuel (V) ici). (b) Une autre possibilité est d'envisager que ce changement de cohérence neuronale pourrait avoir lieu dans les régions intégratives de plus haut niveaux (régions pariétales ou temporales supérieures (M)). (c) Ces changements dans la synchronisation des oscillations cérébrales des régions sensorielles (A et V) pourraient être associés à l'augmentation de l'activité oscillatoire des régions intégratives (M). Ce changement pourrait être le reflet d'interactions ascendantes et descendantes entre les régions uni- et multi-sensorielles. (d et e). De manière plus générale, l'existence d'interactions beaucoup plus complexes pourrait être envisagée entre les régions frontales, les régions temporo-pariétales, les aires unisensorielles et des structures sous corticales (f ; Figure extraite de Senkowski et al., 2008).***

Le système moteur – En plus d'être impliqué dans la perception auditive de la parole (voir section A-4), le système moteur semble jouer un rôle important dans l'intégration de la parole audio-visuelle. Effectivement, des études suggèrent que ce phénomène de fusion serait en partie modulé par le système moteur de la parole (incluant notamment la partie postérieure du gyrus frontal inférieur et le cortex prémoteur ventral adjacent, principalement dans l'hémisphère gauche). Une augmentation des activités motrices a été observée lors de la perception audio-visuelle de la parole par rapport à la perception de stimuli unimodaux auditifs et visuels (Callan et al., 2003 ; Skipper et al., 2005, 2007). A l'inverse du pSTS, le système moteur semble plus activé lors de la présentation de stimuli incongruents par rapport à des signaux congruents (Jones et Callan, 2003). Finalement, Callan et son équipe (2003, 2004) ont montré qu'en dégradant le signal auditif ou visuel de parole lors d'une présentation bimodale, une activation plus importante des régions motrices pouvait être observée. Ces études suggèrent ainsi l'existence de mécanismes de simulation motrice lors de la perception audio-visuelle de la parole et soutiennent

l'hypothèse de processus d'appariement et d'intégration entre signaux auditifs et visuels et connaissances procédurales motrices de l'auditeur (Schwartz, Sato et Fadiga, 2008; Schwartz et al., 2012).

Ces résultats sont cohérents avec le modèle neurobiologique de Skipper et collègues (2007; voir partie A-3) qui propose que les informations extraites des signaux auditifs et visuels de parole seraient des hypothèses et non des interprétations définitives concernant le message perçu. Ces hypothèses seraient générées dans le pSTS, lieu de convergence des flux sensoriels, pour ensuite être envoyés vers le cortex prémoteur ventral (PMv) afin d'être appariées aux commandes motrices. De là, des prédictions sensorielles seraient renvoyées par copies d'efférence au pSTS afin de contraindre l'interprétation du message perçu. Ainsi, en plus des aires sensorielles auditives et visuelles activées lors de la présentation d'un stimulus audio-visuel de parole, le pSTS et les régions motrices (particulièrement le cortex prémoteur ventral) formeraient un réseau temporo-pariéto-frontal optimum pour traiter et améliorer l'interprétation du signal multisensoriel entrant.

PARTIE THÉORIQUE - C

LA NATURE *PRÉDICTIVE* DE LA PERCEPTION DE LA PAROLE

Les deux premières sections de ce chapitre nous ont permis de montrer que la parole n'était pas qu'auditive. Avec l'aide du signal visuel de parole et des connaissances procédurales motrices que nous avons des gestes articulatoires, le traitement auditif est amélioré, facilité voire perturbé. Nous allons voir ici les mécanismes qui permettent à notre cerveau d'utiliser et de combiner toutes ces informations afin de prédire le contenu des signaux entrants. Dans un premier temps, nous expliquerons brièvement certaines théories du codage prédictif, d'un point de vue général d'abord puis son application au domaine de la parole. Nous verrons ensuite les mécanismes neuronaux à l'origine de ces processus prédictifs lors de la perception de la parole audio-visuelle.

1. Théorie du codage prédictif

Il faut voir le cerveau non pas comme un système « passif » entrée-sortie dans lequel les entrées sensorielles sont traitées de manière hiérarchique : des aires sensorielles primaires, secondaires, associatives jusqu'aux aires de plus haut niveau, mais plutôt comme un système « proactif », capable d'émettre des prédictions et d'en vérifier la validité en les comparant avec les entrées sensorielles à chaque sous-niveau de traitement. Le sens de circulation des informations serait donc « bidirectionnel », avec dans un sens ascendant le traitement des entrées sensorielles, de la périphérie vers le cortex et dans un sens descendant, les prédictions générées par le cerveau sur la nature de ces entrées sensorielles, à chaque niveau de traitement (voir Figure 9).

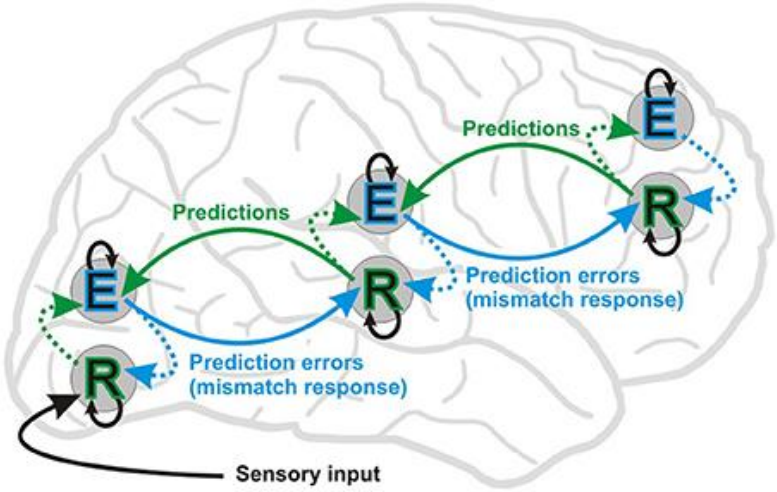

Figure 9 : Trajet d'une information sensorielle à différents niveaux de traitements et à travers deux populations neuronales types : les unités d'erreurs (E) et les unités de représentation (R). Dans ce modèle, les connections ascendantes (traits pleins bleus) transportent les erreurs de prédictions et les connections descendantes (traits pleins verts) transportent les prédictions (Figure extraite de Stefanics et al., 2014) .

La prédiction fournirait un cadre optimal pour minimiser conjointement le temps de traitement de ces informations et le coût du traitement neuronal associé en utilisant les représentations préalablement stockées en mémoire. Ainsi, à partir des entrées sensorielles, le cerveau infèrerait un modèle interne du monde extérieur qui permettrait par la suite de créer des prédictions sur ces entrées sensorielles. Il génèrerait en permanence ces anticipations/prédictions et lorsqu'il y aurait une violation du modèle, c'est-à-dire lorsque les traitements de ces entrées sensorielles seraient différents du modèle, il produirait alors des messages d'erreur ou de surprise. La théorie du codage prédictif (Friston, 2005) stipule que les activités se propageant de façon ascendante (de type « bottom-up ») et descendante (« top-down ») permettraient l'estimation d'erreur de prédiction, c'est-à-dire la différence entre ce que nous anticipons (les prédictions) et ce que nous percevons réellement (les entrées sensorielles). Une actualisation systématique du modèle interne du monde extérieur se ferait grâce aux erreurs de prédictions : tant que les prédictions seraient correctes, le modèle serait renforcé, en revanche, si en cas de prédictions inexactes, le modèle serait réactualisé pour ne pas reproduire cette erreur par la suite et ainsi anticiper au mieux les événements à venir.

Nous allons voir dans un premier temps comment cette hypothèse du codage prédictif peut s'appliquer aux mécanismes qui sous-tendent la parole, puis plus spécifiquement en quoi les indices visuels du signal de parole peuvent contraindre les prédictions faites sur le contenu du signal auditif.

La parole prédictive – Notre cerveau serait donc un système prédictif, il prendrait en compte ce qui a été formulé jusque-là et anticiperait/prédirait, en fonction de notre expérience passée et du contexte, ce qui va être perçu par la suite. Mais pour mieux comprendre l'aspect prédictif de la perception de la parole, il est important dans un premier temps d'expliquer les mécanismes de prédictions qui sous-tendent la production de la parole.

Modèles de production de la parole – Pour comprendre ce couplage perception-action, il est nécessaire de comprendre comment est générée une action – ici des gestes de parole, c'est-à-dire comment fonctionne notre contrôle moteur de la parole. Pour produire un son, trois étapes sont nécessaires : La planification du mouvement, l'exécution du mouvement et le contrôle/la vérification de ce qui a été produit.

Le modèle de Wolpert (1997) : Dans ce modèle du contrôle moteur de la parole, deux modèles internes (c'est-à-dire deux circuits neuronaux différents) seraient impliqués dans la production d'un geste articulatoire (voir Figure 10): le modèle « inverse » qui permettrait l'exécution de l'action préalablement planifiée et le modèle « direct » qui permettrait de prédire les conséquences sensorielles de l'action exécutée. Lors de la planification du geste, le modèle inverse génèrerait les commandes motrices nécessaires pour atteindre les buts articulatoires (et auditifs) désirés et les enverrait au système articulatoire et phonatoire pour que les gestes de parole soient exécutés. En parallèle, le modèle inverse enverrait également des copies de ces commandes motrices (« copies d'efférence ») au modèle direct qui est chargé de générer une prédiction des conséquences auditives et somatosensorielles (décharges corollaires) de ces commandes motrices. Les copies d'efférence permettraient d'anticiper l'action et d'informer le système nerveux central sur les conséquences des actions engagées avant même que l'information sensorielle (conséquence de l'action) soit disponible. Ainsi, le modèle direct aurait pour rôle d'associer les commandes motrices (via

les copies d'efférence) et leurs conséquences en émettant des prédictions basées sur les représentations internes déjà apprises par le modèle.

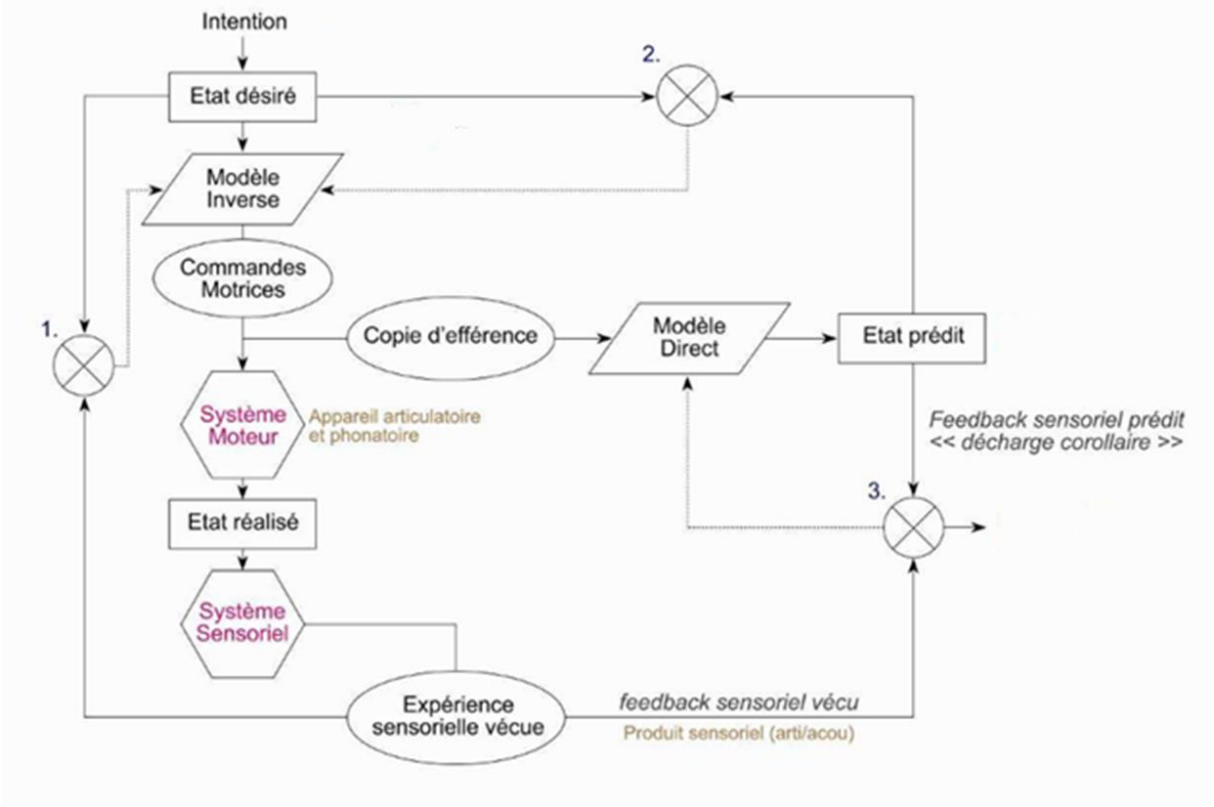

Figure 10 : Schéma du contrôle moteur de la parole de Wolpert (Figure issue de la thèse de Lucile Rapin (2012) adaptée de Wolpert, 1997 et Blakemore, 2003). Les sommateurs 1, 2 et 3 représentent les 3 comparaisons effectuées durant le processus de réalisation d'une action de parole.

Trois comparaisons seraient effectuées durant ce processus de production afin d'ajuster au mieux le geste à exécuter en fonction du but et des modèles/représentations internes :

- Une première comparaison se ferait entre la sortie effective (ce qui est réellement produit) et l'état désiré au départ (les buts) afin d'ajuster si nécessaire les paramètres de l'action en cours (voir le sommateur 1 de la Figure 8).
- Une seconde comparaison se ferait entre l'état désiré (les buts) et l'état prédit afin d'ajuster les commandes motrices avant la réalisation de l'action. Ce processus est plus rapide car il ne nécessite pas d'attendre l'exécution du geste (voir le sommateur 2 de la Figure 8).
- Et enfin, une troisième comparaison se ferait entre la sortie effective (ce qui est réellement produit) et la sortie prédite afin d'ajuster le modèle direct (et les représentations internes déjà existantes) pour les prédictions futures (voir le sommateur 3 de la Figure 8).

Le modèle DIVA (1994, 2011) : Un second modèle est également intéressant pour appréhender la nature prédictive de la perception de la parole. Guenther et collègues (1994, 2011) proposent en effet un modèle neurobiologique sensorimoteur de la production de la parole qui vise à rendre compte des interactions entre les aires motrices, somatosensorielles et auditives durant la production de parole (voir Figure 11).

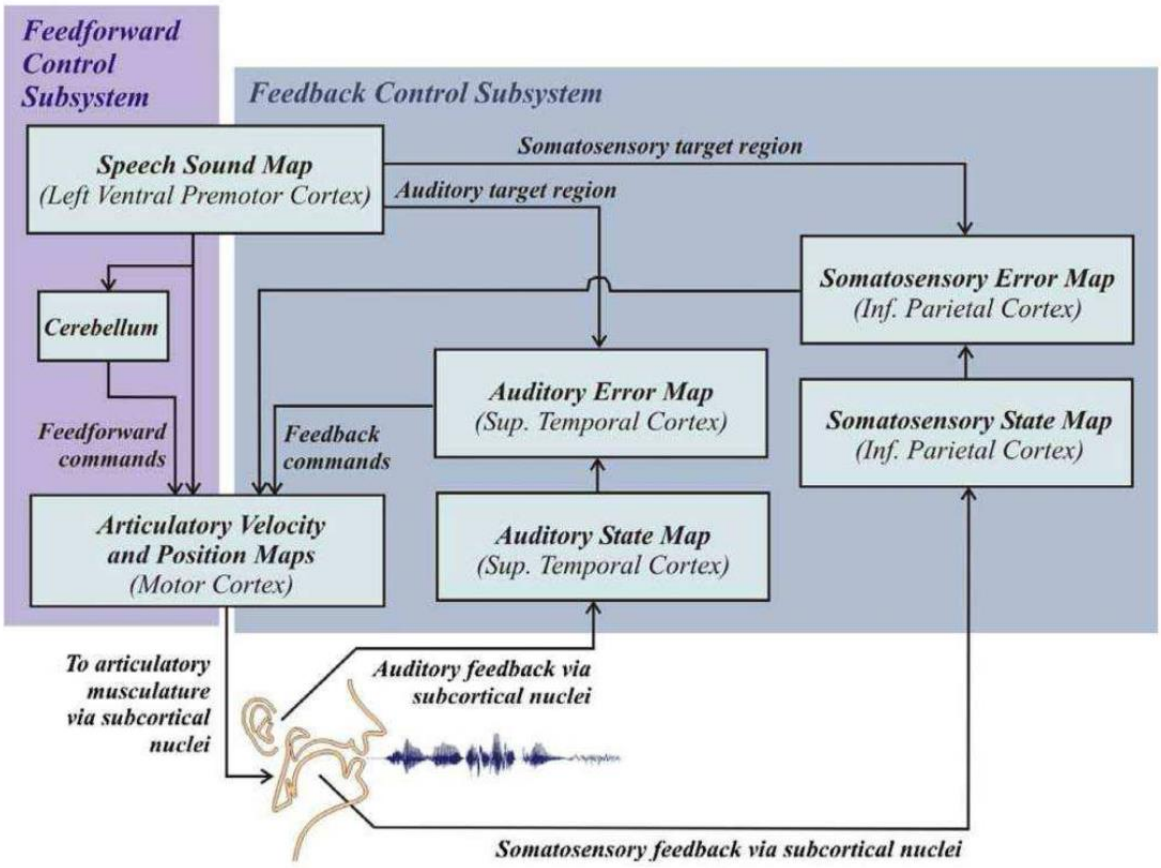

Figure 11 : Schéma du modèle DIVA du contrôle moteur de la parole (figure extraite de Guenther et Vladusich, 2012). Deux systèmes de contrôles différents : le système direct (feedforward, en violet) et le système rétroactif (feedback, en gris).

Dans ce modèle, appelé DIVA (pour "Directions Into Velocities of Articulators"), il existerait deux systèmes de contrôles bien distincts : un premier système « direct » (ou « proactif », « feedforward ») responsable de l'exécution des mouvements des articulateurs jouerait un rôle clé dans les processus d'acquisition du langage. Ce serait en effet ce système qui permettrait d'apprendre les relations entre les représentations phonémiques, motrices, auditives et somatosensorielles. Et un second système de « rétroaction » (« feedback ») qui vérifierait la bonne réalisation de l'objectif de départ en comparant la réponse attendue et la réponse effective. Ainsi, pour parler, un locuteur activerait une carte de sons de parole (« speech sound map ») localisée dans le cortex prémoteur ventral gauche), initiant ainsi deux systèmes de contrôles en parallèle : le système direct qui initierait les plans moteurs du geste à produire afin d'exécuter le programme moteur, et en parallèle le système de contrôle du feedback qui recevrait des informations relatives aux conséquences sensorielles des commandes motrices au sein de cartes d'erreurs auditives et somatosensorielles ("auditory and somatosensory error maps"). Ces cartes d'erreurs permettraient de comparer les prédictions sensorielles avec les signaux entrants pour transmettre au cortex moteur primaire, en cas d'erreurs, des informations nécessaires à l'ajustement de l'acte moteur via une carte de contrôle rétroactif ("feedback control map").

La prédiction dans les modèles de perception de la parole – Les modèles de perception de la parole s'appuient sur des principes similaires à ceux discutés dans le cadre des modèles de production, ce qui souligne encore une fois ce lien étroit entre les mécanismes de

production et de perception. Prenons par exemple deux des modèles que nous avons déjà évoqués plus tôt dans ce chapitre : le modèle d'analyse par synthèse de Skipper et collègues (2007 ; voir partie A-3) et le modèle à double voie antérieure/postérieure de Scott, Rauschecker et collègues (2003 ; voir partie A-3). Ces deux modèles, en plus de traitements purement sensoriels, font tous deux référence à une voie dorsale (sensorimotrice), impliquée dans les mécanismes d'appariement entre représentations sensorielles et motrices. D'après ces modèles, le circuit neuronal utilisé pour produire les gestes de parole serait en partie recruté pour simuler les commandes motrices à l'origine du son perçu. De là, des copies d'efférence (c'est-à-dire les conséquences/prédictions sensorielles de l'action produite, de manière similaire aux copies envoyées par le modèle inverse du modèle de Wolpert (1997) ou reçues par le système rétroactif du modèle de Guenther (1994, 2011)) vont être renvoyées, soit directement au cortex auditif (dans le cas du modèle de Skipper et collègues), soit à une interface sensorimotrice située au niveau de l'IPL (dans le cas du modèle de Scott, Rauschecker et collègues) afin d'être comparées aux hypothèses phonémiques émises au préalable (Skipper et al., op. cit.) ou directement aux entrées sensorielles effectives (Scott et al., op. cit.) dans le but de contraindre l'interprétation finale du signal auditif.

Le rôle des prédictions motrices serait d'autant plus important lorsque le signal auditif est dégradé. Elles permettraient en effet d'avoir accès aux différentes représentations motrices qui pourraient être à l'origine du signal sonore produit, contraignant ainsi l'espace des interprétations du message linguistique perçu. Mais il est plus aisé de percevoir la parole à travers toutes les modalités sensorielles disponibles. En effet, l'intégration multisensorielle de la parole améliorerait la perception du signal auditif de parole en réduisant, dans le cas d'un signal bruité par exemple, l'ambiguïté des stimuli. Nous allons montrer dans la section suivante comment le cerveau utilise les informations visuelles pour prédire le signal auditif de parole.

Intégration multisensorielle – Les informations visuelles contenues dans le signal de parole nous permettent de désambiguïser le message linguistique transmis dans un environnement parfois bruité. Non seulement elles nous permettent de combler les manques du signal acoustique dégradé, mais elles améliorent et facilitent l'interprétation finale du message. Nous pouvons attribuer au signal visuel la propriété de "prédire" la parole auditive, et ce, en partie du fait de sa précédence naturelle sur les informations auditives. En effet, dans la littérature il a été longtemps admis que l'information des lèvres précédait l'information auditive d'environ 150ms (van Wassenhove et al., 2007 ; Chandrasekaran et al., 2009).

Cette précédence est visible au niveau syllabique et vocalique, mais elle est toutefois débattue dans le cas d'un flux de parole continu. SSchwartz et Savariaux (2014) ont en effet montré qu'il peut effectivement exister une forte anticipation visuelle, mais qu'elle n'est essentiellement présente que dans le cas d'une préparation du geste articulatoire en début de séquence de sons, durant laquelle aucun son n'est encore produit. Dans le cas d'un flux de parole continu, les auteurs parlent plutôt de gestes co-modulaires. Ces gestes seraient modulés en même temps que le signal sonore avec des phénomènes d'asynchronie entre les informations visuelles et auditives plus complexes que prétendu jusqu'alors. Dans leur expérience, ils ont enregistré des syllabes audio-visuelles CV isolées (composées d'une des consonnes (C) suivantes : /p t k b d g m n/, suivies de la voyelle (V) /a/) ainsi que des enchainements ininterrompus de plusieurs syllabes audio-visuelles (du type VCVCVCV) prononcées par un locuteur français. Ils ont ensuite fait une analyse acoustique et visuelle

sur la base de données enregistrées afin d'obtenir le début du signal acoustique ainsi que la trajectoire des lèvres pour chaque syllabe isolée et enchainée. Les auteurs ont ainsi confirmé la précédence du visuel sur l'auditif lors de la perception de syllabes isolées, avec une avance allant de 150ms à 400ms, ce qui est cohérent avec les résultats de van Wassenhove et collègues (2007) et Chandrasekaran et collègues (2009). Cependant lors de la perception d'une séquence de plusieurs syllabes (impliquant des gestes co-modulaires), la précédence du visuel est réduite (70ms) et l'auditif est parfois même plus précoce que le visuel (20ms ; dans le cas de consonnes voisées notamment).

Notons également un phénomène intrinsèque au flux de parole continu : la coarticulation. Cette notion est à distinguer des gestes co-modulaires. En effet, la coarticulation est relative à la succession des gestes articulatoires et non à la relation temporelle entre les informations visuelles et auditives. Elle peut s'expliquer par la propriété physique d'un objet en mouvement, qui ne peut passer d'un état à un autre instantanément, mais qui doit effectuer cette modification progressivement. Ainsi en parole, la forme du conduit vocal pour un son donné ne peut se transformer instantanément en une autre forme pour produire un autre son. Le changement doit se faire progressivement. La configuration nécessaire pour produire le phonème suivant va donc influencer la forme du conduit vocal du phonème qui est en train d'être produit, par un effet d'anticipation du geste articulatoire à venir, et le phonème futur sera lui aussi teinté par la configuration qu'avait le conduit vocal précédemment. Or, ces phénomènes d'anticipation sont parfois visibles avant d'être audibles, et parfois au contraire audibles avant d'être visibles (Troille et al., 2010).

C'est probablement pour cette raison qu'un certain nombre d'études ont observé que le phénomène d'intégration tolérait jusqu'à 250ms de désynchronisation auditive et visuelle sans perturber l'intégration optimale. Dans leur expérience, van Wassenhove et collègues (2007) ont présenté deux jeux de stimuli désynchronisés aux participants (des syllabes incongruentes de type McGurk et des syllabes congruentes) avec deux consignes différentes : une tâche d'identification de la syllabe perçue et une tâche de jugement de la synchronisation des stimuli présentés. Leurs résultats montrent que quelle que soit la tâche demandée, l'identification audio-visuelle des stimuli ainsi que le jugement temporel toléraient jusqu'à 250 ms de désynchronisation audio-visuelle pour des syllabes congruentes ou non. Cependant, cette asynchronie n'est pas tolérée de la même façon en cas de précédence du visuel sur l'auditif ou, a contrario, en cas de précédence de l'auditif sur le visuel. Ils ont en effet montré qu'une précédence du visuel était mieux tolérée qu'une précédence du signal auditif, ce qui paraît cohérent avec le décours temporel naturel de la parole audio-visuelle.

L'asynchronie audio-visuelle n'est donc pas nécessaire pour que le cerveau émette des prédictions. Nous avons vu qu'une entrée auditive seule produisait déjà des prédictions motrices sur le contenu du signal auditif. Nous avons également vu que le signal visuel pouvait améliorer ces prédictions, bien qu'il ne soit pas systématiquement en avance sur les événements auditifs présents. Les résultats mentionnés ont également mis en évidence la large tolérance à l'asynchronie des mécanismes d'intégration audio-visuelle, c'est-à-dire qu'ils peuvent se réaliser de façon optimale malgré un décalage relativement important entre le visuel et l'auditif. Nous allons voir dans une seconde partie en quoi les données électrophysiologiques peuvent permettre d'étudier les prédictions émises à partir des informations visuelles sur le signal auditif entrant.

2. Mécanismes neuronaux du codage prédictif

Dans cette section, nous allons nous intéresser aux mécanismes neuronaux du codage prédictif de la perception de la parole. Nous l'avons vu, le cerveau n'est pas un système « passif » et unidirectionnel, il existe des mécanismes de prédictions qui, à travers des flux « bottom-up » (visuel -> auditif) et « top-down » (auditif -> visuel) entre les aires sensorielles, permettent de contraindre l'interprétation finale du signal auditif entrant. Nous avons également mentionné que nos connaissances procédurales motrices pouvaient elles aussi améliorer les traitements auditifs et visuels. Nous pouvons donc supposer qu'il existerait également un échange d'information entre les régions sensorielles auditives et visuelles et le système moteur, à travers de possibles flux bi-directionnels entre ces différents centres de traitement.

L'électro-encéphalographie (ou EEG) est une méthode d'exploration cérébrale non-invasive qui permet d'enregistrer l'activité électrique du cerveau afin d'étudier le décours temporel des processus de traitements des informations. Elle est donc l'outil idéal pour observer les mécanismes neuronaux du codage prédictif, notamment l'influence du signal visuel sur les traitements auditifs de la parole. Nous allons décrire dans cette partie les différents marqueurs typiques de l'intégration de syllabes audio-visuelles : une modulation de l'amplitude et de la latence représentée par les pics d'activation des potentiels évoqués auditifs (PEAs) N1 et P2. Nous verrons dans le même temps les études EEG *princeps* qui ont participé à la compréhension des mécanismes électrophysiologiques de l'intégration audio-visuelle de la parole.

Amplitude et latence – Bien que nous allons utiliser ici le terme « d'intégration précoce », il est important de noter que les potentiels évoqués auditifs (PEAs) que nous allons mentionner par la suite sont des potentiels évoqués en réalité « tardifs » au sein de la chaîne de traitements de décodage acoustique, c'est-à-dire qui apparaissent au minimum 50 ms après le début du stimulus acoustique (a contrario des potentiels « précoces » qui reflètent des mécanismes sous-corticaux et qui apparaissent moins de 10 ms après le début du signal acoustique). Ces potentiels évoqués sont des réponses neuronales à une stimulation sensorielle extérieure. Il existe différents pics d'activations, comme la P50 (pic positif apparaissant 50 ms après le début de signal acoustique), les N1/P2 (pics respectivement négatif et positif apparaissant 100 et 200 ms après le début du signal acoustique), la P300 et la N400 (pics respectivement positif et négatif apparaissant 300 et 400 ms après le début du signal acoustique). Dans la suite du manuscrit nous nous focaliserons sur les caractéristiques des PEAs N1 et P2, marqueurs typiques de l'intégration audio-visuelle précoce (voir aussi des résultats similaires obtenus par le paradigme de la Mismatch Negativity, Colin et al., 2002). Une modulation de leur amplitude ou de leur latence est un indice pour étudier l'influence des informations visuelles sur l'activité cérébrale induite par une entrée auditive. Dans cette sous-section, nous étudierons les PEAs N1 et P2 pris ensemble (complexe N1/P2). Nous les dissocierons dans la section suivante.

Amplitude - Des études en EEG ont montré que la présentation d'un stimulus audio-visuel de parole était généralement accompagnée d'une diminution de l'amplitude des PEAs, par rapport à la présentation uniquement auditive du même stimulus (Klucharev, Möttönen et Sams, 2003 ; Besle et al., 2004 ; van Wassenhove, Grant et Poeppel, 2005; Stekelenburg et Vroomen, 2007 ; Pilling, 2009 ; Vroomen et Stekelenburg, 2009). Cependant, cette analyse n'est pas suffisante pour déterminer si ces PEAs liés à une présentation audio-visuelle sont

uniquement le résultat de traitements auditifs et visuels séparés et indépendants ou si, au contraire, ils sont le reflet d'une interaction entre les traitements des informations auditives et visuelles. Afin d'éclaircir cette question, Klucharev, Möttönen et Sams (2003) ont comparé les réponses évoquées auditives enregistrées lors d'une présentation de syllabes visuelles, auditives ou audio-visuelles. En analysant les signaux EEG relatifs à la présentation audio-visuelle (AV) et à la somme de signaux EEG relatifs à une présentation auditive et visuelle (A+V), ils ont montré une réduction de l'amplitude en perception AV par rapport à la somme des conditions A+V. Leurs résultats montrent qu'il existe bien des interactions audio-visuelles qui sont mises en évidence par une diminution de l'amplitude des réponses évoquées auditives (voir également Besle et al., 2004 et van Wassenhove et al., 2005 pour des résultats similaires).

Une explication possible de cette diminution d'amplitude pourrait venir des prédictions visuelles envoyées dans le cortex auditif, créant ainsi une « désactivation » (selon les termes de van Wassenhove et al.) ou une « dépression » (selon les termes de Besle et al.). En accord avec cette hypothèse, il est à noter que selon la théorie du codage prédictif, les prédictions sensorielles permettraient de réduire le nombre de candidats acoustiques possibles et de là une diminution des réponses neuronales auditives.

L'attention pourrait également avoir un rôle dans cette réduction de l'amplitude. En effet, nous pouvons supposer une différence dans l'attention portée à l'information auditive lorsque le stimulus est uniquement auditif (l'attention est entière), par rapport à un stimulus audio-visuel pour lequel l'attention serait partagée entre les informations auditives et visuelles générant ainsi une réponse plus faible dans le cortex auditif. Afin d'éclaircir cette nouvelle hypothèse, Pilling (2009) a entrepris de tester l'effet de l'attention sur la perception audio-visuelle de stimuli synchrones ou non par rapport à une perception auditive seule. En effet, l'intégration audio-visuelle ne pourrait se faire que durant une période précise (d'environ 250ms comme vu précédemment), donc si le décalage entre le signal auditif et le signal visuel dépasse cette fenêtre temporelle, et si la réduction de l'amplitude est un marqueur de cette intégration, alors cette réduction d'amplitude observée en condition audio-visuelle naturelle devrait être largement réduite en condition audio-visuelle désynchronisée. En revanche, si l'attention contribue à la diminution de l'amplitude, alors l'effet ne devrait être que partiellement modifié, du fait d'une attention visuelle équivalente entre les stimuli audio-visuels synchrones ou non. Leurs résultats concernant les stimuli synchrones montrent une diminution classique de l'amplitude des PEAs en condition AV par rapport à A. En revanche la désynchronisation des informations auditives et visuelles a supprimé cette différence d'amplitude, montrant ainsi que le phénomène d'intégration est bien à l'origine de cette réduction de l'amplitude et qu'elle n'est pas fortement teintée par d'autres mécanismes cognitifs comme l'attention.

L'amplitude est donc un marqueur spécifique de l'intégration audio-visuelle. Cependant, d'après van Wassenhove et collègue (2005), l'amplitude serait indépendante de la saillance perceptive des syllabes, donc du contenu du message. En effet dans leur étude, les auteurs ont analysé la différence des PEAs auditifs et audio-visuels (A-AV) pour l'amplitude en fonction des scores de reconnaissance visuelle des trois types de syllabes présentées /pa/, /ta/ et /ka/. Or, ces syllabes n'ont pas la même saillance visuelle (/p/ est une consonne bilabiale qui est très bien reconnue : environ 95% de réponses correctes, tandis que /k/ est une consonne vélaire, peu visible, qui n'est correctement identifiée que dans 65% des cas). Leurs résultats montrent que la différence d'amplitude est constante, malgré une différence

de saillance visuelle importante entre les syllabes présentées. L'amplitude reflèterait donc un re-phasage de l'activité du cortex auditif, mais ne transmettrait pas les informations relatives au contenu du signal de parole.

La latence – Nous venons de montrer que l'amplitude des potentiels évoqués auditifs pouvait être un marqueur de l'intégration audio-visuelle de la parole, mais une autre mesure semble elle aussi donner des indications sur ce processus particulier : la latence, c'est-à-dire le moment auquel apparaissent les PEAs après la présentation d'un stimulus auditif, ou audio-visuel. De nombreuses études en EEG sur l'intégration de stimuli de parole ont en effet montré, en plus d'une modulation de l'amplitude, un changement de latence des PEAs lors d'une présentation audio-visuelle par rapport à une présentation auditive seule.

L'ajout des informations liées au lieu d'articulation de la consonne accélèrerait le traitement du signal auditif. Si nous remettons ce résultat dans le cadre général de la théorie du codage prédictif, ce sont les prédictions liées au signal visuel qui, une fois envoyées vers les régions auditives (et possiblement motrices) permettraient d'accélérer le traitement du signal auditif entrant, en réduisant le nombre d'interprétations possibles et en pré-activant les neurones pour qu'ils soient plus rapidement réceptifs au signal entrant.

Il est important de rappeler ici qu'une majorité des études en EEG réalisées sur les mécanismes d'intégration audio-visuelle de parole utilisent des syllabes CV ou des voyelles (pour une revue, voir Baart, 2016), or, comme vu précédemment, la précédence de l'entrée visuelle est relativement importante dans ces cas particuliers (Schwartz et Savariaux, 2014). C'est sans doute en partie pour cela qu'une modulation de la latence est attendue et observée dans la plupart des expériences (Klucharev et al., 2003 ; Besle et al., 2004 ; van Wassenhove et al., 2005 ; Pilling, 2009 ; Winneke et Phillips, 2011 ; Paris et al., 2016 ; Baart et Samuel, 2015 par exemple).

D'après van Wassenhove et collègues (2005), au contraire de l'amplitude, une modulation de la latence reflèterait une différence de traitement de traits relatifs au contenu du message. En effet, dans leur expérience, une de leurs analyses consistait à comparer l'amplitude et la latence en fonction de la saillance visuelle des syllabes présentées (c'est-à-dire des scores de reconnaissance visuelle). Si l'amplitude n'a montré aucune variation, il en est tout autre pour la latence. Ils ont pu observer que plus grande était la saillance perceptive de la syllabe (ici, /pa/ > /ta/ > /ka/) plus importante était la différence de latence A-AV des signaux EEG correspondants. Autrement dit, plus le lieu d'articulation est visible, plus rapide est le traitement auditif du signal pour la condition AV par rapport à A.

Cependant, certaines études ne retrouvent pas cette facilitation temporelle (Kaganovich et Schumaker, 2014 ; Stekelenburg et Vroomen, 2007 par exemple). Cela peut s'expliquer par le fait que la saillance perceptive ne dépend pas uniquement des traits phonétiques du phonème présenté, elle peut également varier en fonction de la qualité des enregistrements vidéo, de la taille du stimulus présenté (le visage en entier vs. les lèvres seulement), ainsi que des idiosyncrasies du locuteur, c'est-à-dire les particularités de productions propres au locuteur. Et ces paramètres sont très variables entre les différentes études.

Nous venons d'expliciter deux marqueurs clés de l'intégration audio-visuelle de la parole : une modulation de la latence et/ou de l'amplitude des PEAs en condition audio-visuelle par rapport à une condition auditive seule (ou à la somme A+V). Nous avons vu que l'amplitude reflétait plutôt un stade de traitement de l'unification perceptuelle en général tandis que la

latence véhiculerait plutôt des informations sur le contenu du message en fonction de la saillance perceptive du signal visuel. Nous allons maintenant nous intéresser plus spécifiquement aux deux potentiels évoqués auditifs typiques de l'intégration précoce de syllabes audio-visuelles : N1 et P2.

Amplitude et latence de N1 vs P2 – Les potentiels évoqués auditifs N1 et P2 sont des pics d'activité négatif (N) ou positif (P) qui apparaissent respectivement 100 ms et 200 ms après le début du stimulus acoustique. Ces PEAs sont enregistrés au niveau des électrodes fronto-centrales, là où leur amplitude est maximale. Nous allons tenter de décrire les propriétés fonctionnelles relatives à la modulation de l'amplitude et de la latence de ces deux potentiels évoqués auditifs.

Amplitude de N1 – Il semblerait que la réduction de l'amplitude du PEA N1 soit modulée par les propriétés temporelles et spatiales du stimulus audio-visuel présenté (Stekelenburg et Vroomen, 2007, 2012). Dans l'expérience la plus récente, ils ont présenté des stimuli audio-visuels de parole (/bi/ et /fu/) ainsi que des stimuli audio-visuels d'actions biologiques (« taper des mains » ou « le choc d'une cuillère contre une tasse »). Ces stimuli étaient présentés avec une congruence spatiale ou non (le son et la vidéo provenaient soit du même endroit : en face du participant, soit d'un endroit différent : le signal audio provenait d'un haut-parleur situé à côté du participant tandis que la vidéo apparaissait au centre de l'écran et non à la périphérie). Après avoir analysé l'amplitude du PEA N1, ils ont observé une réduction plus importante de l'amplitude lorsque les stimuli étaient congruents spatialement. Dans une étude plus ancienne, ils ont également montré que l'amplitude du PEA N1 était affectée par la congruence temporelle des stimuli audio-visuels. En effet, ils ont observé que si les informations visuelles ne précédaient pas le signal auditif, aucune diminution d'amplitude n'apparaissait pour le PEA N1. Ces résultats suggèrent que l'amplitude du PEA N1 reflèterait des processus de traitement spatial et temporel du signal auditif, sans pour autant transmettre des informations sur le contenu.

Amplitude de P2 – Il semblerait que l'amplitude du PAE P2 serait plutôt modulée par la congruence phonétique des stimuli audio-visuels (Stekelenburg et Vroomen, 2007). Dans cette étude, les auteurs ont effectivement montré une augmentation de l'amplitude du PEA P2 lors de la présentation de stimuli incongruents (/bi/ en audio et /fu/ en visuel) par rapport à une présentation de syllabes congruentes (/bi/). Ces résultats sont cohérents avec le processus de liage phonétique qui apparaitrait au moment du pic P2 (Baart et al., 2014). Ces derniers ont montré à des participants des pseudo-mots audio-visuels /tabi/ et /tagi/, qui ont été modifiés en parole sinusoïdale (Remez et al., 1981), reconnue comme de la parole par la moitié des sujets ou comme du bruit par l'autre moitié des sujets. En analysant les potentiels évoqués auditifs N1 et P2, les auteurs ont pu observer une réduction de l'amplitude du PEA P2 uniquement lorsque les participants identifiaient l'onde sinusoïdale comme étant un signal de parole.

D'autre part, afin de distinguer le liage phonétique audio-visuel du processus de détection de congruence, Baart et al. (2014) ont décalé l'apparition de la congruence/incongruence en présentant par exemple /tabi/ en audio et /tagi/ en visuel (ou inversement). La détection de l'incongruence n'apparait que 270ms après le début du signal acoustique puisque ce n'est que le /b/ ou le /g/ qui est modifié, elle apparait donc normalement après le PEA P2. Les résultats ainsi observés sur le PEA P2 sont spécifiquement liés au liage phonétique et non à la détection audio-visuelle de la congruence.

Pris ensemble, ces études nous montrent une certaine spécialisation fonctionnelle des PEA N1 et P2. Le potentiel évoqué auditif N1 serait relatif aux informations spatiales et temporelles, mais ne serait pas spécifique à la parole, tandis que le PEA P2 serait quant à lui plus sensible au contenu du signal (van Wassenhove et al., 2005).

En ce qui concerne la distinction fonctionnelle de la modulation de la latence des PEA N1 et P2, la littérature est moins claire. En effet, Baart (2016) a montré dans une méta-analyse que l'effet de la latence était à peu près réparti équitablement sur les PEA N1 et P2 dans les 20 études mises en commun. Cela suggère qu'elle ne serait pas spécifique à un pic d'activation mais plutôt liée au complexe N1/P2 dans sa globalité. Van Wassenhove et collègues (2005) montrent par exemple une réduction de la latence en fonction de la saillance visuelle sur les deux PEAs, tandis que Stekelenburg et Vroomen en 2007 ne montrent un effet de la latence que sur le pic N1 (voir aussi Baart et al., 2013 pour des résultats similaires). Dans une expérience ultérieure (2012), ils observent un effet de la latence uniquement sur P2 tout comme Kaganovich et Schumaker (2014). Cependant, lors d'une étude en MEG (magnétoencéphalographie), Arnal et son équipe (2009) ont montré une facilitation temporelle de la M100 (équivalent du PEA N1) qui serait proportionnelle au degré de prédictibilité du stimulus (réduction de la latence plus importante lorsque la syllabe est très saillante visuellement comme /ja/, et réduction moindre lorsque la prédictibilité de la syllabe est faible comme pour /ga/). De plus, cette facilitation temporelle de M100 ne semble pas être influencée par l'incongruence audiovisuelle des stimuli. Arnal et ses collègues interprètent ces résultats de la façon suivante : il existerait deux voies de traitement par lesquelles les informations visuelles peuvent faciliter le traitement auditif. La première voie serait une voie cortico-corticale directe du cortex visuel au cortex auditif. Compte-tenu de la précédence du visuel lors de la perception de syllabes isolées, le signal visuel enverrait des prédictions via cette voie directement au cortex auditif. Dans un même temps, les prédictions visuelles seraient également envoyées au STS (sillon temporal supérieur ; interface en partie responsable de l'intégration des différentes modalités). De son côté, le signal auditif entrant enverrait lui aussi les informations auditives au STS. De là, une seconde voie serait activée, qui acheminerait par « feedback » les comparaisons faites entre les prédictions visuelles et l'entrée auditive, du STS jusqu'au cortex auditif, pour contraindre l'interprétation finale. La facilitation temporelle observée sur la M100 serait le reflet de la première voie directe et serait dépendante de la saillance perceptive. Cependant, elle ne serait pas influencée par l'incongruence. Ce premier mécanisme prédictif serait donc uniquement lié aux caractéristiques visuelles et n'aurait pas encore reçu les informations portant sur la correspondance entre l'entrée visuelle et l'entrée auditive, il ne serait donc pas issu du STS. C'est lors du second mécanisme prédictif rétroactif (feedback) qu'une erreur de prédiction serait renvoyée au cortex auditif depuis le STS pour finaliser l'interprétation du signal acoustique.

Il existe d'autres niveaux d'interprétation de ces mécanismes prédictifs, notamment à travers l'étude des rythmes cérébraux. Ce sont eux qui donnent la « cadence » pour la bonne réalisation des processus de traitement des informations entrantes. Ce sont des ondes (appelées aussi oscillations) de très faibles amplitudes (quelques microvolts), représentant l'activité électrique cohérente d'un grand nombre de neurones et qui peuvent être classées en fonction de leur fréquence. Les ondes *Delta*, comprises entre 2 et 4 Hz, les ondes *Thêta*, comprises entre 4 et 8 Hz, les ondes *Alpha*, comprises entre 9 et 13 Hz, les ondes *Bêta*, comprises entre 15 et 20 Hz et les ondes *Gamma*, supérieures à 30 Hz, sont les principales

ondes qui permettent de caractériser l'activité électrique du cerveau. La présence ou l'absence de ces ondes semble déterminer le type de traitement qui est en cours. Or d'après la théorie du codage prédictif (Friston, 2005), les prédictions envoyées par les voies descendantes auraient pour rôle de pré-activer des représentations internes de façon à minimiser le coût neuronal et optimiser le traitement de l'information sensorielle à venir. Il s'agirait donc de re-phaser l'activité oscillatoire de la région cérébrale impliquée dans le processus de traitement pour qu'elle soit dans une condition optimale lui permettant de recevoir l'activité électrique liée à l'information sensorielle entrante, et ainsi assurer la transmission du signal vers les autres régions.

Ainsi, d'après Arnal et collègues, ces prédictions visuelles acheminées à travers la voie directe auraient pour rôle de re-phaser les oscillations du cortex auditif de façon à préparer l'arrivée du signal auditif entrant. La seconde voie (feedback) permettrait quant à elle d'améliorer ce re-phasage en fonction de la congruence audiovisuelle du signal. Si le STS reçoit des informations convergentes des deux modalités, il renvoie au cortex auditif une erreur de prédiction minimisée, réduisant ainsi le coût neuronal du traitement auditif. En revanche, si les informations auditives et visuelles ne sont pas congruentes, l'erreur de prédiction renvoyée est plus importante, augmentant ainsi l'activité du cortex auditif.

Dans cette section, nous avons présenté des études mettant en évidence la nature prédictive de la parole. En nous appuyant sur la théorie du codage prédictif, nous avons discuté des mécanismes d'intégration audio-visuelle de la parole, suggérant l'existence de prédictions sensorielles et motrices permettant de contraindre et ainsi faciliter et accélérer le traitement du signal auditif. Ces mécanismes d'intégration précoces sont observables au niveau des potentiels évoqués auditifs N1 et P2 qui sont caractérisés par une modulation de leur amplitude et de leur latence en fonction de la modalité de présentation. Bien qu'une spécification fonctionnelle de ces marqueurs soit encore difficile à établir, une distinction qui ressort particulièrement est l'opposition entre les traitements liés au contenu du message et ceux associés aux informations temporelles et/ou non spécifiques à la parole.

PARTIE THÉORIQUE - D

LA NATURE *MULTIMODALE*, *SENSORIMOTRICE* ET *PRÉDICTIVE* DE LA PERCEPTION DE LA PAROLE

Dans une première partie, nous avons abordé la nature sensorimotrice de la parole. A travers l'existence d'un système miroir, nous avons discuté d'un possible couplage perception-action lors de la perception de la parole. Puis nous avons détaillé certaines théories, en faveur ou non de ce lien perceptivo-moteur, depuis les théories motrices et auditives jusqu'aux théories perceptivo-motrices. Nous avons ensuite illustré, à l'aide de différents modèles neurobiologiques, les interactions entre les entrées sensorielles et les connaissances procédurales motrices au niveau neuroanatomique. Finalement, nous avons montré un certain nombre de résultats empiriques en faveur d'un rôle fonctionnel et causal de notre système moteur dans les mécanismes de perception de la parole.

Dans une seconde partie, nous nous sommes confrontés à la nature multisensorielle de la parole et aux indices visuels exploitables lors de la lecture labiale. Nous avons montré que ces indices étaient complémentaires au signal sonore, et qu'ils facilitaient, jusqu'à un certain point, la perception du signal auditif de parole, que ce signal soit bruité, masqué, ambigu ou complexe. Nous avons mis en avant le caractère spontané et irrépressible de la lecture labiale au point de créer parfois une illusion perceptive lorsque les signaux auditifs et visuels étaient incongruents. Finalement, nous avons focalisé notre attention sur deux régions clés de l'intégration audio-visuelle de la parole : le pSTS et le système moteur de la parole.

Dans une troisième partie, nous avons mis en avant la nature prédictive de la parole, caractérisée par des échanges entre les modèles internes que nous avons du monde extérieur, les prédictions qu'ils génèrent et les entrées sensorielles que reçoit notre cerveau. Au niveau électrophysiologique, nous avons pu voir que les prédictions liées au signal visuel de parole permettaient de faciliter les mécanismes d'intégration audio-visuelle en modulant l'amplitude et/ou la latence des potentiels évoqués auditifs N1 et/ou P2, considérés comme les marqueurs-types du liage audio-visuel.

Dans cette dernière partie, nous présenterons une revue d'études portant sur la perception tactile de la parole, sur la reconnaissance visuelle des gestes linguaux, sur la perception de nos propres productions et sur les mécanismes d'intégration multimodale chez les personnes âgées. Ces études prises ensembles permettent d'apporter un éclairage différent quant à la question de la nature multimodale, sensorimotrice et prédictive de la perception de la parole. Elles constituent le socle, le point de départ, des questionnements et de la réalisation des études menées lors de ce travail de thèse. Ainsi, percevoir les mouvements de nos articulateurs par le toucher, voir et décoder les mouvements linguaux d'un locuteur ou encore percevoir visuellement nos propres productions de parole permettent de questionner l'apport de nos connaissances motrices quant aux gestes articulatoires de parole et le possible transfert de mécanismes prédictifs et d'intégration audio-visuelle à d'autres sources sensorielles. Enfin, nos connaissances motrices vues comme une aide à la

perception de la parole peuvent être envisagées comme un support pour pallier un déclin sensoriel dû au vieillissement afin de préserver les mécanismes d'intégration. Les quatre sections que nous allons présenter ci-après seront représentatives des expériences menées durant cette thèse, les articles relatifs à ces travaux seront présentés dans la partie expérimentale de la présente thèse.

1. Perception et intégration audio-tactile de la parole (toucher ce que l'on ne peut voir ni entendre)

Les interactions audio-tactiles sont courantes, chaque objet a une texture, une forme, et sa manipulation engendre la plupart du temps un son propre à cet objet et aux matériaux qui le composent. Ainsi, les informations tactiles et auditives issues d'actions comme toquer à une porte, applaudir, se gratter, ou jouer d'un instrument de musique vont être perçues simultanément. Mais dans le cas spécifique de la parole, il est rare d'utiliser ses propriétés tactiles pour la percevoir. Rappelons que la production de la parole consiste en une mise en mouvements de différents articulateurs et résonateurs afin de modifier la forme du conduit vocal pour produire des sons différents. Si certains de ces mouvements sont visibles, nous pouvons supposer qu'une partie de la déformation du conduit vocal pourrait être perçue par le toucher. C'est exactement ce que propose la méthode Tadoma (Alcorn, 1932). Mais il existe aussi d'autres manières de communiquer par le toucher : en adaptant une langue des signes visuelle en une langue des signes tactile par exemple ou bien en utilisant un des alphabets tactiles disponibles. Nous allons décrire brièvement ces différentes méthodes avant de nous attarder plus particulièrement sur la méthode Tadoma. Nous montrerons finalement des études comportementales qui se sont intéressées aux interactions entre les modalités auditive et tactile lors de la perception de la parole.

Communiquer par le toucher – Lorsqu'une ou plusieurs privations sensorielles surviennent, la communication devient fastidieuse et toute information devient précieuse pour comprendre le signal de parole. Chaque personne sourde et/ou malentendante, en fonction de son vécu et de son expérience, de son entourage et de l'aide disponible doit trouver la méthode qui lui convient le mieux, c'est pour cela qu'il n'existe pas qu'un seul moyen pour communiquer et que, bien souvent, plusieurs méthodes sont utilisées de façon complémentaire.

Par exemple, la langue des signes tactile s'inspire de langues des signes visuelles. Elle est adaptée aux personnes sourdes et aveugles et leur permet d'accéder à la forme de la main, son orientation, son placement et ses mouvements en posant leur propre main par-dessus celle du signeur (pour plus d'informations, voir Mesch, 2000). Cette méthode est surtout utilisée par des personnes sourdes devenues aveugles tardivement. Elles ont alors appris une langue des signes visuelle qu'elles ont ensuite adapté au toucher lorsque la vision ne leur apportait plus suffisamment d'éléments pour comprendre.

Lorsqu'on a besoin d'épeler un mot inconnu, un nom de famille, une marque, un des alphabets tactiles peut prendre le relais. On peut le retrouver sous plusieurs formes comme par exemple le tracé des lettres de l'alphabet sur une surface du corps (la main, le bras ou le dos) à l'aide de l'index. Mais il peut aussi prendre une forme plus codifiée, en effectuant différents mouvements et formes sur la paume de la main de la personne sourde-aveugle (cette méthode s'apparente alors plutôt à la dactylologie).

D'autres méthodes encore plus codifiées existent mais exigent un long apprentissage et souvent beaucoup de concentration. C'est le cas de l'alphabet codifié de Lorm, développé en 1881 par Hieronymus Lorm, qui consiste en un ensemble de pressions ou de traits effectués sur un endroit précis de la paume de la main. Cet alphabet peut être utilisé par l'intermédiaire d'un gant sur lequel des zones sont délimitées et attribuées à chacune des lettres de l'alphabet. Ainsi, n'importe quelle personne non initiée peut communiquer en suivant les traits tracés sur le gant porté par la personne sourde-aveugle.

Cependant, aucune des méthodes présentées ici n'utilise les propriétés articulatoires du langage oral et toutes sont assez fastidieuses et lentes à apprendre et à utiliser. Seule la langue des signes tactile récupère réellement le mouvement à la base du signe produit, et ce signe est issu non pas d'une langue orale mais d'une langue signée.

La méthode Tadoma – La méthode Tadoma est une des techniques qui va au plus près de la communication orale par le toucher. Développée par Sophie Alcorn (1932) à la Perkins School for the Blind du Massachussets, elle avait pour but initial d'apprendre aux élèves sourds et aveugles à parler. Le nom de cette méthode provient de celui des deux premiers élèves: Winthrop "Tad" Chapman et Oma Simpson. Son principe est de venir directement percevoir les gestes du conduit vocal qui constituent le signal de parole en posant notre main sur le visage du locuteur. Ainsi, à l'aide du pouce placé à la verticale sur les lèvres, l'auditeur peut ressentir les modifications liées aux mouvements des lèvres, aux vibrations nasales et au flux d'air sortant, tandis qu'avec les autres doigts placés le long de la mâchoire et sur la gorge il recueille des renseignements sur les mouvements de la mandibule et les vibrations des plis vocaux (voir Figure 12).

Figure 12 : Methode TADOMA utilisée par Helen Keller, célèbre écrivaine et conférencière américaine, sourde et aveugle (photo issue du livre « See what I'm saying : the extraordinary power of our five senses » de Rosenblum, L. (2010)).

Grâce à toutes ces informations tactiles disponibles, Reed et ses collègues (1985, 1992) ont montré que des participants sourds-aveugles entraînés à la méthode Tadoma pouvaient reconnaître entre 65% et 85% des mots d'une phrase. Cependant, la complexité de la phrase ou du contexte semblait particulièrement défavorable à la bonne compréhension du discours (pas plus de 60% de reconnaissance pour les meilleurs participants du groupe). Ils ont également montré que les scores les plus élevés étaient obtenus pour un débit de parole assez lent (entre 2,5 et 3,5 syllabes par seconde par rapport à 6 syllabes par seconde pour un débit considéré comme normal). Finalement ils ont montré que les traits de voisement, de nasalité et de lieux d'articulation étaient les mieux identifiés, avec une bonne distinction des phonèmes suivants [/p/, /b/, /m/], [/t/, /d/, /n/] et [/k/, /g/].

Cependant, ces études ont été réalisées sur des sujets sourds-aveugles entraînés, et la question d'une possible sensibilité tactile plus grande chez ces personnes a été soulevée par

Sato et collègues en 2010. Leur expérience n'a cependant montré aucune différence de sensibilité entre des personnes voyantes non entraînées et des personnes aveugles non entraînées à la méthode Tadoma lors de la perception de stimuli très simples comme des voyelles de type VCV. De son côté, Morin, en 2011 a montré qu'un entraînement quasi quotidien à la méthode Tadoma pouvait améliorer la perception des consonnes bilabiales (/p/, /b/, /m/ sont déjà très bien perçues sans entraînement) et alvéolaires (/t/, /d/, /n/ sont très mal perçues au début de l'expérience, mais une nette amélioration a été observée au bout d'un mois et demi d'entraînement), cependant les scores d'identifications restent médiocres pour les consonnes articulées à l'arrière du conduit vocal comme /k/ et /g/, même à la fin de l'entraînement. Les consonnes nasales sont quant à elles de mieux en mieux perçues au fil des séances.

Cependant, à l'instar du signal visuel de parole, les informations tactiles ne sont pas suffisantes pour percevoir le message linguistique dans sa totalité. On peut alors se demander si, tout comme la vision, l'information tactile peut influencer l'audition. Nous allons voir dans la section suivante des études qui se sont penchées sur cette question, à l'aide de la méthode Tadoma, d'un jet d'air ou de l'étirement de la peau du visage par exemple. Nous verrons également que des illusions tactiles peuvent être obtenues par l'ajout d'un son incongruent.

Interactions audio-tactile – Un certain nombre d'études se sont intéressées à l'influence de l'information tactile sur le signal auditif de parole en utilisant une procédure similaire à la méthode Tadoma. Fowler et Dekle (1991) sont les premiers à montrer que l'absence de connaissance sur une modalité comme la perception tactile des mouvements de parole n'empêchait pas la fusion des informations haptiques et auditives, mais ce pour un faible nombre de sujets seulement. De façon intéressante, ils ont observé que l'influence de ces deux modalités était bi-directionnelle, c'est-à-dire que non seulement sentir les mouvements du conduit vocal affectait le jugement de la syllabe entendue, mais percevoir la syllabe auditivement affectait également le jugement de la syllabe manuellement perçue.

De leur côté, Gick et collègues en 2008 ont montré, malgré de grandes variations interindividuelles, que l'information tactile permettait d'améliorer de 10% la perception de la parole dans un milieu bruité par rapport à une perception auditive seule. Un gain identique à celui obtenu en comparant une modalité visuo-tactile à la perception auditive seule. Une amélioration perceptive similaire en cas de perception auditive bruitée a également été retrouvée par Sato et collègues en 2010 suite à une tâche d'identification phonémique dans trois conditions différentes : auditive seul, audio-tactile congruente et audio-tactile incongruente. Cependant, à la différence de Fowler et Dekle (1991), ils n'ont pas observé d'illusion de type McGurk suite à la présentation des stimuli audio-tactiles incongruents. Ces résultats suggèreraient selon eux des mécanismes d'intégration différents selon les modalités de présentation : audio-visuelle ou audio-tactile.

Mais l'accès direct aux mouvements du conduit vocal, grâce à la méthode Tadoma, n'est pas la seule manière d'influencer tactilement la perception auditive d'un stimulus de parole. Par exemple, Gick et collègues en 2009 ont montré qu'un simple jet d'air envoyé sur la main ou sur le cou pouvait modifier la perception de la parole. En effet une amélioration ou une perturbation a été observée lors de l'application d'un jet d'air en fonction de la nature « aspirée » ou non de la syllabe présentée (en anglais, /pa/ et /ta/ sont considérées comme aspirées, le flux d'air améliore leur perception, tandis que /ba/ et /da/ sont considérées

comme non aspirées et leur perception est alors perturbée par le flux d'air). En 2010, ils ont de plus montré que cette intégration audio-tactile ne pouvait se réaliser que dans une fenêtre temporelle relativement restreinte et correspondant à la fenêtre réelle de perception d'un jet d'air de parole (le jet d'air ne doit pas apparaître plus tôt que 50ms et pas plus tard que 200ms après l'onset acoustique). Ce résultat montre une certaine similitude avec les mécanismes d'intégration audio-visuelle qui ne supportent pas non plus un grand décalage entre le signal visuel et le signal auditif (Schwartz et Savariaux, 2014 ; van Wassenhove et al., 2005).

Enfin, Ito et collègues, en 2009, ont montré par une autre technique l'influence des informations tactiles sur la perception auditive de la parole. Lors de leur étude, les participants ont été soumis à différentes perturbations somatosensorielles relatives à l'ouverture de la mandibule, lors d'une tâche de jugement de mots. Les premier et second formants (F1 et F2) des mots « head » et « had » ont été synthétisés de façon à créer un continuum composé de 10 items allant de « head » à « had » et les participants avaient pour tâche d'identifier chacun des stimuli présentés comme faisant partie de l'une ou l'autre des catégories. Leurs résultats montrent que l'application d'une perturbation somatosensorielle relative à l'ouverture de la mandibule interférait avec la perception catégorielle de ces deux stimuli. Plus la perturbation était associée à une ouverture importante de la mandibule, plus la perception des sujets était biaisée vers les stimuli de type "had" dont la production de la voyelle ouverte implique une ouverture mandibulaire plus importante que celle des stimuli de type "head".

Mais il n'y a pas qu'en parole que les chercheurs ont pu mettre en évidence des interactions audio-tactiles. Ainsi, Schürmann et ses collègues (2004) ont montré que la perception de l'intensité d'un son pouvait être modulée par l'application simultanée d'une légère vibration sur la main. Dans leur étude, un son de référence était présenté aux sujets avec une intensité définie, puis un deuxième son leur parvenait alors qu'ils avaient pour consigne soit de toucher un tube vibrant simultanément avec le second son soit ne pas le toucher. Les sujets devaient alors ajuster l'intensité du deuxième son grâce à un potentiomètre pour qu'elle égale celle du son de référence. Les résultats suggèrent que la vibration ressentie lorsque les sujets touchaient le tube vibrant modifie la perception de l'intensité du son puisqu'ils choisissaient généralement des intensités plus faibles quand ils touchaient le tube que lorsqu'ils ne le touchaient pas. Cela supposerait que la vibration facilite l'intégration audio-tactile. De leur côté, Jousmäki et Hari (1998) ont observé une modification de la sensation tactile des mains lors de la présentation d'un son modifié artificiellement (le son de base était celui d'un frottement de main). Les sujets rapportaient alors une sensation de dessèchement des mains, d'où le nom de «Parchment-skin illusion ». Finalement, Bresciani et collègues (2005), à travers la présentation d'un nombre de beeps cohérents ou non avec le nombre de tapes reçu sur l'index, ont montré que le nombre de beeps entendu influençait le nombre de tapes perçu. Ces résultats soulignent de nouveau une forte interaction entre les informations tactiles et auditives même en dehors du domaine de la parole.

Pris ensemble, ces études démontrent que la parole peut aussi être perçue par le biais d'informations tactiles et proprioceptives, notamment en cas de non-accès aux informations auditives et visuelles. La perception tactile permet de récupérer certains mouvements du conduit vocal, bien qu'à l'instar de la vision certains phonèmes articulés trop en arrière du conduit vocal ne peuvent être perçus correctement en touchant uniquement le visage comme c'est le cas pour la méthode Tadoma. Cette capacité à discriminer tactilement un

certain nombre de phonèmes sans aucun entraînement peut suggérer qu'un accès direct aux gestes articulatoires combiné à nos connaissances procédurales motrices de ces gestes nous permet de percevoir le langage dans une modalité peu commune. D'autre part, des études ont également montré une interaction entre les modalités auditive et tactile, les informations tactiles venant faciliter la perception auditive du message dans le cas de stimuli congruents, ou perturber la reconnaissance du signal auditif dans le cas de stimuli incongruents. Cette influence serait bi-directionnelle, et l'intégration de ces deux flux sensoriels présenterait des similitudes avec les mécanismes d'intégration audio-visuelle. Dans cette thèse, nous avons voulu tester plus avant les similitudes et les éventuelles différences des processus d'intégration précoce audio-visuels et audio-tactiles à travers deux expériences en électroencéphalographie. Les articles relatifs à ces expériences seront présentés dans la section A de la partie expérimentale de ce manuscrit.

2. Perception et intégration audio-visuo-linguale de la parole

Nous venons de voir que l'on pouvait utiliser des informations tactiles – sans expérience préalable – relatives aux gestes du conduit vocal pour faciliter la perception auditive de la parole. Cette aptitude perceptive pourrait s'expliquer par notre habitude d'associer les informations tactiles et auditives au quotidien ainsi que par l'utilisation de nos connaissances motrices portant sur les gestes de parole. Nous allons maintenant examiner notre capacité à utiliser des informations visuelles inconnues, issues de mouvements de parole pourtant réalisés quotidiennement, à travers un organe clé en production de parole : la langue. Nous verrons qu'à l'aide de différentes techniques d'imagerie, nous pouvons accéder visuellement aux gestes de la langue, gestes produits à l'intérieur de notre conduit vocal. Nous verrons également que nous sommes en partie capables d'utiliser ces nouvelles informations pour faciliter la perception auditive de la parole sans doute grâce aux connaissances motrices dont nous disposons.

La langue en parole – La langue est un organe complexe constitué de pas moins de dix-sept muscles différents qui lui permettent de jouer un rôle central dans la mastication, la déglutition ainsi que lors de la production du langage. Elle est formée de deux parties principales, la *racine* de la langue qui forme comme son nom l'indique l'ancrage de la langue dans la cavité buccale au niveau de l'os hyoïde, et le *corps* de la langue qui comprend l'apex – la pointe de la langue – et le dos de la langue. Dans l'alphabet phonétique international (ou API), les phonèmes consonantiques des langues du monde sont classés suivant deux caractéristiques : le mode d'articulation (c'est-à-dire le type d'obstruction du passage de l'air utilisé, les résonateurs utilisés pour y parvenir ainsi que le voisement) et le lieu d'articulation (c'est-à-dire l'endroit où s'effectue l'obstruction du flux d'air et l'articulateur impliqué). La langue étant la partie la plus mobile du conduit vocal, c'est principalement elle qui sera à l'origine de l'obstruction totale ou partielle du flux d'air, en fonction de l'endroit où le contact se fait entre une partie de la langue et une zone de la cavité buccale. Voici quelques exemple d'occlusions totales du conduit vocal au contact de la langue : pour les alvéolaires /t/, /d/, /n/, c'est la pointe de la langue (ou apex) qui se pose derrière les dents sur la partie que l'on appelle les alvéoles, pour les vélaires /k/ et /g/, c'est le dos de la langue qui entre en contact avec le palais mou. Voici d'autres exemples d'occlusions partielles au contact de la langue : les fricatives palato-alvéolaires /ʃ/ et /ʒ/ dont le contact des bords de la langue avec les dents situées sur le côté créé un espace étroit au centre pour laisser le flux d'air s'échapper, ou encore la fricative uvulaire /ʁ/ qui résulte d'un contact partiel de l'arrière du dos de la langue avec la luette. Pour ce qui est des phonèmes vocaliques, l'aperture désigne

la position verticale de la langue par rapport au palais couplée avec l'ouverture mandibulaire (voyelles ouvertes ou fermées), le lieu d'articulation désigne l'emplacement de la langue dans le conduit (voyelles antérieures ou postérieures), et l'arrondissement des lèvres désigne la forme des lèvres (étirées ou arrondies). La voyelle /i/ est un exemple de voyelle fermée antérieure, pour laquelle la langue est positionnée à l'avant du conduit vocal et à proximité du palais, tandis que le /a/ est une voyelle ouverte centrale, pour laquelle la langue est au plus bas de la cavité buccale tout en étant placée au centre du conduit vocal, à contrario du /u/ par exemple, qui est une voyelle fermée postérieure et arrondie, pour laquelle la langue est proche du palais mais situé à l'arrière du conduit vocal et que les lèvres sont arrondies. La langue est un organe complexe qui est à l'origine d'une grande partie des sons que l'on peut produire, et pourtant sa position dans le conduit vocal ne nous permet pas d'avoir accès à ses mouvements visuellement de façon naturelle. Son étude est cependant très importante pour décrire les sons des langues du monde qui peuvent s'avérer parfois beaucoup plus complexes que ceux que nous avons en français. Différentes techniques d'investigation ont été mises en place, et nous allons présenter les trois principales dans la section suivante.

Les différents moyens de montrer les mouvements de la langue – Deux techniques sont principalement utilisées pour visualiser les mouvements de la langue : l'imagerie ultrasonore ou un avatar 3D. Nous allons les décrire brièvement.

Le principe de l'imagerie ultrasonore est d'utiliser les propriétés des ondes ultrasonores en fonction des milieux traversés et des matériaux rencontrés. On envoie une onde ultrasonore à l'aide d'une sonde sur la zone que l'on souhaite étudier et cette même sonde récupère les échos renvoyés par les différents milieux/matériaux rencontrés par l'onde. En connaissant les propriétés de réflexion, de réfraction, de diffusion et d'atténuation de chacun des milieux, on peut ainsi réinterpréter en termes de distance le temps recueilli entre l'émission de l'onde sonore et la réception des échos. Cela permettra d'obtenir une image en deux dimensions sous forme de coupe de la structure explorée (pour plus d'information, voir la thèse de doctorat de Thomas Hueber, 2009 et Hueber et al., 2008). Grâce à une sonde placée sous le menton du locuteur on peut ainsi obtenir une image de profil de la langue dans le conduit vocal (voir Figure 13).

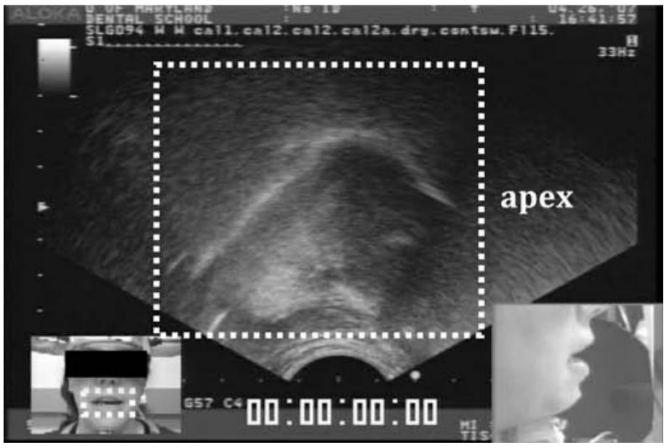

Figure 13 : Exemple d'une image ultrason de la cavité buccale. La langue (encadrée par les pointillés) est vue de profil, le dos de la langue est représenté par le trait plus clair avec l'apex à droite de l'image. Une vue des lèvres de face (bas gauche) et de profil (bas droite) du participant est également montrée (figure extraite de Hueber, 2009).

Les techniques développées aujourd'hui, combinant un dispositif d'enregistrement ultrason et une caméra, permettent d'enregistrer les mouvements de la langue en même temps que les mouvements du visage du locuteur et le son lié à ces productions. Cette méthode donne la possibilité d'étudier les mouvements réels des différents articulateurs du conduit vocal. Cependant, la résolution est nettement moins bonne que celle obtenue lors d'une acquisition d'images par résonance magnétique (IRM) et pour une personne non-initiée à cette technique, il est difficile d'interpréter correctement une image obtenue par ultrason (certaines études sur la perception des mouvements de la langue ont entrepris de tracer un contour de la langue sur l'image ultrason pour en faciliter la reconnaissance, voir d'Ausilio et al., 2014, par exemple).

L'Opti-speech (http://www.vulintus.com/; voir Figure 14) est une autre façon de présenter les mouvements de la langue grâce à un avatar (un visage virtuel) transparent en 3D dont la langue est apparente et colorée pour la mettre en valeur.

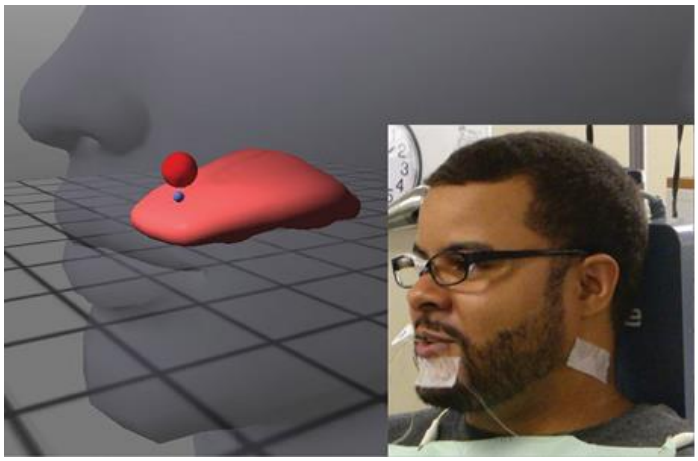

Figure 14 : Illustration du système « Opti-speech » composé de l'avatar 3D montrant la langue ainsi qu'une photo du participant et des capteurs positionnés sur son visage (Figure extraite de Katz et Mehta, 2015).

Ce système est utilisé au départ pour fournir un retour visuel des mouvements de la langue dans le cadre d'une rééducation de la parole. Il est donc nécessaire d'avoir un accès direct aux gestes de la langue du patient. Pour cela les expérimentateurs utilisent un système EMA (ElectroMagnetic Articulography) qui est composé de cinq capteurs placés généralement sur la langue, les lèvres et/ou la mâchoire et d'un récepteur qui enregistre le positionnement des capteurs en fonction d'un capteur de référence fixe attaché à des lunettes. Ces coordonnées une fois récupérées vont être envoyées en temps réel au logiciel qui gère l'avatar et permettront ainsi de bouger la langue virtuelle de l'avatar en fonction des mouvements réels du sujet (voir Katz et al., 2014 pour une description plus détaillée de l'Opti-speech et Katz et al., 2015 pour une expérience de perception utilisant cette méthode).

La tête parlante (Badin et al., 2010) est un visage artificiel, reconstitué à partir de différents modèles 3D des organes de la parole (langue, mâchoire, lèvres, vélum, visage etc…) construits à partir d'images par résonance magnétique ou d'images par rayons X de la tête d'un sujet.

Un logiciel permet de modifier précisément les paramètres de tous ces organes de façon à leur donner la forme réelle qu'ils auraient lors de la production d'un son de parole par

exemple. Nous pouvons donc sélectionner les organes que nous souhaitons afficher à l'écran, comme par exemple la langue. La tête parlante sera alors vue de profil tout comme l'avatar de l'Optispeech ou les images ultrasonores décrites précédemment. A la différence de ces deux techniques, ce visage virtuel permet de visualiser tous les organes les uns par rapport aux autres et de choisir les informations que nous souhaitons montrer (voir Figure 15).

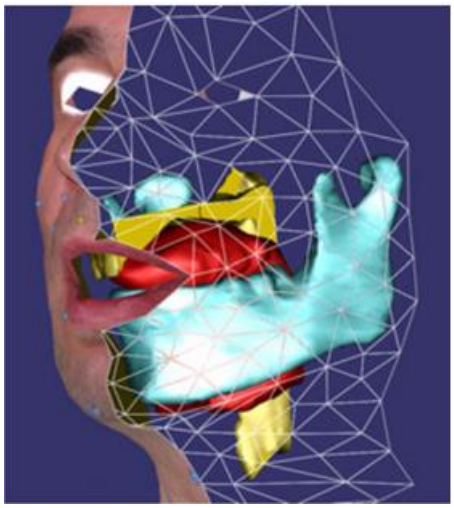

Figure 15 : Exemple de la tête parlante virtuelle. Le visage, la mâchoire, la langue, le palais et le conduit vocal sont visibles (Figure issue de Badin et al., 2010).

Ces trois systèmes de visualisation du conduit vocal ne sont qu'une sélection d'exemples parmi d'autres. Ils ont chacun leurs intérêts et leur limites, mais ils constituent une véritable avancée technique pour mieux montrer le fonctionnement interne de notre appareil articulatoire. Ils permettent, chacun à leur façon, de mettre en place des expériences uniques sur la perception visuelle de la langue en recherche fondamentale, pour l'apprentissage d'une langue étrangère ou en milieu clinique pour la rééducation des mouvements de la langue après une chirurgie par exemple.

Perception des mouvements de la langue – Voir les mouvements de la langue qui sont habituellement peu visibles voire totalement cachés peut aider à l'apprentissage d'une langue seconde. Il a déjà été établi que voir en plus d'entendre un locuteur s'exprimant dans une langue étrangère facilitait la compréhension du message linguistique et l'apprentissage de la langue utilisée (pour une revue, voir Marian, 2009). Katz et Mehta (2015) ont donc voulu tester si un retour visuel des productions de l'apprenant facilitait son apprentissage des sons de la langue seconde. Pour cela, un retour visuel des mouvements de la langue du participant, enregistrés grâce à un système de capture de mouvements (EMA), a été proposé aux participants durant une session d'entraînement à la prononciation du son cible de la langue étrangère. La mesure de la justesse des productions avec ou sans retour visuel des mouvements de la langue a permis de montrer un gain de rapidité dans l'apprentissage d'un nouveau phonème. Ainsi, voir leurs propres productions a permis aux participants de réajuster plus rapidement/facilement leurs gestes articulatoires pour parvenir à la cible.

De même, la vue des mouvements de la langue peut également faciliter dans une certaine proportion la perception auditive d'un son de parole. Wik et Engwall (2008) se sont intéressés à la reconnaissance de mots dans des phrases dégradées par du bruit dans trois conditions de présentation différentes : une présentation acoustique seule, une

présentation audio-visuelle avec une vue de face d'une tête synthétique et une présentation audio-visuelle avec à la fois la vue de face du visage synthétique et la vue de profil de ce visage dans lequel les mouvements de la langue sont rendus visibles par transparence. Leurs résultats sont assez mitigés, ils ne trouvent en effet pas de gain réel à percevoir les mouvements de la langue en plus des mouvements des lèvres, mais une faible amélioration a tout de même été observée lors de la perception de plosives palatales, c'est-à-dire de phonèmes articulés un peu plus à l'arrière du conduit vocal. Les auteurs suggèrent alors qu'on peut extraire certains traits articulatoires de la perception visuelle des mouvements de la langue, mais que cette modalité de présentation ne permet pas d'améliorer la perception audio-visuelle classique d'un son accompagné du visage du locuteur. De leur côté, Badin et collègues (2010) ont trouvé des résultats similaires, faibles mais présents lors d'une tâche d'identification à choix forcé de syllabes VCV (voyelle-consonne-voyelle) lors de la présentation d'une vue de profil d'un visage synthétique pour lequel la langue était visible.

Enfin, d'Ausilio et collègues (2014) se sont intéressés à la perception de stimuli audio-visuels congruents et incongruents, avec pour information visuelle les mouvements de la langue relatifs ou non au percept auditif. Ils ont présenté aux participants des vidéos d'images ultrasons représentant les mouvements de la langue vue de profil, accompagnés d'une syllabe auditive congruente ou non. Les résultats montrent une identification plus rapide des syllabes auditives lorsque les mouvements de la langue sont présentés simultanément, et cette facilitation temporelle est plus importante dans le cas d'une congruence entre la syllabe auditive et visuelle.

Dans cette section, nous nous sommes intéressés à un des articulateurs les plus utilisés pour produire des sons de parole mais qui ne fournit cependant que très peu d'informations visuelles lors de la lecture labiale. Nous avons une expérience motrice forte des mouvements linguaux de parole mais notre expérience visuelle de cet organe est extrêmement limitée. Grâce à différentes techniques d'imagerie, des chercheurs ont pu enregistré puis visualisé les mouvements de la langue et étudié l'influence que peut avoir l'observation visuelle de ces mouvements sur la perception de stimuli auditifs. Les résultats ont démontré que, malgré une faible connaissance visuelle préalable des mouvements de cet articulateur, l'association de ces informations au signal auditif de parole permet d'en améliorer quelque peu la perception. Ces études suggèrent indirectement un recrutement de nos connaissances motrices des gestes articulatoires pour percevoir la parole. Dans cette thèse, nous nous sommes intéressés aux réseaux neuronaux de la perception audio-visuelle labiale et audio-visuelle linguale de la parole. L'article relatif à cette expérience en IRMf sera présenté dans la section B de la partie expérimentale de ce manuscrit.

3. Perception et intégration de nos propres actions

Lors de la toute première section de cette thèse, il a été montré qu'un grand nombre d'études suggéraient que l'expérience motrice de l'observateur jouerait un rôle non négligeable dans la perception de mouvements humain. En parole notamment, plusieurs théories et modèles neurobiologiques proposent un appariement des entrées sensorielles avec les connaissances motrices que possède déjà l'auditeur pour traiter le signal de parole. Ces connaissances motrices ont été acquises à force d'expériences et de répétitions de mouvements que nous avons exécutés (et entendus) tout au long de notre vie. Si nous suivons cette logique d'association entre les entrées sensorielles et les connaissances

procédurales motrices, nous pouvons attendre un meilleur appariement entre des signaux de parole que nous aurions nous-mêmes produits et les gestes moteurs que nous avons en mémoire et qui nous ont permis de les produire. Cela se traduirait par une meilleure reconnaissance de nos propres actions. Nous allons voir que les expériences sur les actions humaines démontrent en effet une meilleure reconnaissance de nos propres mouvements, bien que dans le cas de la perception de parole les résultats soient plus mitigés.

Perception de nos propres actions – Nous avons tous déjà fait l'expérience de reconnaître quelqu'un de loin juste à sa démarche. Beardworth et Buckner (1981) ont montré qu'à l'aide uniquement de points lumineux placés à des endroits clés du mouvement, nous étions capables de reconnaitre notre propre marche parmi des vidéos de marche de personnes inconnues. Ces résultats montrent que malgré le fait que nous n'ayons aucune expérience visuelle de ces points lumineux et de notre marche, nous sommes non seulement capables de reconstituer un mouvement à partir des points lumineux, mais également d'y reconnaitre les idiosyncrasies propres à nos mouvements.

Loula et ses collaborateurs en 2005 ont voulu tester dans quelle mesure le système moteur et l'apprentissage perceptif contribuaient à l'analyse visuelle de mouvements humains. Pour cela, ils se sont basés sur trois hypothèses : 1) Si les expériences motrices influencent l'analyse visuelle de l'action alors l'observateur devrait être plus sensible à ses propres mouvements, 2) si l'expérience visuelle détermine la sensibilité aux mouvements humains alors l'observateur devrait être plus sensible aux mouvements des personnes qui lui sont familières et 3) si les expériences visuelles et motrices influencent toutes deux l'analyse visuelle des mouvements humains alors le sujet devrait mieux reconnaitre ses propres actions et celles de ses connaissances par rapport à celles d'un étranger. Les participants étaient donc soumis au visionnage de trois types de stimuli – toujours en utilisant le paradigme des points lumineux : leurs propres mouvements (grande expérience motrice), les mouvements de leurs amis (grande expérience visuelle) et les mouvements de personnes étrangères (aucune expérience visuelle ou motrice). Leurs résultats révèlent une meilleure identification des actions réalisées par le sujet, puis par les personnes qui lui sont proches et enfin, avec un score proche de la chance, par les personnes inconnues, montrant ainsi une influence de l'expérience motrice et de l'expérience visuelle dans la perception d'actions humaines.

Face à ces résultats, Knoblich et collègues (2001) montrent cependant qu'un grand apprentissage moteur, comme pour l'écriture par exemple, ne joue pas toujours un rôle principal dans la reconnaissance d'un caractère déjà appris ou non. Leur expérience était composée de deux parties (production et perception) : une première durant laquelle le participant devait réaliser une production écrite de caractères familiers (lettres de l'alphabet latin) ou non familiers (caractères inconnus) sans retours visuels, et une seconde durant laquelle le participant devait juger les productions perçues comme étant les siennes ou celles de quelqu'un d'autre. Les résultats ne montrent aucune différence de reconnaissance entre les caractères familiers (déjà appris) et non familiers (nouveaux). Cependant ils trouvent tout de même une meilleure reconnaissance de notre propre écriture (indépendamment de la familiarité des caractères présentés). Ces résultats suggèrent ici que l'apprentissage moteur intervient dans l'identification globale du mouvement, mais qu'il n'est peut-être pas suffisamment précis pour discriminer une partie seulement.

Identifier ses propres actions se traduit aussi par une synchronisation plus efficace du fait d'une reconnaissance accrue des particularités présentes dans le mouvement et de son rythme. Repp et son équipe, en 2004, ont ainsi demandé à des pianistes professionnels de jouer une première voix, puis de jouer une seconde voix en écoutant la première voix jouée par eux-mêmes ou par un autre pianiste. Les résultats observés ont montré que les pianistes parvenaient mieux à se synchroniser avec eux-mêmes suggérant une mise en résonance de ce qu'ils entendent avec le codage de leurs propres actions.

Prises ensembles, ces études montrent que malgré une faible connaissance visuelle de nos actions, nous sommes meilleurs pour percevoir nos propres mouvements par rapport à ceux de quelqu'un qui nous est familier – et a fortiori de quelqu'un d'inconnu. Cela vient soutenir l'idée selon laquelle il y aurait un couplage entre nos connaissances motrices et les gestes perçus, et que ce couplage serait particulièrement efficace lorsque l'objet de la perception proviendrait des mouvements que nous avons nous-mêmes produits. Mais nous allons voir que nous sommes également plus rapides pour identifier notre propre visage et notre propre voix. Cependant le peu d'études qui se sont intéressées à la perception de nos propres signaux de parole montrent des résultats mitigés.

Perception de notre propre parole – Se percevoir soi-même passe aussi par la reconnaissance de notre visage et de notre voix. Keenan et collègues (2000) ont observé lors de tâches d'identification de visages statiques et propres ou non aux participants une facilitation temporelle d'identification pour leurs propres visages. Des chercheurs ont également montré, par électroencéphalographie (Keyes et al., 2009), une différence temporelle dans le traitement des visages, plus précoce (environ 170ms après le début du stimulus) pour notre propre visage et plus tardif pour les visages familiers et inconnus (environ 250ms). Ces deux études suggèrent un traitement spécifique et plus rapide de la vue de notre propre visage.

Une étude en IRM fonctionnelle réalisée par Uddin et al. en 2005 sur l'observation de son propre visage par rapport à celui d'une personne familière vient conforter cette idée d'un processus particulier, relatif au traitement du « soi ». Elle montre en effet une plus grande activation d'un réseau fronto-pariéto-occipital incluant le gyrus frontal inférieur, le gyrus occipital inférieur et le lobule pariétal inférieur lorsque les participants percevaient leur propre visage (voir aussi Kaplan et al., 2008). Ce résultat suggère donc qu'il existerait des aires spécifiques aux représentations abstraites du « soi ».

De tous ces résultats, il pourrait être supposé une meilleure reconnaissance de sa propre voix. Mais face à l'implication du système moteur, une objection peut être faite cependant car nous entendons notre propre voix depuis la naissance, mais de façon déformée à cause de la conduction osseuse notamment. L'effet « soi » pourrait donc provenir d'un apprentissage auditif plutôt que de nos connaissances motrices.

Pour éclaircir cela, le signal visuel de parole apparaît pertinent, car si nous n'avons pas l'habitude de nous regarder parler, en revanche nous connaissons bien les idiosyncrasies contenues dans le signal de parole des personnes qui nous sont familières. Deux résultats sont donc possibles : soit l'expérience sensorielle prime, alors nous devrions mieux reconnaitre des stimuli visuels de personnes qui nous sont proches par rapport à nos propres productions, soit c'est l'expérience motrice qui est à l'origine de cet effet « soi » et nous devrions être meilleurs pour reconnaître nos propres productions.

Dans leur expérience précédemment citée, Keenan et collègues (2000) se sont également intéressés à la perception de notre propre voix. Ils ont montré que les participants étaient plus rapides à identifier leur propre voix que celle d'un inconnu (voir aussi Rosa et al., 2008). Graux et ses collègues en 2012 ont mené une étude de mismatch negativity (MMN) des potentiels évoqués auditifs et de la P3a, une onde positive connue pour être une mesure objective des capacités à reconnaitre différentes voix et maximale sur les électrodes fronto-centrales (Titova et Näätänen, 2001). Ils ont trouvé une MMN similaire pour les 2 conditions testées (sa propre voix vs. la voix d'autres personnes inconnues). En revanche une réponse pré-MMN apparaissant environ 70ms après de début du stimulus semble refléter un traitement relatif à sa propre voix (pas de pic pour les autres stimuli) et la P3a suivant la MMN montre bien une dissociation entre les deux conditions, avec une amplitude plus réduite pour le traitement de sa propre voix. De plus, en 2014, ils montrent la même différence entre des stimuli relatifs aux propres productions du sujet et aux productions de personnes qui lui sont familières. Il y aurait donc, similairement au traitement de notre propre visage, un traitement spécifique de nos propres productions vocales, ce qui soutient l'hypothèse selon laquelle il existerait un réseau neuronal spécifique pour le traitement du « soi » (voir Buckner et al., 2008; Frith and Frith, 1999; Lombardo et al., 2009; Northoff et al., 2006).

Intéressons-nous maintenant non plus à la seule identification de notre voix, mais bien à l'identification visuelle de nos propres productions de gestes de parole. D'après une expérience de Tye-Murray et collègues en 2012, nous serions meilleurs pour lire sur nos propres lèvres que sur les lèvres de quelqu'un d'autre et cette capacité ne serait pas reliée au fait d'être un bon « lecteur» ou non, ou bien d'être plus « lisible» ou non. Cependant, Aruffo et Shore (2012) n'ont trouvé aucun effet visuel du « soi » lors de la présentation de stimuli McGurk. En effet, si l'écoute de sa propre voix diminuait la force de l'illusion audio-visuelle, l'effet McGurk restait en revanche inchangé lors de la perception de son propre visage ou de celui d'un inconnu. Enfin, Bernier et collègues (2012) ont tenté également de démontrer un avantage perceptif sur la perception visuelle de stimuli de parole propres ou non aux sujets. Bien que plusieurs études aient été réalisées, ces chercheurs n'ont cependant trouvé aucun avantage à percevoir ses propres productions par rapport à celle d'une autre personne (comme Aruffo et Shore, 2012). Cette absence de résultats dans ces deux dernières études pourrait s'expliquer en partie par la différence des stimuli utilisés dans les expériences. En effet Tye-Murray et collègues ont choisi une tâche d'identification de mots dans des phrases tandis qu'Aruffo et collègues ainsi que Bernier et collègues ont demandé à leurs participants d'identifier des syllabes. La différence ainsi rapportée pourrait être due à des informations plus nombreuses dans l'articulation de mots ou de phrases (notamment supra-segmentales, liées possiblement aux caractéristiques rythmiques de la phrase) que dans la coarticulation bisyllabique.

Pour conclure, ces études montrent que des gestes produits par l'observateur lui-même étaient mieux perçus du fait d'une correspondance entre les connaissances motrices de l'observateur et les gestes produits et ce malgré une expérience visuelle moindre de ces mouvements. De plus, un traitement particulier et plus rapide du « soi » s'observe lors de la perception de notre propre visage et notre propre voix. Le traitement de la parole n'échappe pas à cette facilitation, cependant il semblerait que cet avantage visuel du « soi » ne soit pour l'instant visible qu'au niveau de la phrase. Dans cette thèse, nous avons voulu tester un possible effet du soi lors de l'intégration précoce des informations auditives et visuelles de

stimuli de parole très simples de type CV. L'article relatif à cette expérience en électro-encéphalographie sera présenté dans la section C de la partie expérimentale de ce manuscrit.

4. Préservation des mécanismes d'intégration avec l'âge

Le vieillissement sensoriel commence très tôt pour la vision et finit par atteindre tous les sens. L'acuité visuelle diminue précocement, mais c'est à un âge avancé que l'on constate de façon plus systématique ses effets néfastes sur le quotidien de la personne. A cela s'ajoute une presbyacousie, c'est-à-dire une diminution de l'acuité auditive due à l'âge qui rend les situations de communication très difficiles pour les personnes âgées et leur entourage. Nous allons voir dans un premier temps que tout ce qui nous permet habituellement de percevoir correctement la parole va se fragiliser avec les années, notamment notre capacité à distinguer de la parole dans le bruit et à lire sur les lèvres pour améliorer la perception du signal auditif. Cependant, nous verrons que l'intégration audio-visuelle est préservée chez les personnes âgées grâce à des mécanismes compensatoires.

Pertes sensorielles avec l'âge – La perte d'audition devient réellement gênante lorsqu'on observe une diminution de la compréhension de la parole dans des environnements bruités (Wong et al., 2009 ; Anderson et al., 2011). Cette difficulté provient de deux problèmes majeurs : une dégradation du système auditif (depuis la cochlée jusqu'aux aires auditives) qui se traduit notamment par une perte de la perception des hautes fréquences et une difficulté croissante à séparer les différents flux de parole pour extraire le signal cible. Mais il est à noter que cette difficulté à percevoir la parole dans le bruit se retrouve même chez des personnes âgées n'ayant pas de pertes auditives (Cruickshanks et al., 1998; Gordon-Salant et Fitzgibbons, 1993). Cela suggère que la compréhension de la parole dans le bruit ne peut pas être évaluée uniquement à l'aide d'un audiogramme qui ne teste que la perception de sons purs non bruités. Getzmann et collègues (2015) ont cherché à comprendre quels mécanismes cérébraux pouvaient être déficients pour augmenter ainsi la difficulté à percevoir de la parole dans le bruit. Ils ont analysé les corrélats électrophysiologiques d'un groupe de personnes âgées ayant de bonnes performances en perception dans le bruit, d'un groupe de personnes âgées ayant des mauvaises performances et d'un groupe contrôle de jeunes adultes. Ils ont ainsi observé un déficit du contrôle attentionnel (avec une P2 plus tardive), ainsi qu'une réduction des processus de traitement de la parole (avec une réduction de la N400) chez toutes les personnes âgées par rapport au groupe de jeunes adultes. La P2 pourrait refléter, en plus de l'intégration audio-visuelle de la parole que nous avons déjà abordée, une certaine allocation attentionnelle (Potts, 2004). La N400, quant à elle, est supposée être le reflet du traitement de stimuli significatifs (ou potentiellement significatifs) typiquement liés à la perception du langage. En revanche, les séniors ayant de bonnes performances montrent une meilleure allocation de l'attention ainsi qu'un plus grand contrôle inhibiteur (un complexe P2/N2 plus grand) que leurs pairs ayant des performances moindres en perception de la parole dans le bruit. Une perte sensorielle pure ne peut donc expliquer en totalité les problèmes que rencontrent les séniors pour comprendre un signal de parole dans un milieu bruité. Des déficits plus généraux semblent impliqués comme l'attention par exemple.

Mais nous avons vu dans la section B de cette partie théorique que le signal visuel de parole pouvait servir de support pour améliorer la compréhension de la parole, notamment dans le bruit. Qu'en est-il chez les personnes âgées qui présentent bien souvent, en plus de

problèmes d'audition, des problèmes de vue ? Le véritable problème n'est pas la perte sensorielle, qui peut se corriger souvent facilement à l'aide de lunettes ou d'appareils auditifs, mais les déficits cognitifs plus généraux dus à – ou aggravé par – ces pertes sensorielles. La lecture labiale en est l'exemple le plus frappant : malgré un apprentissage rigoureux et long et une utilisation quasi quotidienne de cette capacité, avec l'âge, nous devenons de moins bons lecteurs labiaux. C'est en effet ce qu'ont montré Dancer et ses collègues (1994) en étudiant les performances de lecture labiale d'un groupe de personnes âgées (60-69 ans) et d'un groupe de jeunes adultes (20-29 ans). Les résultats montrent que les séniors sont nettement moins bons que les jeunes pour comprendre visuellement des phrases de tous les jours. Tye-Murray et collègues ont répliqué ces résultats en 2007 en étendant l'évaluation des performances en lecture labiale à des mots isolés puis à des syllabes du type VCV. Ces deux équipes de chercheurs s'étaient également intéressées à l'effet du genre sur les performances en lecture labiale, partant de la constatation que les hommes montreraient un déclin de l'acuité auditive plus grand que les femmes. Les résultats sont cependant mitigés car Dancer et collègues montrent en effet que les femmes âgées ont de meilleurs scores de reconnaissance que les hommes, tandis que Tye-Murray et collègues ne trouvent aucun effet du genre dans leurs résultats.

Pris ensemble, ces résultats montrent qu'avec l'âge la communication devient un véritable challenge. Une accumulation des difficultés, tant au niveau auditif que visuel, entraîne une perturbation de la compréhension du message linguistique. Cependant, nous allons voir que les séniors sont capables de tirer parti au maximum des informations disponibles et de compenser ces déficits par différents mécanismes.

Intégration audio-visuelle préservée – En dépit d'un déficit sensoriel et cognitif, la capacité d'unifier un percept visuel et un percept auditif semble préservée chez les séniors. Laurienti et collègues (2006) ont montré une facilitation temporelle à percevoir un stimulus audio-visuel (une association d'un disque de couleur et du nom de la couleur du disque prononcée à voix haute) par rapport à un stimulus auditif seul (le nom de la couleur prononcé à voix haute) chez les personnes âgées comme chez les jeunes, de plus les séniors ont montré un gain plus important à percevoir les deux modalités plutôt qu'une seule. D'autre part, l'intégration multisensorielle qui en résulte améliore la rapidité de perception de la modalité la mieux perçue en condition seule (ici la modalité visuelle). Cela suggère que cette fusion n'est pas le résultat d'un compromis entre la vitesse et la précision de la réponse.

Cienkowski et Carney (2002) se sont intéressés à l'intégration des informations auditives et visuelles chez les personnes âgées à travers l'identification de stimuli McGurk. Les séniors sont aussi bons que les jeunes adultes pour intégrer le signal visuel et le signal auditif de parole (i.e. le nombre de réponses relatives à un percept fusionné n'est pas différent à travers les deux groupes). Cependant une différence entre les personnes âgées et les jeunes est apparue au niveau du choix des réponses. En effet, les auteurs ont remarqué que les séniors se basaient plus sur l'information visuelle lorsque les stimuli étaient incongruents tandis que les jeunes orientaient leur choix plutôt vers l'information auditive. Ils ont également retrouvé une diminution des capacités en lecture labiale chez les personnes âgées qui pourtant n'affecte pas leur capacité à intégrer des informations multisensorielles. Finalement, Tye-Murray et collègues (2010) ont également montré une meilleure reconnaissance des stimuli audio-visuels par rapport aux stimuli auditifs seuls chez les personnes âgées.

De leur côté, Sekiyama et son équipe (2014) se sont intéressés à l'influence du visuel sur le signal auditif de parole. Ils ont ainsi comparé une population de personnes âgées sans perte auditive et une population de jeunes adultes lors d'une tâche de perception de syllabes dans quatre conditions différentes : auditive seule, visuelle seule, audio-visuelle congruente et audio-visuelle incongruente (du type McGurk). Les résultats de la première expérience montrent que les séniors sont plus sensibles à l'information visuelle de parole que les jeunes pour un rapport signal/bruit du signal acoustique identique. Dans une seconde expérience, ils ont ajusté le rapport signal/bruit afin que les séniors et les jeunes obtiennent les mêmes scores en perception auditive. Il s'agissait de mettre tous les participants à un même niveau de perception. Les résultats observés confirment ceux de leur première expérience. D'autre part, contrairement à Cienkowski et Craney (2002), ils n'ont trouvé aucune différence de performance en lecture labiale entre les deux groupes.

Ainsi, l'intégration audio-visuelle de signaux de parole semble préservée voire améliorée chez les personnes âgées, malgré le déclin sensoriel auquel elles sont souvent confrontées. Elles présentent un biais vers l'information visuelle bien que leur capacité en lecture labiale soit parfois moindre par rapport aux jeunes adultes. Ces résultats démontrent qu'elles prêtent plus attention aux indices visuels, et qu'elles peuvent les associer au signal auditif pour améliorer sa perception malgré des difficultés à extraire suffisamment d'information du signal visuel pour le comprendre seul. Cela suggère l'existence de mécanismes compensatoires qui prennent le relais lorsque les informations sensorielles ne sont plus suffisantes pour décoder la totalité du message linguistique. C'est ce que nous allons voir dans la prochaine section.

Mécanismes compensatoires – En l'absence de perte auditive, les personnes âgées montrent tout de même des troubles de la perception de la parole dans le bruit. Wong et collègues (2009) ont réalisé une étude portant sur les activations cérébrales liées à la perception dans le bruit chez les personnes âgées. Ils ont ainsi présenté une série de mots dans trois conditions différentes : signal auditif clair, signal auditif + bruit de conversations de plusieurs locuteurs avec un rapport signal/bruit de +20 dB, signal auditif + bruit de conversations de plusieurs locuteurs avec un rapport signal/bruit de -5dB. Les résultats comportementaux montrent une réduction des scores perceptifs uniquement lorsque le signal auditif est très bruité (la condition avec un rapport signal/bruit de -5dB) pour les personnes âgées par rapport aux jeunes adultes. L'observation des résultats IRMf montre une diminution de l'activité cérébrale dans le cortex auditif qui s'accompagne en revanche d'une augmentation des régions cognitives plus générales comme les aires liées à la mémoire de travail ou à l'attention. Il semblerait alors que pour compenser la diminution de l'efficacité de traitement des aires auditives, le cerveau des personnes âgées recrute des régions frontales impliquées normalement dans des mécanismes cognitifs plus généraux.

Cette notion de mécanismes compensatoires fait écho à l'hypothèse de déclin-compensation (Declin-Compensation Hypothesis) qui s'oppose à l'hypothèse de la cause commune (Common Cause Hypothesis). L'hypothèse de la cause commune suggère un déclin général des régions cérébrales, aussi bien des régions sensorielles que des régions cognitives de plus haut niveau. Or les résultats de l'équipe de Wong montrent une diminution de l'activité des régions sensorielles, mais une sur-activation des régions préfrontales. Le déclin des traitements neuronaux n'est donc pas généralisé, et le recrutement plus important d'autres régions reflèterait la mise en place de mécanismes compensatoires.

Le modèle HAROLD (Hemispheric Asymmetry Reduction in OLDer adults) proposé par Cabeza en 2002 illustre bien ce phénomène. Il se base sur des recherches effectuées dans les domaines de la mémoire, de l'attention, des fonctions exécutives, de la perception visuelle et des traitements langagiers chez les personnes âgées (Grossman et al., 2002; Stebbins et al., 2002; Reuter-Lorrenz et al., 2000; Madden et al., 1999; Grady, 1994) et propose qu'avec l'âge, l'activité du cortex préfrontal devient moins latéralisée, le cerveau recrutant alors les régions homologues de l'autre hémisphère pour compenser le déclin des traitements neuronaux durant les tâches de mémoire épisodique. Des travaux ont étendu ce phénomène à d'autres régions cérébrales comme le gyrus temporal supérieur (STG) et le pôle temporal lors de jugements sémantiques de phrases (Berlingeri et al., 2013).

Dans cette dernière section, nous nous sommes intéressés à la perception et à l'intégration de la parole chez les personnes âgées. Nous avons vu que les interactions sociales devenaient parfois un véritable problème pour cette population à cause d'une perte sensorielle ainsi que d'une fatigue attentionnelle (et neuronale) diminuant leur capacité de traitement de la parole. Mais nous avons également montré les étonnantes facultés du cerveau à s'adapter à sa propre situation à travers la preuve d'une conservation des mécanismes d'intégration de la parole grâce à un recrutement de régions habituellement non utilisées, notamment préfrontales. Nous avons vu plus tôt que le système moteur pouvait être un support dans la perception de la parole, particulièrement en conditions difficiles. On aurait pu supposer alors une implication des régions motrices pour compenser ce déclin sensoriel naturel chez les personnes âgées. Cependant à ce jour, aucune étude n'a réellement démontré de rôle majeur du système moteur comme support dans l'intégration des informations auditives et visuelles de parole. Dans cette thèse, nous nous sommes intéressés au rôle des régions motrices dans les mécanismes de perception uni- et multisensorielle de la parole au cours du vieillissement. L'article relatif à l'expérience en stimulation magnétique transcrânienne menée sur ce sujet sera présenté dans la section D de la partie expérimentale de cette thèse.

PARTIE EXPÉRIMENTALE - A

ETUDES EEG SUR LA PERCEPTION AUDIO-TACTILE

Dans ce chapitre, nous présenterons, sous la forme de deux articles scientifiques publiés dans *Neuropsychologia* et *Frontiers in psychology* en 2014, deux études en électro-encéphalographie (EEG) que nous avons réalisées sur l'intégration audio-visuelle et audio-tactile de la perception de la parole (relatives à l'état de l'art proposé dans la première section D-1 de la partie théorique de cette thèse). Ces deux expériences ont été effectuées au sein du laboratoire Gipsa-lab de Grenoble, sous la direction de Marc Sato et Coriandre Vilain et avec la collaboration de Camille Cordeboeuf, au sein de la plateforme EEG NeuroSpeech de ce laboratoire.

A- Treille, A., Cordeboeuf, C., Vilain, C. & Sato, M. (2014a). ***Haptic and visual information speed up the neural processing of auditory speech in live dyadic interactions.*** **Neuropsychologia, 57: 71-77**

B- Treille, A., Vilain, C. & Sato, M. (2014b). ***The sound of your lips: electrophysiological cross-modal interactions during hand-to-face and face-to-face speech perception*****. Frontiers in psychology, 5(420): 1-9**

Nous avons vu que la parole n'était pas uniquement auditive et/ou visuelle, mais que nous pouvions également la percevoir grâce à la récupération d'informations tactiles liées aux gestes articulatoires, en posant notre main sur le visage du locuteur. A travers ces deux études EEG, nous nous sommes intéressés aux possibles interactions multimodales précoces en comparant les potentiels évoqués auditifs (PEAs) N1 et P2 lors de la présentation de syllabes auditives, audio-visuelles et audio-tactiles (technique adaptée de la méthode Tadoma). Compte-tenu de la précédence des informations visuelles et tactiles lors de la prononciation de stimuli de parole isoléstels que des syllabes de type CV, une modulation des PEAs N1 et/ou P2 pourrait suggérer l'utilisation d'indices visuels et tactiles prédictifs pour faciliter le traitement auditif de la parole. Nous avons donc cherché à savoir si les informations auditives et tactiles (rarement utilisées pour percevoir de la parole) suivaient les mêmes mécanismes que ceux utilisés lors de l'intégration audio-visuelle (Expérience A). Nous avons ensuite voulu préciser la nature de ces indices visuels et tactiles, c'est-à-dire s'ils étaient ou non spécifiques à la saillance perceptive des stimuli présentés (Expérience B).

Nous nous sommes donc intéressés à la modulation de l'amplitude et de la latence des PEAs N1 et P2 lors d'une tâche d'identification syllabique (/pa/ *vs.* /ta/ dans l'expérience A et /pa/ *vs.* /ta/ *vs.* /ka/ dans l'expérience B en condition auditive seule, audio-visuelle et audio-tactile. Chacune des syllabes présentées au participant était prononcée en direct par l'expérimentatrice, fournissant ainsi une variabilité proche de celle que nous pourrions avoir lors d'une situation de communication naturelle.

Nos deux études ont montré une réduction de l'amplitude des PEAs en condition audio-visuelle et audio-tactile par rapport à une présentation auditive seule ainsi qu'une facilitation temporelle des PEAs lors de l'ajout des modalités visuelle et tactile par rapport à la condition auditive seule. Cependant, ces modulations retrouvées dans l'expérience B ne

sont pas corrélées avec la saillance perceptive des syllabes présentées. Du fait de la grande variabilité de nos stimuli prononcés en direct par l'expérimentatrice, cette absence de corrélation avec la saillance perceptive pourrait suggérer que la prédictibilité sensorielle n'est pas si claire en condition d'interaction « naturelle ». Ainsi, sans pour autant contredire l'hypothèse d'indices prédictifs présents dans les informations visuelles et tactiles, l'extraction de ces indices pourrait dépendre fortement de la variabilité des stimuli de parole utilisés.

Pour conclure, nos deux études ont permis de montrer que l'intégration audio-tactile de syllabes était en partie similaire à l'intégration audio-visuelle de la parole, avec une réduction de l'amplitude et de la latence des PEAs N1 et/ou P2 par rapport à une condition d'écoute seule. Nos résultats soulignent de plus un impact de la nature et du nombre de stimuli et d'exemplaires utilisés et une certaine prudence dans l'interprétation des résultats expérimentaux d'études utilisant peu d'exemplaires différents d'un même stimulus, la réalité conversationnelle pouvant être parfois plus variable.

Neuropsychologia 57 (2014) 71–77

Contents lists available at ScienceDirect

Neuropsychologia

journal homepage: www.elsevier.com/locate/neuropsychologia

Haptic and visual information speed up the neural processing of auditory speech in live dyadic interactions

Avril Treille *, Camille Cordeboeuf, Coriandre Vilain, Marc Sato

GIPSA-LAB, Département Parole and Cognition, CNRS and Grenoble Université, Grenoble, France

ARTICLE INFO

Article history:
Received 8 July 2013
Received in revised form
2 February 2014
Accepted 4 February 2014
Available online 11 February 2014

Keywords:
Audio–visual speech perception
Audio–haptic speech perception
Multisensory interactions
EEG

ABSTRACT

Speech can be perceived not only by the ear and by the eye but also by the hand, with speech gestures felt from manual tactile contact with the speaker's face. In the present electro-encephalographic study, early cross-modal interactions were investigated by comparing auditory evoked potentials during auditory, audio–visual and audio–haptic speech perception in dyadic interactions between a listener and a speaker. In line with previous studies, early auditory evoked responses were attenuated and speeded up during audio–visual compared to auditory speech perception. Crucially, shortened latencies of early auditory evoked potentials were also observed during audio–haptic speech perception. Altogether, these results suggest early bimodal interactions during live face-to-face and hand-to-face speech perception in dyadic interactions.

1. Introduction

Interactions between auditory and visual modalities are beneficial in daily conversation. Visual speech information is known to effectively improve speech intelligibility in noise (Benoît, Mohamadi, & Kandel, 1994; Sumby & Pollack, 1954), the understanding of a semantically complex acoustic statement (Reisberg, McLean, & Goldfield, 1987) or a foreign language (Navarra & Soto-Faraco, 2005). Furthermore, seeing incongruent articulatory gestures may also modify auditory speech perception (McGurk & MacDonald, 1976). The fact that visual input may facilitate or even change the perceiver's auditory experience thus provides clear evidence for audio–visual integration in speech processing.

Despite no current agreement between theoretical models of audio–visual speech perception regarding the processing level at which the acoustic and visual speech signals fuse to a unified speech percept (for a review, see Schwartz, Robert-Ribes, and Escudier (1998)), recent electro-encephalographic (EEG) and magneto-encephalographic (MEG) studies demonstrate that early auditory evoked potentials N1 and P2 are attenuated (Arnal, Morillon, Kell, & Giraud, 2009; Besle, Fort, Delpuech, & Giard, 2004; Klucharev, Möttönen, & Sams, 2003; Pilling, 2010; Stekelenburg & Vroomen, 2007; van Wassenhove, Grant, & Poeppel, 2005; Vroomen & Stekelenburg, 2010) and speeded up (van Wassenhove et al., 2005) when an auditory syllable is accompanied by visual information from the speaker's face. The speeding-up and amplitude suppression of auditory evoked potentials is thought to reflect early multisensory integrative mechanisms. Given the temporal precedence of visible speech movements on the auditory signal for isolated syllables, the observed effects on early auditory evoked potentials might be due to the increased temporal predictability of the onset of the auditory stimulus (Stekelenburg & Vroomen, 2007; Vroomen & Stekelenburg, 2010) and/or might reflect specific visual phonetic prediction of the incoming auditory syllable (Arnal et al., 2009; Arnal, Wyart, & Giraud, 2011; Arnal & Giraud, 2012; van Wassenhove et al., 2005).

From these studies, one fundamental issue is whether early cross-modal speech interactions only depend on well-known auditory and visual modalities or, rather, might also be triggered by other sensory modalities, namely the auditory and haptic modalities. Audio–haptic interactions are indeed frequently experienced in daily life, with auditory and tactile stimuli often perceived simultaneously (for instance, when we scratch ourselves, rub our hands together, knock at a door, or play a musical instrument). As in the McGurk audiovisual illusion (McGurk & MacDonald, 1976), incongruities between audio and tactile inputs may even result in unexpected percepts (Jousmäki & Hari, 1998). Regarding speech, past researches on the Tadoma method demonstrate that deaf-blind individuals can understand spoken language remarkably well through the haptic modality (Alcorn, 1932; Norton et al., 1977). In this method, speech is received by placing a hand on the face of the talker in order to monitor orofacial speech movements. Interestingly, a few behavioral studies also

* Correspondence to: Avril Treille, GIPSA-LAB, UMR CNRS 5216, Grenoble Université, 1180, avenue centrale, BP 25, 38040 Grenoble Cedex 9, France. Tel.: +33 476 827 784; fax: +33 476 824 335.
E-mail address: avril.treille@gipsa-lab.inpg.fr (A. Treille).

http://dx.doi.org/10.1016/j.neuropsychologia.2014.02.004
0028-3932 © 2014 Published by Elsevier Ltd.

provide evidence for audio–tactile speech interaction in individuals without sensory impairement, with inexperienced participants presented with syllables heard and felt from manual tactile contact with a speaker's face (Fowler & Dekle, 1991; Gick, Jóhannsdóttir, Gibraiel, & Mühlbauer, 2008; Sato, Cavé, Ménard, & Brasseur, 2010). Fowler and Dekle (1991) demonstrated the influence of tactile information on speech perception in a completely untrained population, with felt syllables affecting judgments of the syllable heard and, conversely, acoustic syllables affecting judgments of the syllable felt. Interestingly, they also found evidence for audio–haptic McGurk-type illusion but only in few participants (but see Sato et al. (2010). Gick et al. (2008) further showed that manual tactile information improves both auditory and visual speech intelligibility in noise. Similarly, Sato et al. (2010) demonstrated that manual tactile information relevant to recovering speech gestures enhances auditory speech perception in case of degraded acoustic information and that audio–tactile interactions occur similarly in blind and sighted untrained listeners.

The present electro-encephalographic study aimed at further investigating early cross-modal interactions through dyadic interactions between a listener and a speaker. We compared auditory evoked components in individuals without sensory impairement, not experienced in the Tadoma method, during auditory, audio–visual and audio–haptic speech perception during a forced-choice task between /pa/ and /ta/ syllables. To this aim, participants were seated at arm's length from an experimenter and they were instructed to manually categorize each syllable presented auditorily, visually and/or haptically.

Cross-modal speech interactions are usually thought to primarily depend on auditory and visual modalities, and have typically been attributed to the frequency with which event specific information from these two modalities are jointly encountered in daily conversation. To explore whether perceivers might integrate tactile information in auditory speech perception in a similar way as they do in visual information, we tested whether haptic and visual information from speech gestures both attenuate and speed-up early auditory evoked responses compared to auditory speech perception. Such evidence for early cross-modal interactions during both face-to-face and hand-to-face speech perception would further suggest that sensory information from speech gestures conveys predictive temporal and/or phonetic information to the incoming auditory speech input and would emphasize the multimodal nature of speech perception.

2. Methods

2.1. Participants

Two groups of fourteen and fifteen healthy adults, native French speakers, participated in the study (EEG experiment: 7 females, mean age of 34 years ± 11 years; behavioral experiment: 8 females, mean age of 28 years ± 9 years). All participants were right-handed, had normal or corrected-to-normal vision and reported no history of speaking, hearing or motor disorders. None of them was experienced in the Tadoma method.

2.2. Experimental procedure

2.2.1. EEG experiment

Early cross-modal speech interactions and auditory evoked components were first evaluated in an EEG experiment. The experimental procedure was adapted from the Tadoma method and similar to that previously used by Fowler and Dekle (1991), Gick et al. (2008) and Sato et al. (2010). Participants were individually tested in a sound-proof room and were seated at arm's length from a female experimenter (see Fig. 1A). They were told that they would be presented with /pa/ or /ta/ syllables either auditorily, visually, audio-visually, haptically, or audio-haptically over the hand–face contact.

Five modalities of presentation were tested. In the auditory modality (A), participants were instructed to keep their eyes closed and to listen to each syllable overtly produced by the experimenter. In the audio–visual modality (AV), they were asked to also look at the experimenter's face. In the audio–haptic modality (AH), they were asked to keep their eyes closed with their right hand placed on the experimenter's face (the thumb placed lightly and vertically against the experimenter's lips and the other fingers placed horizontally along the jaw line in order to help distinguishing both lip and jaw movements). The visual-only (V) and haptic-only (H) modalities were similar to the AV and AH modalities except that the experimenter silently produced each syllable. Because of no reliable acoustical triggers (see below), EEG data were not analyzed in the visual-only and haptic-only modalities.

The experimenter faced the participant and a computer screen placed behind the participant. On each trial, the computer screen specified the syllable to be produced. To this aim, the syllable was printed three times on the computer screen at 1 Hz, with the last display serving as the visual go-signal to produce the syllable. The inter-trial interval was 3 s. The experimenter previously practiced and learned to articulate each syllable in synchrony with the visual go-signal, with an initial neutral closed-mouth position and maintaining an even intonation, tempo and vocal intensity.

A two-alternative forced-choice identification task was used, with participants instructed to categorize each perceived syllable by pressing on one of two keys corresponding to /pa/ or /ta/ on a computer keyboard with their left hand. In order to dissociate sensory/perceptual responses from motor responses on EEG data, a brief single audio beep was delivered 600 ms after the visual go-signal (expecting to occur in synchrony with the experimenter production). Participants were told to produce their responses only after this audio go-signal.

The experiment included five individual experimental sessions related to each modality of presentation (A, V, H, AV, AH). Before each session, participants were informed about the modality of presentation. In each session, every syllable (/pa/ or /ta/) was presented 40 times in a randomized sequence for a total of 80 trials. The order of the modality of presentation and the response key designation were fully counterbalanced across participants. Before the experiment, participants performed few practice trials in all modalities. They received no instructions concerning how to interpret visual and haptic information but they were asked to pay attention to both modalities during bimodal presentation. Because the experimental procedure was quite taxing for the experimenter and the participants, short breaks were offered between each experimental session.

Presentation software (Neurobehavioral Systems, Albany, CA) was used to control the visual stimuli for the experimenter, the audio stimuli (beep) for the participant and to record key responses. In addition, all experimenter productions were recorded for off-line analyses.

2.2.2. Behavioral experiment

In order to test the temporal precedence of visible/tactile speech movements on the auditory signal for isolated syllables, reaction times (RTs) in a control behavioral experiment were evaluated in another group of fifteen participants during auditory, audio–visual and audio–haptic speech perception. Visual-only and haptic-only modalities were not included in the experiment because of no reliable acoustical triggers to estimate RTs. Importantly, the experimental procedure was perfectly identical to that used in the EEG experiment (with notably the same experimenter/speaker) except that the audio-go signal was removed and participants were instructed to categorize each perceived syllable as quickly as possible with their left hand. As in the EEG experiment, participants performed few practice trials in all modalities and were asked to pay attention to both modalities during bimodal presentation.

The experiment included three individual experimental sessions related to each modality of presentation (A, AV, AH). Before each session, participants were informed about the modality of presentation. In each session, every syllable (/pa/ or /ta/) was presented 20 times in a randomized sequence for a total of 40 trials. The order of the modality of presentation and the response key designation were fully counterbalanced across participants. Before the experiment, participants performed few practice trials in all modalities.

2.3. EEG acquisition

EEG data were continuously recorded from 64 scalp electrodes (Electro-Cap International, INC., according to the international 10–20 system) using the Biosemi ActiveTwo AD-box EEG system operating at a sampling rate of 256 Hz. Two additional electrodes served as reference (Common Mode Sense [CMS] active electrode) and ground (Driven Right Leg [DRL] passive electrode). One other external reference electrode was at the top of the nose. The electrooculogram measuring horizontal (HEOG) and vertical (VEOG) eye movements were recorded using electrodes at the outer canthus of each eye as well as above and below the right eye. Before the experiment, the impedance of all electrodes was adjusted to get low offset voltages and stable DC.

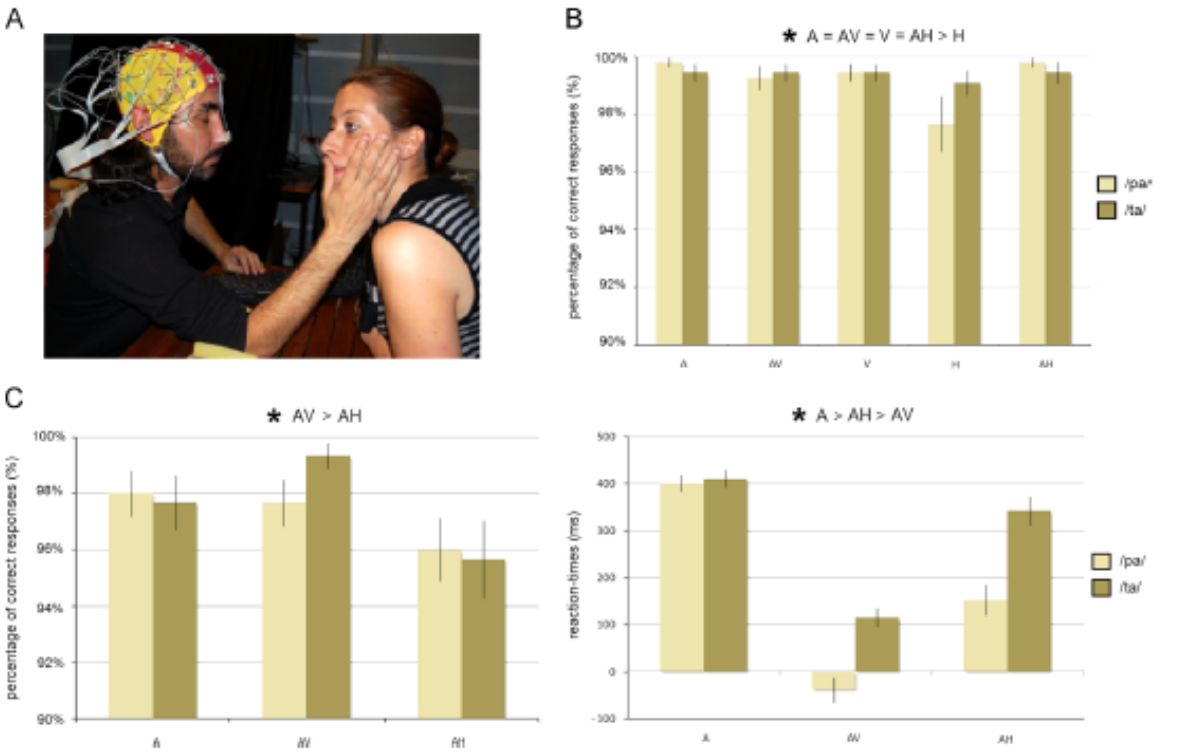

Fig. 1. (A) Experimental design used in the audio–haptic modality. Participants were asked to keep their eyes closed with their right hand placed on the experimenter's face and to categorize with their left hand each perceived syllable. (B) Mean percentage of correct identification for /pa/ and /ta/ syllables in each modality of presentation in the EEG experiment. (C) Mean percentage of correct identification and RTs (in ms) for /pa/ and /ta/ syllables in each modality of presentation in the behavioral experiment. Error bars represent standard errors of the mean.

2.4. Data analyses

2.4.1. Behavioral analyses

In both experiments, the proportion of correct responses was individually determined for each participant, each syllable and each modality. In addition, in the behavioral experiment, RTs were calculated from the consonantal onset of each produced syllable (see acoustical analyses). Two-way repeated-measure ANOVAs were performed on these data with the modality (A, V, H, AV, AH in the EEG experiment; A, AV, AH in the behavioral experiment) and the syllable (/pa/, /ta/) as within-subjects variables.

2.4.2. Acoustical analyses

In both experiments, acoustical analyses were performed on the experimenter's recorded syllables in order to determine the individual syllable onsets serving as acoustical triggers for EEG and RT analyses in the EEG and behavioral experiments, respectively. In the EEG experiment, because the experimenter silently produced the syllables in the V and H modalities, acoustical analyses were only performed for A, AV and AH modalities.

All acoustical analyses were performed using Praat software (Boersma & Weenink, 2013). A semi-automatic procedure was first devised for segmenting the experimenter's recorded syllables in the A, AV and AH modalities (5160 utterances). This procedure involved the automatic segmentation of each syllable based on an intensity and duration algorithm detection. Based on minimal duration and low intensity energy parameters, the algorithm automatically identified pauses between each syllable and set the syllable's boundaries on that basis. For each syllable, these boundaries were further hand-corrected, based on waveform and spectrogram information. Omissions and wrong productions were identified and removed from the analyses (less than 1%).

The individual syllable onsets served as acoustical triggers for EEG and RT analyses. In addition, to determine possible production differences between modalities of presentation in the EEG experiment, the mean duration, relative intensity and f0 values (calculated from a period defined as ±25 ms of the maximum peak intensity of each syllable) averaged over /pa/ and /ta/ syllables, as well as the mean delay between the visual go-signal and the produced syllable were then calculated for each participant and each modality. These data were entered into one-way repeated-measure ANOVAs with the modality (A, AV, AH) as the within-subjects variable.

2.4.3. EEG analyses

Because of no reliable acoustical triggers, EEG data were not analyzed in the visual-only and haptic-only modalities. EEG data in the A, AV and AH modalities were processed using the EEGLAB toolbox (Delorme & Makeig, 2004) running on Matlab (Mathworks, Natick, MA, USA). Since N1/P2 auditory evoked potentials have maximal response over fronto-central sites on the scalp (Näätänen & Picton, 1987; Scherg & Von Cramon, 1986), and in line with previous EEG studies on audio–visual speech perception and auditory evoked potentials (e.g. Pilling (2010), Stekelenburg and Vroomen (2007), van Wassenhove et al. (2005), Vroomen and Stekelenburg (2010)), EEG data preprocessing and analyses were conducted on 6 representative frontal and central electrodes (F3, Fz, F4, C3, Cz, C4). EEG data were first re-referenced off-line to the nose recording and band-pass filtered using a two-way least-squares FIR filtering (1–20 Hz). Data were then segmented into epochs of 1000 ms including a 100 ms prestimulus baseline (from −500 ms to −400 ms to the acoustic syllable onset, individually determined from the acoustical analyses). Epochs with an amplitude change exceeding ±100 μV at any channel (including HEOG and VEOG channels) were rejected (on average, 2% ± 3%).

Because of an insufficient number of trials per syllable for reliable EEG analyses, responses from /pa/ and /ta/ syllables were averaged together. For each participant and each modality, the EEG waveforms of the six electrodes were first carefully inspected by two experimenters. Two temporal windows were then defined in order to include N1 and P2 peaks for all electrodes (on average, 90–130 ms for N1 and 180–220 ms for P2). From these temporal windows, maximal amplitude and peak latency of auditory N1 and P2 evoked responses were then determined for the 6 electrodes. The mean latencies for N1 and P2 peaks were of 116 ms (±6 ms)/209 ms (±7 ms), 111 ms (±5 ms)/209 ms (±7 ms) and 105 ms (±6 ms)/201 ms (±7 ms) in the A, AV and AH modalities, respectively. Three-way repeated-measure ANOVAs were performed on N1 and P2 amplitude and latency with the modality (A, AV, AH), the rostro-caudal position (frontal, central) and the medio-lateral position (left, middle, right) of the electrodes as within-subjects variables.

3. Results

For all the following analyses, the significance level was set at $p=.05$ and Greenhouse–Geisser corrected (for violation of the sphericity assumption) when appropriate. When required, posthoc analyses were conducted with Newman–Keuls tests.

3.1. Behavioral analyses (see Fig. 1)

3.1.1. EEG experiment (see Fig. 1B)

Overall, the mean proportion of correct responses was of 99%. The main effect of modality of presentation was significant ($F(4,52)=3.63$, $p<.01$), with more correct responses in the A, V, AV and AH modalities than in the H modality (as shown by post-hoc analyses, all comparisons significant; on average, A: 100%, V: 99%, AV: 99%, AH, 100%, H: 98%). No significant effect of the syllable or interaction was observed. These results thus confirm a near perfect identification of the perceived syllables in all modalities, although a slightly lower accuracy in the H modality was observed (on average, 2%).

3.1.2. Behavioral experiment (see Fig. 1C)

The mean proportion of correct responses was of 97%. The main effect of modality of presentation was significant ($F(2,28)=3.34$, $p=.05$), with more correct responses in the AV than in the AH modality (as shown by post-hoc analyses; on average, A: 98%, AV: 99%, AH: 96%. No significant effect of the syllable or interaction was observed.

Regarding RTs, the main effect of modality was significant ($F(2,28)=94.81$, $p<.001$), with faster RTs observed in the AH than in the A modality, and in the AV modality than in the AH modality (as shown by post-hoc analyses, all comparisons significant; on average, A: 406 ms, AV: 39 ms, AH: 248 ms). Faster RTs were also observed for /pa/ compared to /ta/ syllables ($F(1,14)=74.10$, $p<.001$). Finally, the interaction between the modality and the syllable was reliable ($F(2,28)=36.38$, $p<.001$), with no differences between /pa/ and /ta/ syllables in the A modality but faster RTs for /pa/ than for /ta/ in both the AV and AH modalities (as shown by post-hoc analyses; on average, A-/pa/: 401 s, A-/ta/: 411 ms, AV-/pa/: −38 ms, AV-/ta/: 116 ms, AH-/pa/: 153 ms, AH-/ta/: 343 ms).

In sum, despite near perfect syllable recognition (although lower in the AH modality), faster RTs in both the AV and AH modalities provide evidence for the temporal precedence of the dynamic configurations of the articulators on the auditory signal. RTs observed in the AV modality were around 400 ms faster than in in the auditory modality. Compared to previous studies showing that the visual signal typically precedes the onset acoustic speech signal by tens to a few hundred milliseconds (see van Wassenhove et al. (2005)), this indicates that movements of the experimenter/speaker were here highly anticipated (likely due to the experimental procedure and, more specifically, to the 1 Hz visual go-signals for the experimenter). It is also worthwhile noting that negative RTs for /pa/ syllables demonstrate that participants under time pressure (and with subsequent auditory feedback) even recognized the syllable on a visual basis. Interestingly, the haptic advantage was not as strong as the visual one. Given the causal relationship between visual and haptic onsets, this might indicate less natural and more complex processing to extract relevant speech information. Finally, faster RTs for for /pa/ than for /ta/ syllables in both the audio–visual and audio–haptic modalities are likely due to the strong visual and haptic perceptual saliencies of bilabial movements.

3.2. Acoustical analyses

The mean delay between the visual go-signal and the produced syllable was constant across A, AV and AH modalities ($F(2,26)=0.86$; on average, A: +3 ms, AV: +14 ms, AH: −12 ms). Similarly,

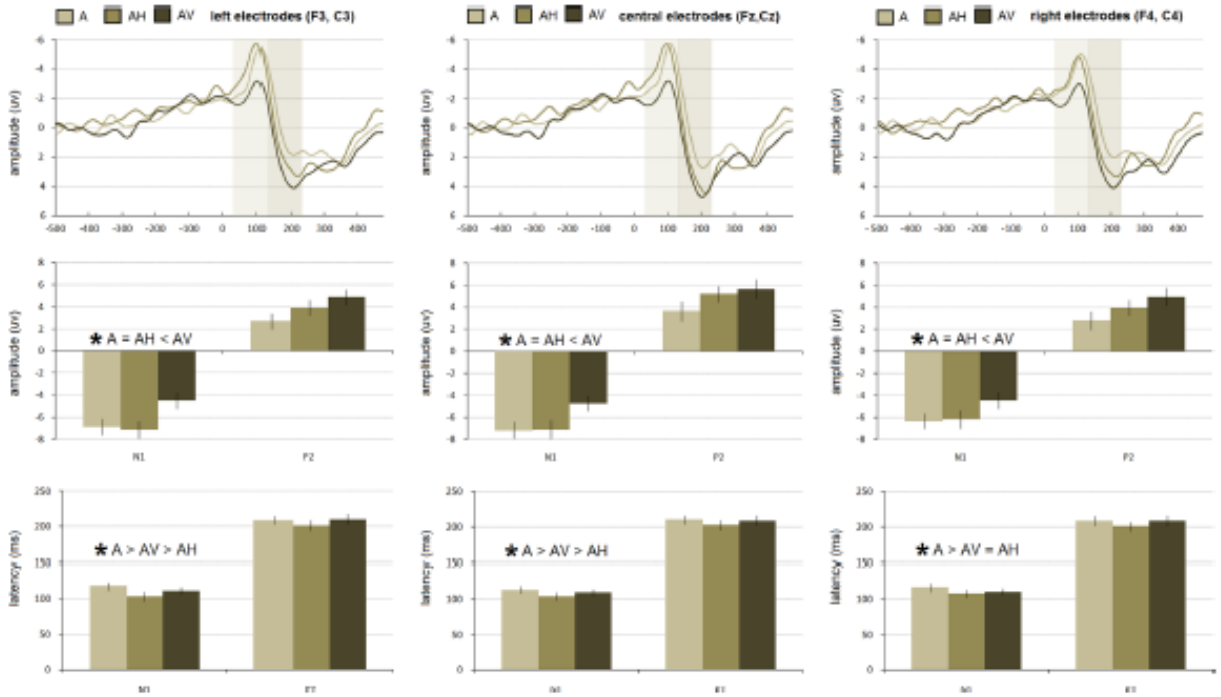

Fig. 2. Grand-average auditory evoked potentials (top), mean amplitude (in μV, middle) and mean latency (in ms, bottom) of N1 and P2 components averaged over left (F3, C3), central (Fz, Cz) and right (F4, C4) electrodes. EEG data were segmented into epochs of 1000 ms, from −500 ms to +500 ms to the acoustic syllable onset. Because of no reliable acoustical triggers, EEG data were not analyzed in the visual-only and haptic-only modalities. Error bars represent standard errors of the mean.

no differences were observed between modalities on the mean syllable duration ($F(2,26)=1.12$; on average, A: 194 ms, AV: 192 ms, AH: 197 ms) and intensity ($F(2,26)=3.47$; on average, A: 73 dB, AV: 73 dB, AH: 71 dB). However, the main effect of f0 was significant ($F(2,26)=5,93$, $p<.01$), with a lower f0 in the AH modality compared to A and AV modalities (as shown by post-hoc analyses, all comparisons significant; A: 241 Hz, AV: 240 Hz, AH: 237 Hz). These results show that, on average, the experimenter articulated the syllables in synchrony with the visual go-signal and maintained a quite constant intonation, tempo and vocal intensity across modalities and participants. Lower f0 observed in the AH modality is likely due to the participant's contact with the experimenter's face. However, this difference remains quite low (on average, 3 Hz) and, in our view, cannot explain latency and amplitude differences observed on EEG data between A, AV and AH modalities. Although the N1/P2 complex is known to depend on acoustic features of the speech signal, in our best knowledge, no EEG study demonstrated significant N1 and/or P2 latency and amplitude changes related to such a small f0 difference (i.e., 3 Hz) between isolated speech sounds. In addition, it can be noted that no f0 difference was observed between A and AV modalities despite significant difference on N1 amplitude and latency on EEG data between these modalities (see below).

3.3. EEG analyses—N1 amplitude (see Fig. 2—Middle)

The main effect of medio-lateral position was significant ($F(2,26)=6.49$, $p<.005$), with a reduced negative N1 amplitude observed in right electrodes as compared to left and middle electrodes (as shown by post-hoc analyses, all comparisons significant; on average, left: −6.17 μV, middle: −6.35 μV, right: −5.66 μV). Of more interest is the significant effect of modality ($F(2,26)=12.84$, $p<.001$), with a reduced negative N1 amplitude observed for the AV modality as compared to both A and AH modalities (as shown by post-hoc analyses, all comparisons significant; on average, A: −6.80 μV, AH: −6.84 μV, AV: −4.55 μV). The interaction between the modality and the medio-lateral position of electrodes was also reliable ($F(4,52)=4.33$, $p<.005$). For both A and AH modalities, a reduced negative N1 amplitude was observed in right electrodes as compared to both left and middle electrodes (as shown by post-hoc analyses, all comparisons significant; on average, A-left: −6.87 V, A-middle: −7.17 μV, A-right: −6.36 μV, AH-left: −7.16 μV, AH-middle: −7.13 μV, AH-right: −6.18 μV). However, for the AV modality, no differences were observed (on average, AV-left: −4.49 V, AV-middle: −4.73 μV, AV-right: −4.44 μV). No other effect or interactions were found to be significant.

The main effect of medio-lateral position confirms that auditory N1 had a central maximum (Näätänen & Picton, 1987; Scherg & Von Cramon, 1986) and possibly indicate left-lateralized auditory responses. Of more interest, the main effect of modality appears in line with previous EEG studies on audio–visual speech perception and confirm a visually-induced amplitude suppression of the auditory evoked N1 component. Interestingly, no haptically-induced amplitude suppression was observed, with similar amplitude for A and AH modalities and a higher negative N1 amplitude observed for the AH modality compared to the AV modality.

3.4. EEG analyses—N1 latency (see Fig. 2—Bottom)

The main effect of modality was significant ($F(2,26)=4.62$, $p<.02$), with a shorter negative N1 peak latency observed for the AH modality compared to the A modality (as shown by post-hoc analyses, all p's $<.05$; on average, A: 116 ms, AH: 105 ms, AV: 111 ms). The fact that the main effect did not provide evidence for shorter auditory evoked responses in the AV modality compared to the A modality is probably due response variability between the medio-lateral position of the electrodes. Indeed, a significant interaction between the modality and the medio-lateral position of electrodes ($F(4,52)=5.89$, $p<.001$) further demonstrate a shorter negative N1 peak latency for both AH and AV modalities compared to the A modality. Posthoc analyses showed that, in the left and middle electrodes, a shorter negative N1 peak latency was observed for the AH modality compared to the AV modality, and for the AV modality compared to the A modality (all p's $<.05$; on average, A-left: 118 ms, AV-left: 111 ms, AH-left: 104 ms, A-middle: 113 ms, AV-middle: 110 ms, AH-middle: 104 ms). In the right electrodes, a shorter negative N1 peak latency was observed for both the AH and AV modalities compared to the A modality (all p's $<.05$; on average, A-right: 116 ms, AV-right: 110 ms, AH-right: 108 ms). No other effects or interactions were significant. Altogether, these results appear in line with previous EEG studies with a visually-induced speeding-up of the auditory evoked N1 component. Similarly, a shorter negative N1 peak latency was also observed for the AH modality compared to the A modality, and compared to the AV modality in the left and middle electrodes.

3.5. EEG analyses—P2 amplitude and latency (see Fig. 2—Middle and bottom)

The analysis on P2 amplitude showed a significant effect of the medio-lateral position ($F(2,26)=14.56$, $p<.001$), with an higher positive P2 amplitude observed in middle electrodes as compared to both left and right electrodes (all p's $<.05$; on average, left: 3.83 μV, middle: 4.81 μV, right: 3.88 μV). In the same line to what was found with N1 amplitude, this effect confirms that auditory P2 had a central maximum. No other effects or interactions were found to be significant. Finally, regarding P2 peak latency, no effects or interactions were significant.

4. Discussion

In the present study, early cross-modal interactions were investigated by comparing early auditory evoked potentials and behavioral performance during auditory, audio–visual and audio–haptic speech perception in dyadic interactions between a listener and a speaker. Congruent with the possibility that face-to-face and hand-to-face dyadic interactions speed up the processing of auditory speech, reduced latency of early auditory evoked responses and faster reaction times were observed during both audio–visual and audio–haptic speech perception compared to auditory speech perception.

Before we discuss these results, it is important to consider a clear limitation of the present study. Since face-to-face and hand-to-face speech perception were here examined in live dyadic interactions, we used individual syllable onsets of the experimenter's productions as acoustical triggers for EEG and RT analyses. For the visual-only and haptic-only modalities, the use of electromyographic and/or visual recordings of the experimenter's lip movement would not allowed to determine such reliable triggers, due to the variability or temporal limitation of these signals (but see Leotta, Rabinowitz, Reed, and Drulach (1988), for a synthetic Tadoma system which allows realistic and precisely timed synthetic facial inputs). Because of no reliable triggers, these two unimodal conditions were not analyzed in the EEG experiment and were not included in the behavioral RT experiment. For that reason, we could not use an additive model (i.e., $AV \neq A+V$) to determine whether the results observed in the audio–visual and audio–haptic modalities simply come from a superposition of the sum of the unimodal signals or truly reflect crossmodal interactions. Notably, without EEG recordings in the visual-only and

haptic-only modalies, it is impossible to determine whether the observed N1 and P2 auditory evoked potentials in the audio–visual and audio–haptic modalities are not contamined by visual and haptic ERPs. From that question, although auditory ERPs are rarely observed in the visual-only modality in fronto-central electrodes (see Besle et al. (2004), Pilling (2010), van Wassenhove et al. (2005)), haptic ERPs (notably, the P30, P40, P100 and N140 components) are known to also occur in fronto-central electrodes (e.g., Desmedt and Tomberg (1989)). Importantly, it should be noted that RTs observed in the behavioral experiment strongly suggest that visual, and therefore haptic, speech movements of the experimenter/speaker started well before (around 400 ms) the acoustical onset of the syllables. Given the difference between sensory onsets, it is therefore unlikely that haptic (and visual) ERPs might arise at the same time-latency of auditory ERPs in the audio–haptic (and audio–visual) modality on fronto-central electrodes. Hence, although our results cannot fully demonstrate early bimodal integration mechanisms in the audio–visual and audio–haptic modalities, they strongly suggest that haptic and visual information speed up the neural processing of auditory speech.

In spite of this important limitation, our results appears fully in line with previous EEG studies on audio–visual speech perception (Besle et al., 2004; Klucharev et al., 2003; Pilling, 2010; Stekelenburg & Vroomen, 2007; van Wassenhove et al., 2005; Vroomen & Stekelenburg, 2010), with N1 auditory evoked potentials attenuated and speeded up during audio–visual compared to auditory speech perception. Given the temporal precedence of visible speech movements on the auditory signal for isolated syllables (the visual signal, and therefore the haptic signal, preceding the onset acoustic speech signal by tens to a few hundred milliseconds during individual syllable production; see van Wassenhove et al. (2005)), the speeding-up and amplitude suppression of auditory evoked potentials likely reflect early multisensory integrative mechanisms. Despite near perfect recognition, this temporal precedence of the dynamic configurations of the articulators on the auditory signal is attested in the behavioral experiment. Indeed, faster RTs were observed in both the audio–visual and audio–haptic modalities compared to the unimodal auditory modality.

Crucially, although participants were not experienced with audio–haptic speech perception, haptic information was also found to speed up auditory speech processing, with a shorter latency of N1 auditory evoked potentials in audio–haptic compared to auditory speech perception. Compared to a strong visually-induced N1 amplitude suppression observed in the present experiment and in previous studies on audio–visual speech perception, no haptically-induced N1 amplitude suppression was however observed. Although speculative, one possibility is that this difference might partly be explained by higher attentional demands in the audio–haptic modality, which is well known to enhance amplitude of early auditory evoked potentials (Näätänen, 1992; Giard et al., 2000).

Importantly, two qualitatively different integrative mechanisms, although not mutually exclusive, can been proposed to explain these findings. A first mechanism implies that early cross-modal audio–visual and audio–haptic interactions are not speech specific and rather depend on the temporal relationship of sensory input. Congruent with this hypothesis, Stekelenburg and Vroomen (2007) demonstrated visually-induced amplitude suppression and latency reduction of auditory-evoked N1 responses during audio–visual perception of both speech and non speech actions, like clapping hands. They further showed that early cross-modal interactions are not restricted to actions but can be observed also with artificial stimuli if their timing is made predictable (like two moving disks predicting a pure tone; Vroomen and Stekelenburg (2010)). It can be also noted that audio–tactile interactions are not restricted to speech and have been previously observed with simultaneous presented tone and vibration (e.g., Lütkenhöner, Lammertmann, Simões, and Hari (2002)). A second mechanism implies the existence of specific phonetic prediction, extracted from the visual signal, of the incoming auditory speech target. From that view, early auditory-visual latency facilitation has been shown to correlate with visual identification of the speech stimuli, with stronger latency facilitation coupled with higher visual identification. In a previous study by van Wassenhove et al. (2005), auditory-visual facilitation effects were shown to systematically vary according to the identification scores observed in the visual modality. In their study, a higher visual accuracy was observed for /pa/ compared to /ta/ syllables, and for /ta/ compared to /ka/ syllables. Consistent with an articulator-specific facilitation, latency of auditory evoked potentials were found to be shorter for /pa/ then for /ta/ syllables, and for /ta/ than for /ka/ syllables (for similar results using MEG, see also Arnal et al. (2009)). It is worthwhile noting that a potential limit of these studies comes from the use of a limited number of tokens used to represent each syllable, repeatedly presented to the participants and possibly enhancing stimulus predictability. Conversely, one clear limit of the present study comes from the use of a forced-choice task between /pa/ and /ta/ syllables and a ceiling effect on perceptual scores. Although our results do not contradict the hypothesis that sensory inputs convey phonetic predictive information with respect to the incoming auditory speech input, future studies using a larger sample of syllables are therefore required to test whether haptic information is used to specifically predict the incoming auditory syllable.

Despite the above-mentioned limitations of this study, the present and previously observed audio–haptic advantages in untrained participants may have profound implications for further understanding the basis of cross-modal speech integration (Fowler & Dekle, 1991; Gick et al., 2008; Sato et al., 2010). Since cross-modal speech interactions have previously been attributed to the frequency with which event specific information from auditory and visual modalities are jointly encountered in daily conversation, evidence for audio–haptic speech interactions, although less natural, further emphasize the multisensory and predictive nature of speech perception. It is first important to recall that, apart from speech, auditory and tactile stimuli are usually perceived simultaneously in daily life (as when knowing at a door). Although largely unexplored compared to audio–visual interactions, previous studies have provided evidence for an efficient integration of auditory and somatosensory processing at both the behavioral and neural levels (e.g., Desmedt and Tomberg (1989), Jousmäki and Hari (1998)). Importantly, the integration of sensory inputs can be viewed as a key feature of action control, as when we first knock on a door and then adjust the force applied in the subsequent knock (Wolpert, Ghahramani, & Jordan, 1995). Regarding speech, audio–haptic interactions likely depend on the temporal precedence of the dynamic configurations of the articulators on the auditory signal (as attested in the behavioral experiment). As previously mentioned, despite less natural and apparently more complex processing to extract relevant speech information from the haptic modality, this temporal precedence might increase the temporal predictability of the onset of the auditory stimulus and/or provide specific phonetic prediction of the incoming auditory syllable. The possibility that the brain might extract predictive temporal and/or phonetic relevant information for auditory processing when being provided with such information, although we rarely if ever touch the speaker's face to understand speech, raised important question on the representational format of speech. From that point, haptic perception is likely to be partly driven by listener's knowledge of speech production (Sato et al., 2010; Schwartz, Ménard, Basirat, & Sato, 2012), as for example

when relating the tactile sensation of lip protusion to the production of /pa/ syllables. Although speculative, thanks to the temporal precedence of tactile inputs, the observed audio–haptic interactions might also partly arise from auditory, visual and/or motor imagery processes and, possibly, a crossmodal mapping between the related sensory and motor speech representations.

In conclusion, our results suggest early integrative mechanisms between auditory, visual and haptic modalities. Since audio–haptic and audio–visual speech interactions were never assessed at the brain level in dyadic interactions between a listener and a speaker, these results provide new insights on the multisensory nature of speech perception.

Acknowledgments

This study was supported by research grant from the Centre National de la Recherche Scientifique (CNRS). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding agency. We thank Jean-Luc Schwartz and Marco Congedo for helpful discussions on multisensory speech perception and on EEG analyses.

References

Alcorn, S. (1932). The Tadoma method. *Volta Review*, 34, 195–198.
Arnal, L. H., & Giraud, A. L. (2012). Cortical oscillations and sensory predictions. *Trends in Cognitive Science*, 16(7), 390–398.
Arnal, L. H., Morillon, B., Kell, C. A., & Giraud, A. L. (2009). Dual neural routing of visual facilitation in speech processing. *The Journal of Neuroscience*, 29(43), 13445–13453.
Arnal, L. H., Wyart, V., & Giraud, A. L. (2011). Transitions in neural oscillations reflect prediction errors generated in audiovisual speech. *Nature Neuroscience*, 14(6), 797–801.
Benoît, C., Mohamadi, T., & Kandel, S. D. (1994). Effects on phonetic context on audio–visual intelligibility of French. *Journal of Speech and Hearing Research*, 37, 1195–1203.
Besle, J., Fort, A., Delpuech, C., & Giard, M. H. (2004). Bimodal speech: Early suppressive visual effects in human auditory cortex. *European Journal of Neuroscience*, 20, 2225–2234.
Boersma, P. & Weenink, D. (2013). *Praat: doing phonetics by computer. Computer program*, Version 5.3.42., retrieved 2 March 2013 from (http://www.praat.org/).
Delorme, A., & Makeig, S. (2004). EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics. *Journal of Neuroscience Methods*, 134, 9–21.
Desmedt, J. E., & Tomberg, C. (1989). Mapping early somatosensory evoked potentials in selective attention: Critical evaluation of control conditions used for titrating by difference the cognitive P30, P40, P100 and N140. *Electroencephalography and Clinical Neurophysiology*, 74(5), 321–346.
Fowler, C., & Dekle, D. (1991). Listening with eye and hand: Crossmodal contributions to speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, 17, 816–828.
Giard, M. H., Fort, A., Mouchetant-Rostaing, Y., & Pernier, J. (2000). Neurophysiological mechanisms of auditory selective attention in humans. *Front. Bioscience*, 5, 84–94.
Gick, B., Jóhannsdóttir, K. M., Gibraiel, D., & Mühlbauer, M. (2008). Tactile enhancement of auditory and visual speech perception in untrained perceivers. *Journal of Acoustical Society of America*, 123, 72–76.
Jousmäki, V., & Hari, R. (1998). Parchment-skin illusion: Sound-biased touch. *Current Biology*, 8(6), 190.
Klucharev, V., Möttönen, R., & Sams, M. (2003). Electrophysiological indicators of phonetic and non-phonetic multisensory interactions during audiovisual speech perception. *Brain Research. Cognitive Brain Research*, 18, 65–75.
Leotta, D. F., Rabinowitz, W. M., Reed, C. M., & Drulach, N. I. (1988). Preliminary results of speech-reception tests obtained with the synthetic Tadoma system. *Journal of Rehabilitation Research and Development*, 25(4), 45–52.
Lütkenhöner, B., Lammertmann, C., Simões, C., & Hari, R. (2002). Magnetoencephalographic correlates of audiotactile interaction. *NeuroImage*, 15(3), 509–522.
McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264, 746–748.
Näätänen, R. (1992). *Attention and Brain Function*. Hillsdale, NJ: Lawrence Erlbaum Associates.
Näätänen, R., & Picton, T. W. (1987). The N1 wave of the human electric and magnetic response to sound: A review and an analysis of the component structure. *Psychophysiology*, 24, 375–425.
Navarra, J., & Soto-Faraco, S. (2005). Hearing lips in a second language: Visual articulatory information enables the perception of second language sounds. *Psychological Research*, 71(1), 4–12.
Norton, S. J., Schultz, M. C., Reed, C. M., Braida, L. D., Durlach, N. I., Rabinowitz, W. M., et al. (1977). Analytic study of the Tadoma method: Background and preliminary results. *Journal of Speech and Hearing Research*, 20, 574–595.
Pilling, M. (2010). Auditory event-related potentials (ERPs) in audiovisual speech perception. *Journal of Speech, Language, and Hearing Research*, 52(4), 1073–1081.
Reisberg, D., Mclean, J., & Goldfield, A. (1987). Easy to hear but hard tounderstand: A lipreading advantage with intact auditory stimuli. In: R. Campbell, & B. Dodd (Eds.), *Hearing by eye: The psychology of lipreading* (pp. 97–113). London (UK): Lawrence Erlbaum Associates.
Sato, M., Cavé, C., Ménard, L., & Brasseur, L. (2010). Auditory-tactile speech perception in congenitally blind and sighted adults. *Neuropsychologia*, 48(12), 3683–3686.
Scherg, M., & Von Cramon, D. (1986). Evoked dipole source potentials of the human auditory cortex. *Electroencephalography and Clinical Neurology*, 65, 344–360.
Schwartz, J. L., Ménard, L., Basirat, A., & Sato, M. (2012). The perception for action control theory (PACT): A perceptuo-motor theory of speech perception. *Journal of Neurolinguistics*, 25(5), 336–354.
Schwartz, J.-L., Robert-Ribes, J., & Escudier, P. (1998). Ten years after summerfield: A taxonomy of models for audio–visual fusion in speech perception In: R. Campbell, B. Dodd, & D. Burnham (Eds.), *Hearing by eye II: Advances in the psychology of speechreading and auditory-visual speech* (pp. 85–108). Hove, UK: Psychology Press.
Stekelenburg, J. J., & Vroomen, J. (2007). Neural correlates of multisensory integration of ecologically valid audiovisual events. *Journal of Cognitive Neuroscience*, 19, 1964–1973.
Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of Acoustical Society of America*, 26, 212–215.
van Wassenhove, V., Grant, K. W., & Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proceedings of the National Academy of Sciences of the United States of America*, 102, 1181–1186.
Vroomen, J., & Stekelenburg, J. J. (2010). Visual anticipatory information modulates multisensory interactions of artificial audiovisual stimuli. *Journal of Cognitive Neuroscience*, 22, 1583–1596.
Wolpert, D. M., Ghahramani, Z., & Jordan, M. I. (1995). An internal model for sensorimotor integration. *Science*, 269(5232), 1880–1882.

frontiers in PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 13 May 2014
doi: 10.3389/fpsyg.2014.00420

The sound of your lips: electrophysiological cross-modal interactions during hand-to-face and face-to-face speech perception

Avril Treille*, Coriandre Vilain and Marc Sato

CNRS, Département Parole and Cognition, Gipsa-Lab, UMR 5216, Grenoble Université, Grenoble, France

Edited by:
Riikka Mottonen, University of Oxford, UK

Reviewed by:
Joana Acha, Basque Centre on Cognition, Brain and Language, Spain
Takayuki Ito, Haskins Laboratories, USA

****Correspondence:***
Avril Treille, CNRS, Département Parole and Cognition, Gipsa-Lab, UMR 5216, Grenoble Université, 1180 Avenue Centrale, BP 25, 38040 Grenoble Cedex 9, France
e-mail: avril.treille@gipsa-lab.inpg.fr

Recent magneto-encephalographic and electro-encephalographic studies provide evidence for cross-modal integration during audio-visual and audio-haptic speech perception, with speech gestures viewed or felt from manual tactile contact with the speaker's face. Given the temporal precedence of the haptic and visual signals on the acoustic signal in these studies, the observed modulation of N1/P2 auditory evoked responses during bimodal compared to unimodal speech perception suggest that relevant and predictive visual and haptic cues may facilitate auditory speech processing. To further investigate this hypothesis, auditory evoked potentials were here compared during auditory-only, audio-visual and audio-haptic speech perception in live dyadic interactions between a listener and a speaker. In line with previous studies, auditory evoked potentials were attenuated and speeded up during both audio-haptic and audio-visual compared to auditory speech perception. Importantly, the observed latency and amplitude reduction did not significantly depend on the degree of visual and haptic recognition of the speech targets. Altogether, these results further demonstrate cross-modal interactions between the auditory, visual and haptic speech signals. Although they do not contradict the hypothesis that visual and haptic sensory inputs convey predictive information with respect to the incoming auditory speech input, these results suggest that, at least in live conversational interactions, systematic conclusions on sensory predictability in bimodal speech integration have to be taken with caution, with the extraction of predictive cues likely depending on the variability of the speech stimuli.

Keywords: audio-visual speech perception, audio-haptic speech perception, multisensory interactions, EEG, auditory evoked potentials

INTRODUCTION

How information from different sensory modalities, such as sight, sound and touch, is combined to form a single coherent percept? As central to adaptive behavior, multisensory integration occurs in everyday life when natural events in the physical world have to be integrated from different sensory sources. It is an highly complex process known to depend on the temporal, spatial and causal relationships between the sensory signals, to take place at different timescales in several subcortical and cortical structures and to be mediated by both feedforward and backward neural projections. In addition to their coherence, the perceptual saliency and relevance of each sensory signal from the external environment, as well as their predictability and joint probability to occur, also act on the integration process and on the representational format at which the sensory modalities interface (for reviews, see Stein and Meredith, 1993; Stein, 2012).

Audio-visual speech perception is a special case of multisensory processing that interfaces with the linguistic system. Although one can extract phonetic features from the acoustic signal alone, adding visual speech information from the speaker's face is known to improve speech intelligibility in case of a degraded acoustic signal (Sumby and Pollack, 1954; Benoît et al., 1994; Schwartz et al., 2004), to facilitate the understanding of a semantically complex statement (Reisberg et al., 1987) or a foreign language (Navarra and Soto-Faraco, 2005), and to benefit hearing-impaired listeners (Grant et al., 1998). Conversely, in laboratory settings, adding incongruent visual speech information may interfere with auditory speech perception and even create an illusory percept (McGurk and MacDonald, 1976). Finally, as in other cases of bimodal integration, audio-visual speech integration depends on the perceptual saliency of both the auditory (Green, 1998) and visual (Campbell and Massaro, 1997) speech signals, as well as their spatial (Jones and Munhall, 1997) and temporal (van Wassenhove et al., 2003) relationships.

At the brain level, several magneto-encephalographic (MEG) and electro-encephalographic (EEG) studies demonstrate that visual speech input modulates auditory activity as early as 50–100 ms in the primary and secondary auditory cortices (Sams et al., 1991; Klucharev et al., 2003; Lebib et al., 2003; Besle et al., 2004; Hertrich et al., 2007; Winneke and Phillips, 2011). Importantly, it has been shown that both the latency and amplitude of auditory evoked responses (N1/P2, M100) are attenuated and speeded up during audio-visual compared to auditory-only speech perception (Klucharev et al., 2003; Besle et al., 2004; van Wassenhove et al.,

2005; Stekelenburg and Vroomen, 2007; Arnal et al., 2009; Pilling, 2010; Vroomen and Stekelenburg, 2010; Baart et al., 2014; Treille et al., 2014). Moreover, N1/P2 latency facilitation also appears to be directly function of the visemic information, with the higher visual recognition of the syllable, the longer latency facilitation (van Wassenhove et al., 2005; Arnal et al., 2009). Since the visual speech signal preceded the acoustic speech signal by 10s or 100s of milliseconds in these studies, the observed speeding-up and amplitude suppression of auditory evoked potentials might both reflect non-speech specific temporal (Stekelenburg and Vroomen, 2007; Vroomen and Stekelenburg, 2010) and phonetic (van Wassenhove et al., 2005; Arnal et al., 2009) visual predictions of the incoming auditory syllable (for recent discussions, see Arnal and Giraud, 2012; van Wassenhove, 2013; Baart et al., 2014).

Interestingly, speech can be perceived not only by the ear and by the eye but also by the hand, with orofacial speech gestures felt and monitored from manual tactile contact with the speaker's face. Past studies on the Tadoma method provide evidence for successful communication abilities in trained deaf-blind individuals through the haptic modality (Alcorn, 1932; Norton et al., 1977). A few behavioral studies also demonstrate the influence of tactile information on auditory speech perception in untrained individuals without sensory impairment, especially in case of noisy or ambiguous acoustic signals (Fowler and Dekle, 1991; Gick et al., 2008; Sato et al., 2010). In a recent EEG study (Treille et al., 2014), electrophysiological evidence of cross-modal interactions was found during both audio-visual and audio-haptic speech perception, through the course of live dyadic interactions between a listener and a speaker. In this study, participants were seated at arm's length from an experimenter and they were instructed to manually categorize /pa/ or /ta/ syllables presented auditorily, visually and/or haptically. In line with the above-mentioned EEG/MEG studies, N1 auditory evoked responses were attenuated and speeded up during live audio-visual speech perception. Crucially, haptic information was also found to speed up auditory speech processing as early as 100 ms. Given the temporal precedence of the dynamic configurations of the articulators on the auditory signal, as attested in a behavioral control experiment, the observed audio-haptic interactions in the listener's brain raise the possibility that the brain use predictive temporal and/or phonetic relevant tactile information for auditory processing, despite less natural processing to extract relevant speech information from the haptic modality. From this possibility, however, a clear limit of this study comes from the use of a simple two-alternative forced-choice identification task between /pa/ and /ta/ syllables and an insufficient number of trials for reliable EEG analyses per syllable.

To further explore whether perceivers might integrate tactile information in auditory speech perception as they do with visual information, the present study aimed at replicating the observed bimodal interactions during live face-to-face and hand-to-face speech perception (Treille et al., 2014). As observed in previous studies on audio-visual speech perception (van Wassenhove et al., 2005; Arnal et al., 2009), we also specifically tested whether modulation of N1/P2 auditory evoked potentials during both audio-visual and audio-haptic speech perception might depend on the degree to which the haptic and visual signals predict the incoming auditory speech target. To this aim, the experimental procedure was adapted from the Tadoma method and similar to that previously used by Treille et al. (2014), except the use of a three-alternative forced-choice identification task between /pa/, /ta/, and /ka/ syllables and a sufficient number of trials for reliable EEG analyses per syllable. A gradient of visual and haptic recognition between the three syllables was first attested in a behavioral experiment, which was a requirement to assess visual and haptic predictability on the incoming auditory signal in a subsequent EEG experiment. In line with previous EEG studies on audio-visual speech integration (van Wassenhove et al., 2005; Arnal et al., 2009), we hypothesized that the higher visual and haptic recognition of the syllable, the stronger latency facilitation in the audio-visual and audio-haptic modalities.

MATERIALS AND METHODS

PARTICIPANTS

Sixteen healthy adults, native French speakers, participated in the study (eight females; mean age ± SD, 29 ± 8 years). All participants were right-handed, had normal or corrected-to-normal vision and reported no history of speaking, hearing or motor disorders. Written informed consent was obtained for all participants and they were compensated for the time spent in the study. The study was approved by the Grenoble University Ethical Committee.

STIMULI

Based on a previous EEG study (van Wassenhove et al., 2005), /pa/, /ta/, and /ka/ syllables were selected in order to ensure precise acoustic onsets (thanks to the unvoiced stop bilabial /p/, alveolar /t/, and velar /k/ stop consonants) crucial for EEG analyses and, importantly, to ensure a gradient of visual and haptic recognition between these syllables (with notably the bilabial /p/ consonant known to be more visually salient than alveolar /t/ and velar /k/ consonants).

EXPERIMENTAL PROCEDURE

The study consisted on one behavioral experiment immediately followed by one EEG experiment. The behavioral experiment was performed in order to ensure a gradient of visual and haptic recognition of /pa/, /ta/, and /ka/ syllables. Importantly, since individual syllable onsets of the experimenter's productions were used as acoustical triggers for EEG analyses, the visual and haptic modalities of presentation were not included in the EEG experiment. In both experiments, Presentation software (Neurobehavioral Systems, Albany, CA, USA) was used to control the visual stimuli for the experimenter, the audio stimuli (beep) for the participant and to record key responses. In addition, all experimenter productions were recorded for off-line analyses in the EEG experiment.

Behavioral experiment

In a first behavioral experiment, participants were individually tested in a sound-proof room and were seated at arm's length from a female experimenter (see **Figure 1A**).

They were told that they would be presented with /pa/, /ta/, or /ka/ syllables either auditorily, visually, audio-visually, haptically, or audio-haptically over the hand-face contact. In the auditory modality (A), participants were instructed to keep their eyes closed and to listen to each syllable overtly produced by the

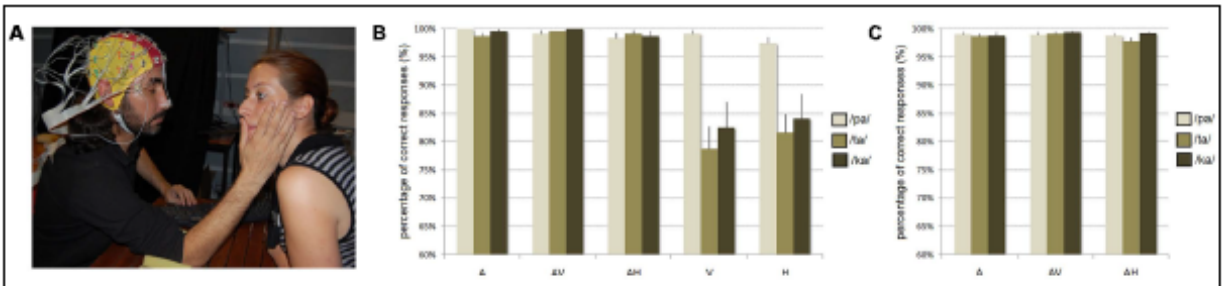

FIGURE 1 | (A) Experimental design used in the audio-haptic (AH) modality. In the haptic (H) and AH modalities, participants were asked to keep their eyes closed with their right hand placed on the experimenter's face and to categorize with their left hand each perceived syllable. In the auditory modality (A), participants were instructed to keep their eyes closed while, in the visual (V) and audio-visual modality (AV), they were asked to also look at the experimenter's face. The behavioral experiment included A, V, H, AV, AH modalities while the EEG experiment only included A, AV, and AH modalities. **(B,C)** Mean percentage of correct identification for /pa/, /ta/, and /ka/ syllables in each modality of presentation in the **(B)** behavioral and **(C)** EEG experiments. Error bars represent standard errors of the mean.

experimenter. In the audio-visual modality (AV), they were asked to also look at the experimenter's face. In the audio-haptic modality (AH), they were asked to keep their eyes closed with their right hand placed on the experimenter's face (the thumb placed lightly and vertically against the experimenter's lips and the other fingers placed horizontally along the jaw line in order to help distinguishing both lip and jaw movements). This experimental procedure was adapted from the Tadoma method and similar to that previously used by Treille et al. (2014). Finally, the visual-only (V) and haptic-only (H) modalities were similar to the AV and AH modalities except that the experimenter silently produced each syllable.

The experimenter faced the participant and a computer screen placed behind the participant. On each trial, the computer screen specified the syllable to be produced. To this aim, the syllable was printed three times on the computer screen at 1 Hz, with the last display serving as the visual go-signal to produce the syllable. The inter-trial interval was 3 s. The experimenter previously practiced and learned to articulate each syllable in synchrony with the visual go-signal, with an initial neutral closed-mouth position and maintaining an even intonation, tempo and vocal intensity.

A three-alternative forced-choice identification task was used, with participants instructed to categorize each perceived syllable by pressing on one of three keys corresponding to /pa/, /ta/, or /ka/ on a computer keyboard with their left hand. A brief single audio beep was delivered 600 ms after the visual go-signal (expecting to occur in synchrony with the experimenter production) with the participants told to produce their responses only after this audio go-signal. This procedure was done in order to dissociate sensory/perceptual responses from motor responses on EEG data in the next experiment. As a consequence, no reaction-times were acquired and only response rate were considered in further analyses.

Every syllable (/pa/, /ta/, or /ka/) was presented 15 times in each modality (A, V, H, AV, AH) in a single randomized sequence for a total of 225 trials. The response key designation were counterbalanced across participants. Before the experiment, participants performed few practice trials in all modalities. They received no instructions concerning how to interpret visual and haptic information but they were asked to pay attention to both modalities during bimodal presentation.

EEG experiment

Because of no possible reliable acoustical triggers in the visual-only and haptic-only modalities, the EEG experiment only included three individual experimental sessions related to A, AV, and AH modalities of presentation. Except this difference and the number of trials, the experimental procedure was identical to that used in the behavioral experiment. In each session, every syllable (/pa/, /ta/, or /ka/) was presented 80 times in a randomized sequence for a total of 240 trials. The order of the modality of presentation and the response key designation were fully counterbalanced across participants. Because the experimental procedure was quite taxing, each experimental session was split into two blocks of around 6 min each, allowing short breaks for both the experimenter and the participants.

EEG ACQUISITION

In the EEG experiment, EEG data were continuously recorded from 64 scalp electrodes (Electro-Cap International, INC., according to the international 10–20 system) using the Biosemi ActiveTwo AD-box EEG system operating at a sampling rate of 256 Hz. Two additional electrodes served as reference (common mode sense [CMS] active electrode) and ground (driven right leg [DRL] passive electrode). One other external reference electrode was at the top of the nose. The electro-oculogram measuring horizontal (HEOG) and vertical (VEOG) eye movements were recorded using electrodes at the outer canthus of each eye as well as above and below the right eye. Before the experiment, the impedance of all electrodes was adjusted to get low offset voltages and stable DC.

DATA ANALYSES

Behavioral analyses

In both the behavioral and EEG experiments, the proportion of correct responses was individually determined for each participant, each syllable and each modality. Two-way repeated-measure ANOVAs were performed on these data with the modality

(A, V, H, AV, AH in the behavioral experiment; A, AV, AH in the EEG experiment) and the syllable (/pa, /ta/, /ka/) as within-subjects variables.

Acoustical analyses

In the EEG experiment, acoustical analyses were performed on the experimenter's recorded syllables in order to determine the individual syllable onsets serving as acoustical triggers for the EEG analyses. All acoustical analyses were performed using Praat software (Boersma and Weenink, 2013). First, an automatic procedure based on an intensity and duration algorithm detection roughly identified each syllable's onset in the A, AV, and AH modalities (11520 utterances). For all syllables, these onsets were further manually and precisely determined, based on waveform and spectrogram information related to the acoustic characteristics of voiced stop consonants. Omissions and wrong productions were identified and removed from the analyses (less than 1%).

EEG analyses

EEG data were processed using the EEGLAB toolbox (Delorme and Makeig, 2004) running on Matlab (Mathworks, Natick, MA, USA). Since N1/P2 auditory evoked potentials have maximal response over central sites on the scalp (Scherg and Von Cramon, 1986; Näätänen and Picton, 1987), EEG data preprocessing and analyses were conducted on three central electrodes (C3, Cz, C4). These electrodes, covering left, middle, and right central sites, were also selected based on previous EEG studies on audio-visual speech perception (e.g., Klucharev et al., 2003; Besle et al., 2004; Pilling, 2010; Treille et al., 2014). EEG data were first re-referenced offline to the nose recording and band-pass filtered using a two-way least-squares FIR filtering (1–20 Hz). Data were then segmented into epochs of 1000 ms (from −500 ms to +500 ms to the acoustic syllable onset, individually determined from the acoustical analyses), with the prestimulus baseline defined from −500 ms to −400 ms. Epochs with an amplitude change exceeding ±60 μV at any channel (including HEOG and VEOG channels) were rejected (on average, less than 10%).

For each participant and each modality, the peak latency of auditory N1 and P2 evoked responses were first determined on the EEG waveform averaged over all electrodes and syllables. For each syllable, two temporal windows were then defined on these peaks ±30 ms in order to individually calculate N1 and P2 amplitude and latency on the related average waveform of C3, Cz, C4 electrodes. Two-way repeated-measure ANOVAs were then performed on N1 and P2 amplitude and latency with the modality (A, AV, AH) and the syllable (/pa/, /ka/, /ta/) as within-subjects variables.

In order to confirm previous EEG/MEG studies demonstrating that P2 and M100 latency reduction in the audio-visual modality vary as a function of the visual recognition of the presented syllable (van Wassenhove et al., 2005; Arnal et al., 2009), additional Pearson's correlation analyses were carried out. These correlation analyses were performed between the individual visual and haptic recognition scores of the three syllables in the behavioral experiment and the related latency facilitation and reduction amplitude observed in the AV and AH modalities in the EEG experiment (leading to 3 × 16 correlation points per measure and per modality). In addition to raw data, these analyses were also performed on individual *Z*-score normalized data, in order to take account of individual differences.

RESULTS

For all the following analyses, the significance level was set at $p = 0.05$ and Greenhouse–Geisser corrected (for violation of the sphericity assumption) when appropriate. When required, *post hoc* analyses were conducted with Newman–Keuls tests.

BEHAVIORAL ANALYSES

Behavioral experiment (see Figure 1B)

Overall, the mean proportion of correct responses was of 94%. The main effect of modality of presentation was significant [$F(4,60) = 33.67$, $p < 0.001$], with more correct responses in A, AV, and AH modalities than in V and H modalities (as shown by *post hoc* analyses, all p's < 0.001). Significant differences were also observed between syllables [$F(2,30) = 15.59$, $p < 0.001$], with more correct responses for /pa/ than for /ta/ and /ka/ syllables (as shown by *post hoc* analyses, all p's < 0.001). Finally, the interaction between the modality and the syllable was also reliable [$F(8,120) = 7.39$, $p < 0.001$]. While no significant differences were observed between syllables in A, AV, and AH modalities (with almost perfect identification for all syllables), more correct responses were observed for /pa/ than for /ta/ and /ka/ syllables in both V and H modalities (as shown by *post hoc* analyses, all p's < 0.001). Altogether, these results thus demonstrate a near perfect identification of /pa/ in all modalities, but a lower accuracy for /ta/ and /ka/ syllables in V and H modalities.

EEG experiment (see Figure 1C)

In the EEG experiment, the mean proportion of correct responses was of 99%. No significant effect of the modality [$F(2,30) = 1.72$], syllable [$F(2,30) = 1.34$] or interaction [$F(4,60) = 0.90$] was observed, with a near perfect identification of all syllables in A, AV, and AH modalities.

EEG ANALYSES

N1 amplitude (see Figures 2 and *3A-left)*

The main effect of modality was significant [$F(2,30) = 9.19$, $p < 0.001$], with a reduced negative N1 amplitude observed in the AV and AH modalities as compared to the A modality (as shown by *post hoc* analyses, $p < 0.001$ and $p < 0.02$, respectively; on average, A: −5.3 μV, AV: −3.1 μV, AH: −4.1 μV). The interaction between the modality and the syllable was also found to be significant [$F(4,60) = 7.23$, $p < 0.001$]. While for /pa/ a significant amplitude reduction was observed in both AV and AH modalities as compared to the A modality, an amplitude reduction was only observed in the AV modality for /ta/ and /ka/ syllables (as shown by *post hoc* analyses, all p's < 0.001, see **Figure 3A**-left). In sum, these results demonstrate a visually induced amplitude suppression for all syllables and, importantly, an haptically induced amplitude suppression but only for /pa/ syllable.

P2 amplitude (see Figures 2 and *3B-left)*

No significant effect of the modality [$F(2,30) = 1.91$], the syllable [$F(2,30) = 1.09$] and their interaction [$F(4,60) = 1.58$] was observed.

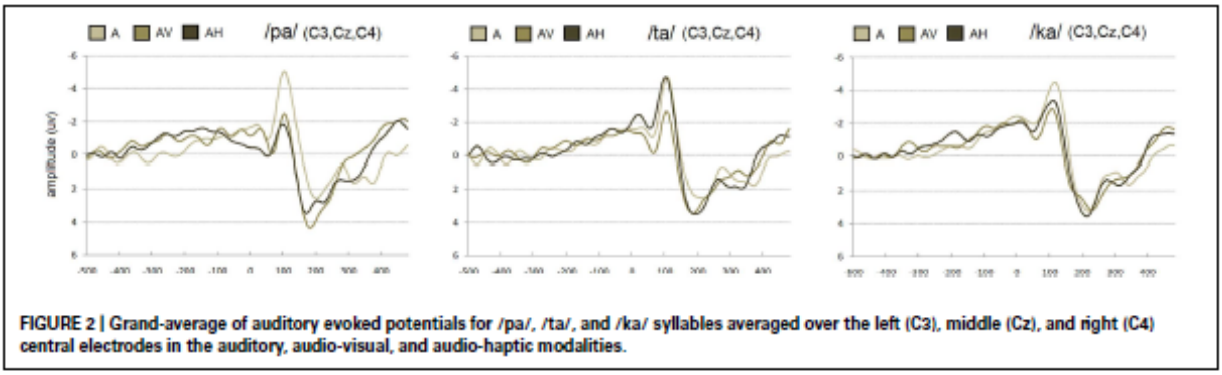

FIGURE 2 | Grand-average of auditory evoked potentials for /pa/, /ta/, and /ka/ syllables averaged over the left (C3), middle (Cz), and right (C4) central electrodes in the auditory, audio-visual, and audio-haptic modalities.

N1 latency (see Figures 2 and 3C-left)

No significant effect of the modality [$F(2,30) = 0.36$], the syllable [$F(2,30) = 3.13$] and their interaction [$F(4,60) = 1.78$] was observed.

P2 latency (see Figures 2 and 3D-left)

The main effect of syllable [$F(2,30) = 4.54$, $p < 0.02$] was reliable, with shorter P2 latencies observed for /pa/ and /ta/ syllables as compared to /ka/ (as shown by *post hoc* analyses, all p's < 0.03; on average, /pa/: 210 ms, /ta/: 211 ms, /ka/: 217 ms). Crucially, the main effect of modality was significant [$F(2,30) = 4.05$, $p < 0.03$], with shorter latencies in AV and AH as compared to the A modality (as shown by *post hoc* analyses, all p's < 0.05; on average, A: 223 ms, AV: 208 ms, AH: 207 ms). In sum, these results thus indicate faster processing of the P2 auditory evoked potential for /pa/ and /ka/ syllables. In addition, a latency facilitation was observed in both AV and AH modalities, irrespective of the presented syllables.

Correlation between perceptual recognition scores (see Figure 3-right)

For raw data, whatever the modality, no significant correlation was however observed for both N1 amplitude (AV: $r = 0.09$, $p = 0.54$; AH: $r = 0.06$, $p = 0.70$), P2 amplitude (AV: $r = 0.25$, $p = 0.09$; AH: $r = -0.09$, $p = 0.53$), N1 latency (AV: $r = -0.06$, $p = 0.71$; AH: $r = 0.11$, $p = 0.45$), and P2 latency (AV: $r = 0.07$, $p = 0.66$; AH: $r = -0.01$, $p = 0.92$). Results on additional correlation analyses on normalized data also failed to demonstrate any significant correlation for both N1 and P2 amplitude (N1-AV: $r = 0.01$, $p = 0.98$; N1-AH: $r = 0.18$, $p = 0.87$; P2-AV: $r = 0.21$, $p = 0.15$; P2-AH: $r = 0.02$, $p = 0.91$) and latency (N1-AV: $r = 0.01$, $p = 0.92$; N1-AH: $r = 0.12$, $p = 0.65$; P2-AV: $r = 0.06$, $p = 0.68$; P2-AH: $r = -0.02$, $p = 0.87$).

DISCUSSION

Two main results emerge from the present study. First, in line with our previous results (Treille et al., 2014), a modulation of N1/P2 auditory evoked potentials was observed during live audio-visual and audio-haptic speech perception compared to auditory speech perception. However, contrary to two previous studies of audio-visual speech perception (van Wassenhove et al., 2005; Arnal et al., 2009), no significant correlation was observed between the latency facilitation observed in the bimodal conditions and the degree of visual and haptic recognition of the presented syllables.

Before we discuss these results, it is first important to consider one potential limitation of the present study. Classically, testing cross-modal interactions requires to determine that the observed response in the bimodal condition differ to the sum of those observed in the unimodal conditions (e.g., AV $\neq$ A + V). However, visual-only and haptic-only modalities were not here tested, due to the technical difficulty to get temporal accurate and reliable triggers for EEG analyses. Notably, because of their temporal limitation and variability, visual and/or surface electromyographic recordings of the experimenter's lip, jaw or tongue movements would not allowed to determine reliable triggers (especially in the case of lip stretching for /ta/ and /ka/ syllables). From the possibility that the observed bimodal neural responses simply come from a superposition of the unimodal signals, it should however be noted that auditory evoked potentials are rarely observed in the visual-only modality in central electrodes (Besle et al., 2004; van Wassenhove et al., 2005; Pilling, 2010). Furthermore, in our previous study and using the same experimental design, we obtained behavioral evidence for a strong temporal precedence of the haptic and visual signals on the acoustic signal (Treille et al., 2014). In our view, it is therefore unlikely that visual and haptic event-related potentials might arise at the same time-latency and at the same central electrodes that N1 and P2 auditory evoked potentials. For these reasons, we here compared neural responses in each bimodal condition to the related unimodal condition (i.e., AV $\neq$ A and AH $\neq$ H), a testing procedure that has previously demonstrated latency facilitation and amplitude reduction of auditory evoked potentials in audio-visual compared to auditory-only speech perception (van Wassenhove et al., 2005; Pilling, 2010).

In spite of this limitation, the observed modulation of N1/P2 auditory evoked potentials in the audio-visual condition strongly suggests cross-modal speech interactions. It is first worthwhile noting that, for each participant, the three syllables were randomly presented in each session in order to minimize repetition effects, and the order of the modality of presentation was fully counterbalanced across participants so that possible overlapping modality effects are unlikely. In addition, auditory-evoked responses were compared between modalities, with the same number of trials and therefore similar possible habituation effects. Although our results

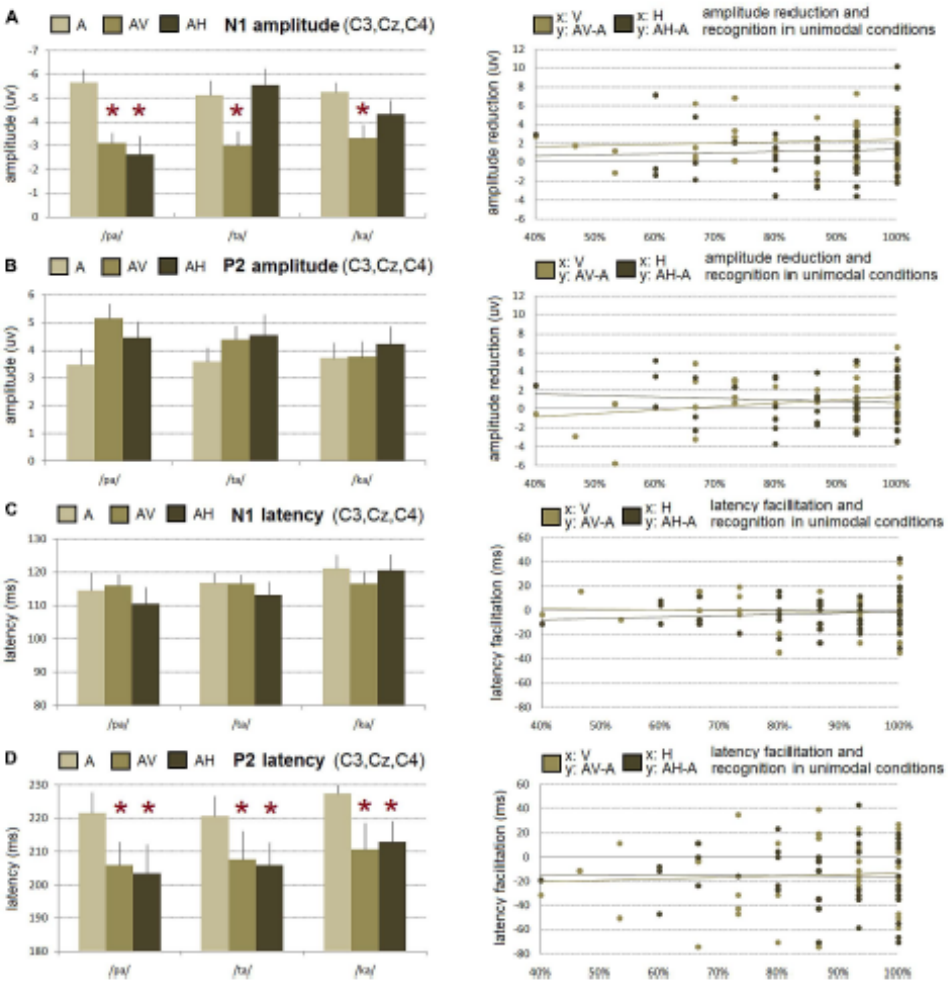

FIGURE 3 | Left. Mean N1 **(A)** and P2 **(B)** amplitude and mean N1 **(C)** and P2 **(D)** latency for /pa/, /ta/, and /ka/ syllables averaged over left (C3), middle (Cz), and right (C4) central electrodes in the auditory (A), audio-visual (AV), and audio-haptic (AH) modalities. Error bars represent standard errors of the mean. * indicates a significant effect. **Right.** Correlation on raw data between the recognition scores observed in the visual-only and haptic-only modalities in the behavioral experiment (*x*-axis) and the reduction amplitude and latency facilitation observed in the audio-visual and audio-haptic modalities in the EEG experiment (*y*-axis). No correlation was significant.

appear globally consistent with previous EEG studies, some differences have however to be mentioned. First, while the observed amplitude reduction was here confined to the N1 auditory evoked potential, as in our previous study (Treille et al., 2014; see also Besle et al., 2004), such a visually induced suppression has been previously observed for both N1 and P2 auditory components (Klucharev et al., 2003; van Wassenhove et al., 2005; Stekelenburg and Vroomen, 2007; Pilling, 2010; Baart et al., 2014) or only for the P2 component (Baart et al., 2014). Second, the observed P2 latency facilitation also contrasts with previous studies showing earlier latencies during audio-visual speech perception for both N1 and P2 peaks (van Wassenhove et al., 2005; see also Pilling, 2010, for a small but not consistent effect) or only for N1 peak (Stekelenburg and Vroomen, 2007; Baart et al., 2014; Treille et al., 2014). From these differences, it is hypothesized that N1 and P2 components as well as latency facilitation and amplitude reduction effects might reflect different aspects and/or stages of audio-visual speech integration. For instance, van Wassenhove et al. (2005) observed a visually induced suppression of both N1 and P2 components independently of the visual saliency of the speech stimuli, but a latency reduction of N1 and P2 peaks depending on the degree of their visual predictability. From their results, they argue for two distinct integration stages: (1) a global bimodal perceptual stage, reflected in the amplitude reduction, independent of the featural content of the visual stimulus and possibly reflecting phase-coupling of auditory and visual cortices, and (2) a featural phonetic stage, reflected in the latency facilitation and stronger for P2, in which articulator-specific and predictive visual information

are taking into account in auditory phonetic processing (for further discussion, see van Wassenhove, 2013). In parallel, Stekelenburg and Vroomen (2007), Vroomen and Stekelenburg (2010), and Baart et al. (2014) also argue for a bimodal, non-speech specific stage in audio-visual speech integration but here thought to be reflected in the N1 latency facilitation and amplitude reduction. Congruent with this hypothesis, they observed an amplitude and a latency reduction of auditory-evoked N1 responses during audio-visual perception for both speech and non-speech actions, like clapping hands (Stekelenburg and Vroomen, 2007), as well as for artificial audio-visual stimul, like two moving disks predicting a pure tone when colliding with a fixed rectangle (Vroomen and Stekelenburg, 2010). In addition, they also provided evidence for a P2 amplitude reduction specifically dependent on the phonetic predictability of the visual speech input (Baart et al., 2014; see also Vroomen and Stekelenburg, 2010). Taken together, although the observed differences across the present and previous studies on N1 and/or P2 latency facilitation and/or amplitude reduction are still a matter of debate (van Wassenhove et al., 2005; Baart et al., 2014), they might both reflect multistage processes in audio-visual speech integration and also derive from specific experimental settings used in these studies.

From that latter possibility, one interesting finding is that the observed latency and amplitude reduction in the EEG experiment, notably for the P2 component, did not significantly depend on the degree of visual recognition of the speech targets in the behavioral experiment. This contrasts with two previous studies reporting latency shifts of auditory evoked responses directly function of the visemic information (van Wassenhove et al., 2005; Arnal et al., 2009). For instance, van Wassenhove et al. (2005) demonstrated a visually induced facilitation of the P2 auditory evoked potential which systematically varied according to the visual-only recognition of the presented syllable (i.e., the more visually salient was the syllable, the more stronger the latency facilitation). While they observed a P2 latency facilitation around 25 ms, 16 ms, and 8 ms for /pa/, /ta/, and /ka/ syllables, respectively, we here observed latency facilitations around 17 ms, 13 ms, and 15 ms for the same syllables. However, correlation scores likely depend on overall differences in recognition scores between syllables which were stronger in previous studies (van Wassenhove et al., 2005; Arnal et al., 2009). Furthermore, one important difference between our experimental setting and those used in these two studies is that audio-visual interactions were here tested during live face-to-face interactions between a speaker and a listener, with a unique occurrence of the presented syllable in each trial. This natural stimulus variability contrasts with the limited number of tokens used to represent each syllable in the previous studies which were repeatedly presented to the participants (i.e., van Wassenhove et al. (2005): one speaker, three syllables, one token per syllable and 100 trials per syllable and per modality; Arnal et al. (2009): one speaker, five syllables, one token per syllable and 54 trials per syllable and per modality). Similarly, another possible experimental factor impacting bimodal speech integration comes from the number of syllable type. From that view, it is worthwhile noting that we did observe a latency facilitation during live face-to-face speech perception in our previous study, using a similar experimental design, but only for the N1 component (Treille et al., 2014). In this study, however, a simple two-alternative forced-choice identification task between /pa/ and /ta/ syllables was used. It is therefore possible that specific phonetic contents of these two syllables were less perceptually dominant in this previous study, with a more global yes-no strategy done in relation to the more salient bilabial movements for /pa/ as compared to /ta/ (for experimental designs only using two distinct speech stimuli, see also Stekelenburg and Vroomen, 2007; Pilling, 2010; Vroomen and Stekelenburg, 2010; Baart et al., 2014). Overall, given the significant P2 latency facilitation, our results do not contradict the hypothesis that visual inputs convey predictive information with respect to the incoming auditory speech input (for a discussion on the sensory predictability of audio-visual speech stimuli, see Chandrasekaran et al., 2009; Schwartz and Savariaux, 2013) nor the fact that visual predictability of the speech stimulus might be reflected in auditory evoked responses. We simply argue that visual predictions on the incoming acoustic signal in audio-visual speech perception might likely be constrained not only by the featural content of the visual stimuli but also by the experimental context and by short-term memory traces and knowledge the listener previously acquired on these stimuli.

As in the audio-visual condition, the observed modulation of N1/P2 auditory evoked potentials during audio-haptic speech perception also clearly suggests cross-modal speech interactions between the auditory and the haptic signals. In this bimodal condition, we also observed a latency facilitation on the P2 auditory evoked potential that did not vary according to the degree of haptic recognition of the speech targets. In addition to this latency facilitation, an N1 amplitude reduction was also observed but only for /pa/ syllable. As previously noted, this latter result fits well with a stronger haptic saliency of the bilabial rounding movements involved in /pa/ syllable (see Treille et al., 2014, for behavioral evidence) and with previous studies on audio-visual integration demonstrating that N1 suppression is strongly dependent on whether the visual signal reliably predicts the onset of the auditory event (Stekelenburg and Vroomen, 2007; Vroomen and Stekelenburg, 2010). As discussed previously, the fact that P2 latency reduction was nevertheless observed for all syllables indirectly argue for distinct integration processes in the cortical speech processing hierarchy (van Wassenhove et al., 2005; Stekelenburg and Vroomen, 2007; Vroomen and Stekelenburg, 2010; Baart et al., 2014).

Taken together, our results provide new evidence for audio-visual and audio-haptic speech interactions in live dyadic interactions (Treille et al., 2014). The fact that the modulation of N1/P2 auditory evoked potentials were quite similar in these bimodal conditions, despite the less natural haptic modality, further emphasizes the multimodal nature of speech perception. As previously mentioned, apart from speech, multisensory integration from sight, sound and haptic modalities naturally occurs in everyday life. Although bimodal speech perception is a special case of multisensory processing that interfaces with the linguistic system, similar integration processes might have been used to extract temporal and/or phonetic relevant information from the visual and haptic speech signals that, together with the listener's knowledge of speech production (for a review, see Schwartz et al., 2012), might have constrained the incoming auditory processing.

REFERENCES

Alcorn, S. (1932). The Tadoma method. *Volta Rev.* 34, 195–198.

Arnal, L. H., and Giraud, A. L. (2012). Cortical oscillations and sensory predictions. *Trends Cogn. Sci.* 16, 390–398. doi: 10.1016/j.tics.2012.05.003

Arnal, L. H., Morillon, B., Kell, C. A., and Giraud, A. L. (2009). Dual neural routing of visual facilitation in speech processing. *J. Neurosci.* 29, 13445–13453. doi: 10.1523/JNEUROSCI.3194-09.2009

Baart, M., Stekelenburg, J. J., and Vroomen, J. (2014). Electrophysiological evidence for speech-specific audiovisual integration. *Neuropsychologia* 65, 115–211. doi: 10.1016/j.neuropsychologia.2013.11.011

Benoît, C., Mohamadi, T., and Kandel, S. D. (1994). Effects on phonetic context on audio-visual intelligibility of French. *J. Speech Hear. Res.* 37, 1195–1203.

Besle, J., Fort, A., Delpuech, C., and Giard, M. H. (2004). Bimodal speech: early suppressive visual effects in human auditory cortex. *Eur. J. Neurosci.* 20, 2225–2234. doi: 10.1111/j.1460-9568.2004.03670.x

Boersma, P., and Weenink, D. (2013). *Praat: Doing Phonetics by Computer. Computer Program*, Version 5.3.42. Available at: http://www.praat.org/ [accessed March 2, 2013].

Campbell, C. S., and Massaro, D. W. (1997). Perception of visible speech: influence of spatial quantization. *Perception* 26, 627–644. doi: 10.1068/p260627

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., and Ghazanfar, A. A. (2009). The natural statistics of audiovisual speech. *PLoS Comput. Biol.* 5:e1000436. doi: 10.1371/journal.pcbi.1000436

Delorme, A., and Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *J. Neurosci. Methods* 134, 9–21. doi: 10.1016/j.jneumeth.2003.10.009

Fowler, C., and Dekle, D. (1991). Listening with eye and hand: crossmodal contributions to speech perception. *J. Exp. Psychol. Hum. Percept. Perform.* 17, 816–828. doi: 10.1037/0096-1523.17.3.816

Gick, B., Jóhannsdóttir, K. M., Gibraiel, D., and Mühlbauer, M. (2008). Tactile enhancement of auditory and visual speech perception in untrained perceivers. *J. Acoust. Soc. Am.* 123, 72–76. doi: 10.1121/1.2884349

Grant, K., Walden, B. E., and Seitz, P. F. (1998). Auditory-visual speech recognition by hearing-impaired subjects: consonant recognition, sentence recognition, and auditory-visual integration. *J. Acoust. Soc. Am.* 103, 2677–2690. doi: 10.1121/1.422788

Green, K. P. (1998). "The use of auditory and visual information during phonetic processing: implications for theories of speech perception," in *Hearing by Eye, II. Perspectives and Directions in Research on Audiovisual Aspects of Language Processing*, eds R. Campbell, B. Dodd, and D. Burnham (Hove: Psychology Press), 3–25.

Hertrich, I., Mathiak, K., Lutzenberger, W., Menning, H., and Ackermann, H. (2007). Sequential audiovisual interactions during speech perception: a whole-head MEG study. *Neuropsychologia* 45, 1342–1354. doi: 10.1016/j.neuropsychologia.2006.09.019

Jones, J. A., and Munhall, K. G. (1997). The effects of separating auditory and visual sources on audiovisual integration of speech. *Can. Acoust.* 25, 13–19.

Klucharev, V., Möttönen, R., and Sams, M. (2003). Electrophysiological indicators of phonetic and non-phonetic multisensory interactions during audiovisual speech perception. *Brain Res. Cogn. Brain Res.* 18, 65–75. doi: 10.1016/j.cogbrainres.2003.09.004

Lebib, R., Papo, D., de Bode, S., and Baudonnière, P. M. (2003). Evidence of a visual-to-auditory cross-modal sensory gating phenomenon as reflected by the human P50 event-related brain potential modulation. *Neurosci. Lett.* 341, 185–188. doi: 10.1016/S0304-3940(03)00131-9

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

Näätänen, R., and Picton, T. W. (1987). The N1 wave of the human electric and magnetic response to sound: a review and an analysis of the component structure. *Psychophysiology* 24, 375–425. doi: 10.1111/j.1469-8986.1987.tb00311.x

Navarra, J., and Soto-Faraco, S. (2005). Hearing lips in a second language: visual articulatory information enables the perception of second language sounds. *Psychol. Res.* 71, 4–12. doi: 10.1007/s00426-005-0031-5

Norton, S. J., Schultz, M. C., Reed, C. M., Braida, L. D., Durlach, N. I., Rabinowitz, W. M., et al. (1977). Analytic study of the Tadoma method: background and preliminary results. *J. Speech Hear. Res.* 20, 574–595.

Pilling, M. (2010). Auditory event-related potentials (ERPs) in audiovisual speech perception. *J. Speech Lang. Hear. Res.* 52, 1073–1081. doi: 10.1044/1092-4388(2009/07-0276)

Reisberg, D., McLean, J., and Goldfield, A. (1987). "Easy to hear but hard tounderstand: a lipreading advantage with intact auditory stimuli," in *Hearing by Eye: The Psychology of Lipreading*, eds R. Campbell and B. Dodd (London: Lawrence Erlbaum Associates), 97–113.

Sams, M., Aulanko, R., Hämäläinen, M., Hari, R., Lounasmaa, O. V., Lu, S. T., et al. (1991). Seeing speech: visual information from lip movements modifies activity in the human auditory cortex. *Neurosci. Lett.* 127, 141–145. doi: 10.1016/0304-3940(91)90914-F

Sato, M., Cavé, C., Ménard, L., and Brasseur, L. (2010). Auditory-tactile speech perception in congenitally blind and sighted adults. *Neuropsychologia* 48, 3683–3686. doi: 10.1016/j.neuropsychologia.2010.08.017

Scherg, M., and Von Cramon, D. (1986). Evoked dipole source potentials of the human auditory cortex. *Electroencephalogr. Clin. Neurol.* 65, 344–360. doi: 10.1016/0168-5597(86)90014-6

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

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

Schwartz, J. L., and Savariaux, C. (2013). "Data and simulations about audiovisual asynchrony and predictability in speech perception," in *Proceedings of the 12th International Conference on Auditory-Visual Speech Processing*, Annecy, France.

Stein, B. E. (2012). *The New Handbook of Multisensory Processing*. Cambridge: MIT Press.

Stein, B. E., and Meredith, M. A. (1993). *The New Handbook of Multisensory Processing*. Cambridge, MA: MIT Press.

Stekelenburg, J. J., and Vroomen, J. (2007). Neural correlates of multisensory integration of ecologically valid audiovisual events. *J. Cogn. Neurosci.* 19, 1964–1973. doi: 10.1162/jocn.2007.19.12.1964

Sumby, W. H., and Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *J. Acoust. Soc. Am.* 26, 212–215. doi: 10.1121/1.1907309

Treille, A., Cordeboeuf, C., Vilain, C., and Sato, M. (2014). Haptic and visual information speed up the neural processing of auditory speech in live dyadic interactions. *Neuropsychologia* 57, 71–77. doi: 10.1016/j.neuropsychologia.2014.02.004

van Wassenhove, V. (2013). Speech through ears and eyes: interfacing the senses with the supramodal brain. *Front. Psychol.* 4:388. doi: 10.3389/fpsyg.2013.00388

van Wassenhove, V., Grant, K. W., and Poeppel, D. (2003). Temporal window of integration in auditory-visual speech perception. *Neuropsychologia* 45, 598–607. doi: 10.1016/j.neuropsychologia.2006.01.001

van Wassenhove, V., Grant, K. W., and Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proc. Natl. Acad. Sci. U.S.A.* 102, 1181–1186. doi: 10.1073/pnas.0408949102

Vroomen, J., and Stekelenburg, J. J. (2010). Visual anticipatory information modulates multisensory interactions of artificial audiovisual stimuli. *J. Cogn. Neurosci.* 22, 1583–1596. doi: 10.1162/jocn.2009.21308

Winneke, A. H., and Phillips, N. A. (2011). Does audiovisual speech offer a fountain of youth for old ears? An event-related brain potential study of age differences in audiovisual speech perception. *Psychol. Aging* 26, 427–438. doi: 10.1037/a0021683

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 02 March 2014; accepted: 21 April 2014; published online: 13 May 2014.
Citation: Treille A, Vilain C and Sato M (2014) The sound of your lips: electrophysiological cross-modal interactions during hand-to-face and face-to-face speech perception. Front. Psychol. 5:420. doi: 10.3389/fpsyg.2014.00420
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology.

PARTIE EXPÉRIMENTALE - B

ETUDE IRMF SUR LA PERCEPTION AUDIO-VISUELLE LINGUALE

Dans ce chapitre, nous présenterons, sous la forme d'un article scientifique publié dans le *Journal of Cognitive Neuroscience* en 2017, une étude en imagerie par résonance magnétique fonctionnelle (IRMf) sur la perception audio-visuelle des mouvements des lèvres ou de la langue (relative à l'état de l'art proposé dans la section D-2 de la partie théorique de cette thèse). Cette expérience a été réalisée au sein du laboratoire Gipsa-lab de Grenoble et de la plateforme IRM (IRMaGe) de Grenoble, sous la direction de Marc Sato et Coriandre Vilain et avec la collaboration de Thomas Hueber (GIPSA-lab) pour l'enregistrement des images ultrasons utilisées pour les stimuli, et de Laurent Lamalle (UMS IRMage, Université Grenoble Alpes) pour les passations IRMf,.

Treille, A., Vilain, C., Hueber, T., Lamalle, L. & Sato, M. (2017). Inside Speech: Multisensory and Modality-specific Processing of Tongue and Lips Speechs Actions. Journal of Cognitive Neuroscience, 29(3):448-466.

Nous avons vu que la perception d'une action dépendrait non seulement de régions sensorielles, mais aussi du système moteur de l'observateur. Cependant, l'influence de l'expérience auditive ou visuelle d'une action sur l'activité sensorimotrice reste très peu étudiée de nos jours. L'étude IRMf présentée ci-après avait pour but de déterminer dans quelle mesure les représentations sensorielles et motrices interagissent durant la perception d'actions de parole relatives aux mouvements des lèvres ou de la langue. Nous avons choisi ces deux articulateurs du fait d'une grande expérience visuelle des mouvements des lèvres d'autrui (notamment en lecture labiale) et, en revanche, d'une faible connaissance visuelle des mouvements de la langue d'autrui.

Lors de cette étude, il était présenté aux participants des actions de parole auditives, visuelles et audio-visuelles. Le signal visuel était relatif soit à la vue de face de mouvements des lèvres préalablement enregistrés par une caméra, soit à une coupe sagittale de mouvements de la langue (la langue était donc vue de profil) enregistrée par un système ultrason. Les participants avaient pour tâche de percevoir passivement des syllabes (/pa/, /ta/ et /ka/).

Nos résultats montrent que la perception visuo-labiale et visuo-linguale entraîne l'activation d'un large réseau sensorimoteur commun. Cependant, une plus grande activation des régions motrices et somatosensorielles a été observée lors de la perception visuelle des mouvements de la langue, tandis que la vue des mouvements des lèvres activait plus fortement les régions auditives et visuelles. De plus, nous avons également observé une corrélation entre l'activation du cortex prémoteur et les scores de reconnaissance visuelle des stimuli linguaux et une corrélation de l'activité des régions visuelles avec les scores de reconnaissance visuelle des stimuli labiaux.

Pris ensemble, ces résultats suggèrent que les traitements unimodaux et multimodaux de la perception des gestes des lèvres ou de la langue s'appuient sur un large réseau neuronal

sensorimoteur commun. Ils suggèrent également que les traitements visuels d'actions audibles mais non visibles (comme celles de la langue) nécessitent une simulation motrice et visuelle des actions pour faciliter leur reconnaissance et/ou l'apprentissage de l'association entre les signaux auditifs et visuels.

Inside Speech: Multisensory and Modality-specific Processing of Tongue and Lip Speech Actions

Avril Treille[1], Coriandre Vilain[1], Thomas Hueber[1], Laurent Lamalle[2,3], and Marc Sato[4]

Abstract

■ Action recognition has been found to rely not only on sensory brain areas but also partly on the observer's motor system. However, whether distinct auditory and visual experiences of an action modulate sensorimotor activity remains largely unknown. In the present sparse sampling fMRI study, we determined to which extent sensory and motor representations interact during the perception of tongue and lip speech actions. Tongue and lip speech actions were selected because tongue movements of our interlocutor are accessible via their impact on speech acoustics but not visible because of its position inside the vocal tract, whereas lip movements are both "audible" and visible. Participants were presented with auditory, visual, and audiovisual speech actions, with the visual inputs related to either a sagittal view of the tongue movements or a facial view of the lip movements of a speaker, previously recorded by an ultrasound imaging system and a video camera. Although the neural networks involved in visual visuo-lingual and visuo-facial perception largely overlapped, stronger motor and somatosensory activations were observed during visuo-lingual perception. In contrast, stronger activity was found in auditory and visual cortices during visuo-facial perception. Complementing these findings, activity in the left premotor cortex and in visual brain areas was found to correlate with visual recognition scores observed for visuo-lingual and visuo-facial speech stimuli, respectively, whereas visual activity correlated with RTs for both stimuli. These results suggest that unimodal and multimodal processing of lip and tongue speech actions rely on common sensorimotor brain areas. They also suggest that visual processing of audible but not visible movements induces motor and visual mental simulation of the perceived actions to facilitate recognition and/or to learn the association between auditory and visual signals. ■

INTRODUCTION

Through life experiences, we learn about which sensory features of actions are most behaviorally relevant for successful categorization and recognition. However, one intriguing question is to know what happens when an action is not accessible to one sensor in the daily experience—typically, accessible via their impact on acoustics but not visible. From this question, this fMRI study aimed at determining multisensory and modality-specific processing of tongue and lip speech actions, with tongue movements of our interlocutor usually "audible" but not visible and lip movements both "audible" and visible.

Motor Resonance in Biological Action Recognition

Although information from different sensory modalities, such as sight and/or sound, is processed in unisensory and multisensory brain areas, several studies have identified a central role for motor representations in action recognition. These results appear in keeping with the long-standing proposal that perception and action are two closely linked processes and with more recent neurophysiological perspectives based on the existence of mirror neurons in nonhuman primates and on an action–perception matching system in humans (for reviews, see Rizzolatti & Craighero, 2004; Rizzolatti, Fogassi, & Gallese, 2001). Mirror neurons are polymodal visuomotor or audio–visuomotor neurons in the ventral premotor and posterior parietal cortices (areas F5 and PF) of the macaque monkey, which have been shown to discharge both when the monkey performs hand or mouth actions and when it views or listens to similar actions made by another individual (e.g., Fogassi et al., 2005; Ferrari, Gallese, Rizzolatti, & Fogassi, 2003; Keysers et al., 2003; Kohler et al., 2002; Gallese, Fadiga, Fogassi, & Rizzolatti, 1996; Rizzolatti, Fadiga, Gallese, & Fogassi, 1996; Di Pellegrino, Fadiga, Fogassi, Gallese, & Rizzolatti, 1992). The existence of mirror neurons thus suggests that action observation partly involves the same neural circuits that are used in action performance. Since then, auditory–vocal mirror neurons have also been recorded in nonmammalian vertebrates (Prather, Peters, Nowicki, &

[1]CNRS UMR 5216 & Grenoble Université, [2]Université Grenoble-Alpes & CHU de Grenoble, [3]CNRS UMS 3552, Grenoble, France, [4]CNRS UMR 7309 & Aix-Marseille Université

Journal of Cognitive Neuroscience X:Y, pp. 1–19
doi:10.1162/jocn_a_01057

Mooney, 2008), and numerous neurophysiological and brain imaging experiments have provided evidence for the existence of a frontoparietal action–perception matching system in humans (Rizzolatti & Craighero, 2004). Altogether, these studies demonstrate that sensory information related to biological movements is not only processed in sensory regions but also in the observer's motor system and partly relies on his or her own motor knowledge.

From that view, a stronger activity in the premotor cortex and the posterior parietal cortex is observed during visual and audiovisual perception of biological movements, compared with nonbiological movements (e.g., Saygin, 2007; Calvert, Campbell, & Brammer, 2000; Howard et al., 1996). Moreover, hearing action-related sounds like knock on the door or hand clapping or more complex auditory material like a piano piece also activates motor and premotor regions (e.g., Lahav, Saltzman, & Schlaug, 2007; Pizzamiglio et al., 2005; Aziz-Zadeh, Iacoboni, Zaidel, Wilson, & Mazziotta, 2004; Haueisen & Knösche, 2001). These results support the long-standing theoretical proposal that specific constraints and regularity in biological motion and kinematics are used in action recognition (Viviani & Stucchi, 1992; Johansson, 1973), even when they are roughly represented by point lights (Loula, Prasad, Harber, & Shiffrar, 2005; Beardsworth & Buckner, 1981). Furthermore, action recognition seems to rely not only on biological features per se but also more specifically on a motor repertoire shared by individuals of the same species and related to their relevant physical and/or communicative ability for perceptual processing. For example, Tai, Scherfler, Brooks, Sawamoto, and Castiello (2004) observed premotor activity during the sight of human hand grasp but not during the sight of the same action performed by a robot, which supports the use of a human biological motor repertoire in action recognition. On their side, Buccino et al. (2004) showed that the observation of a biting action performed by humans, monkeys, or dogs induced motor activity in humans, contrary to what happens during the observation of dog-specific barking movements. Calvo-Merino and colleagues (Calvo-Merino, Grèzes, Glaser, Passingham, & Haggard, 2006; Calvo-Merino, Glaser, Grèzes, Passingham, & Haggard, 2005) also showed that, apart from visual familiarity, the involvement of motor areas during action observation strongly relies on motor learning. They indeed observed, among other parietal and cerebellar regions, stronger premotor cortex activity when male dancers viewed dance movements from their own motor repertoire compared with female dance movements that they often saw but never performed. Although a causal role of the motor system during action recognition is still debated, these fMRI studies suggest a strong correlation between motor activity and action observation.

Motor Resonance Extends to Speech Action

Speech is a special type of biological human actions that interfaces with the linguistic system and requires an accurate control of our speech articulators (i.e., the lips, the tongue, the jaw, the velum, and the larynx). As with other type of actions, such as grasping or walking, several neuroimaging studies suggest that speech recognition is also partly mediated by the motor system. Brain areas involved in the planning and execution of speech actions (i.e., the posterior part of the left inferior frontal gyrus, the premotor and primary motor cortices) have indeed shown neural responses during auditory speech perception (e.g., Pulvermuller et al., 2006; Wilson & Iacoboni, 2006; Wilson, Saygin, Sereno, & Iacoboni, 2004). In addition, repetitive and double-pulse TMS studies also suggest that speech motor regions are causally recruited during auditory speech categorization, especially in case of complex situations (e.g., the perception of acoustically ambiguous syllables or when phonological segmentation or working memory processes are strongly required; Grabski et al., 2013; d'Ausilio, Bufalari, Salmas, & Fadiga, 2011; d'Ausilio et al., 2009; Möttönen & Watkins, 2009; Sato, Tremblay, & Gracco, 2009; Meister, Wilson, Deblieck, Wu, & Iacoboni, 2007). Taken together, these results support the idea that our motor knowledge used to produce speech sounds helps to partly constraint phonetic decoding of the sensory inputs, as proposed in motor and sensorimotor theories of speech perception and language comprehension (Pickering & Garrod, 2013; Schwartz, Ménard, Basirat, & Sato, 2012; Skipper, Van Wassenhove, Nussman, & Small, 2007; Liberman & Mattingly, 1985).

Importantly, speech provides visual as well as auditory information. Although humans are proficient to extract phonetic features from the acoustic signal alone and, to a lesser extent, are capable to partly read on lips when audition is lacking, interactions between auditory and visual modalities are beneficial in speech perception. Neuroimaging studies demonstrate the existence of specific brain areas playing a key role in the audiovisual integration of speech. Notably, activity within unisensory visual and auditory regions (the visual motion-sensitive cortex, V5/MT, and the Heschl's gyrus) as well as within multisensory regions (the posterior parts of the left superior temporal gyrus/STS [pSTS/pSTG]) is modulated during audiovisual speech perception, when compared with auditory and visual unimodal conditions (Skipper et al., 2007; Skipper, Nusbaum, & Small, 2005; Callan et al., 2003, 2004; Calvert et al., 2000). Because pSTS/pSTG displays supra-additive and subadditive responses during congruent and incongruent stimuli presentation, it has been proposed that both visual and auditory speech information are integrated in these high-level multisensory integrative regions and that modulations of neuronal responses within the sensory-specific cortices would then be due to feedback projections from this multisensory region. Such modulations would represent the physiological correlates of the perceptual changes experienced after multisensory integration (e.g., Beauchamp, 2005; Beauchamp, Argall, Bodurka, Duyn, & Martin, 2004; Beauchamp, Lee, Argall, & Martin, 2004; Calvert

et al., 2000). In addition, premotor and motor cortices, known to play a crucial role in speech production, might also play a key role in audiovisual integration mechanisms in speech perception (e.g., Sato, Buccino, Gentilucci, & Cattaneo, 2010; Watkins & Paus, 2004; Calvert & Campbell, 2003; Watkins, Strafella, & Paus, 2003; Campbell et al., 2001). From that view, Skipper and colleagues (2005, 2007) observed stronger activation in speech motor regions during audiovisual speech perception, compared with auditory and visual unimodal conditions. Callan and colleagues (2003, 2004) also demonstrated increased motor activity under adverse listening or viewing conditions during bimodal speech presentation. In addition, increased activity or, on the contrary, subadditive responses in the Broca's area have also been reported during the perception of incongruent compared with congruent audiovisual speech stimuli (Pekkola et al., 2006; Ojanen et al., 2005) or compared with unimodal speech stimuli (Calvert et al., 2000). From these results, multisensory areas and speech motor regions appear as good candidates for brain areas where acoustic and visual speech signals can interact, which suggests a possible integration between incoming sensory signals and speech motor knowledge specific to the listener.

Motor Resonance for Audible but Hidden Actions

If the motor system is indeed involved in multisensory integration, what happens when an action is not accessible to one sensor in the daily experience—typically audible but not visible? We know from the classic studies by Meltzoff and Moore (1977, 1983) that 3-week-old infants, and even newborns, are able to associate from birth a visual action they have never seen, like lip and tongue protrusion, with motor commands, possibly through the use of their proprioceptive system. This indirectly suggests that, in adults, the sensorimotor network could play a role in the visual processing of audible but not visible actions by enabling a transfer of motor knowledge toward an inferred visual experience, possibly combined with past auditory and somatosensory experiences.

Lips and tongue are two perfect articulators to test this specific question. First, we have an excellent somatosensory–motor control of both articulators, notably during speaking. Second, because of their position inside the vocal tract, tongue movements of our interlocutor are usually "audible" but not visible, whereas lip movements are both "audible" and visible. Interestingly, few behavioral studies using virtual tongue movements or ultrasound images of tongue movements demonstrate stronger speech learning with a visual tongue feedback (Katz & Mehta, 2015) and an enhancement of auditory stimuli discrimination when they are matched with related visual tongue movements compared with auditory-only or incongruent audio-visuo-lingual stimuli (d'Ausilio, Bartoli, Maffongelli, Berry, & Fadiga, 2014; Badin, Tarabalka, Elisei, & Bailly, 2010).

To determine the neural networks involved in the perceptual processing of visuo-lingual and visuo-facial actions, an fMRI study on unimodal and multimodal speech perception was conducted. Participants had to recognize auditory, visual, or audiovisual speech stimuli, with the visual presentation related to either a sagittal view of the tongue movements or a facial view of the lip movements of a speaker, with lingual and facial movements previously recorded by an ultrasound imaging system and a video camera. Our first goal was to determine the shared neural correlates of visual and audiovisual tongue and lip movements as well as the neural specificity of lingual perception compared with facial perception. We also examined possible similarities and differences in the integration between audio-visuo-lingual and audio-visuo-facial modalities and the correlation between neural activity and visual syllable recognition scores.

METHODS

Participants

Fourteen healthy adults (seven women and seven men with a mean age of 26 years, ranging from 18 to 44 years), who are native French speakers, participated in the study after giving their informed consent. Two participants were removed from the study because of excessive head movements or technical problems during MRI acquisition. All participants were right-handed according to standard handedness inventory (Oldfield, 1971), had normal or corrected-to-normal vision, and reported no history of speaking, hearing, or motor disorders. The protocol was approved by the Grenoble University Ethical Committee with all participants screened for neurological, psychiatric, and other possible medical problems and contraindications to MRI. None of the participants were experienced with visuo-lingual ultrasound images.

Stimuli

Before the experiment, multiple utterances of /pa/, /ta/, and /ka/ syllables were individually recorded by one male and one female speakers in a soundproof room. These syllables were selected based on previous studies on audiovisual speech perception to ensure a gradient of visuo-labial saliency (with notably the bilabial /p/ consonant known to be more visually salient than alveolar /t/ and velar /k/ consonants). Regarding visuo-lingual saliency, /t/ and /k/ consonants have more visible tongue movement than /p/ because of the involvement of the apex or the dorsum of the tongue during alveolar or velar occlusion (see Figure 1).

Synchronous recordings of auditory, visual, and ultrasound signals were acquired by the Ultraspeech system

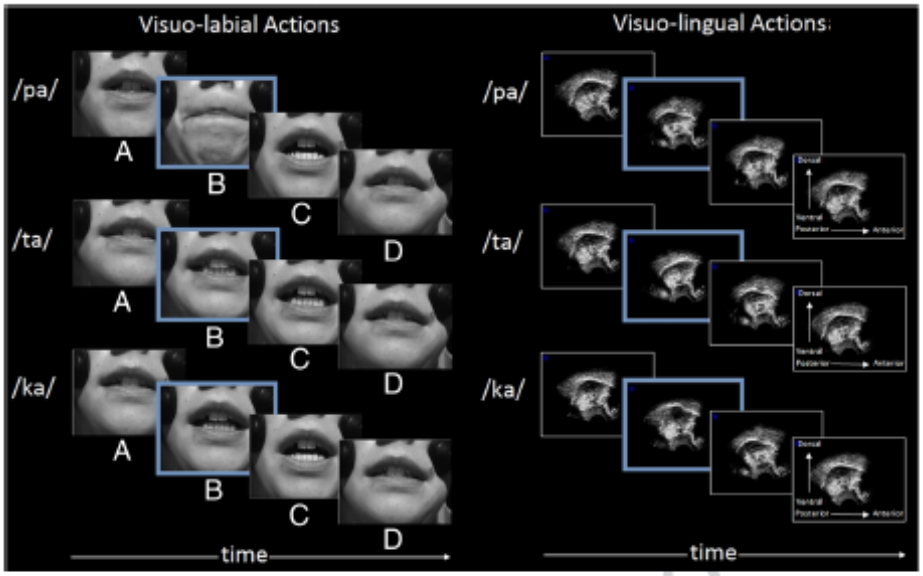

Figure 1. Examples of visual stimuli related to lip and tongue movements for /pa/, /ta/, and /ka/ syllables at four crucial moments: (A) initial neutral position, (B) closure of the vocal tract (in red, /pa/: bilabial occlusion; /ta/: alveolar occlusion, with tongue behind the teeth; /ka/: velar occlusion, with tongue against the palate), (C) vowel production with a maximum opening of the mouth and with the tongue at the back of the vocal tract, and (D) ending neutral position.

(Hueber, Chollet, Denby, & Stone, 2008) composed of a Terason T3000 ultrasound scanner, a 140° microconvex transducer with 128 elements (tongue movements acquired with a sampling rate of 60 fps with a 320 × 240 pixel resolution), an industrial USB color camera (facial movements acquired with a sampling rate of 60 fps with a 640 × 480 pixel resolution), and an external microphone connected to an RME Fireface800 soundcard (audio digitizing at 44.1 kHz with 16-bit quantization recording).

Two clearly articulated /pa/, /ta/, and /ka/ tokens were selected per speaker (with the speaker initiating each utterance from a neutral mid-open mouth position), providing 12 syllables altogether. Sixty stimuli were created consisting of the 12 distinct /pa/, /ta/, and /ka/ syllables related to five conditions: an auditory condition (A), two visual conditions related to either facial (i.e., lip movements) or tongue movements of a speaker (V_F, V_T), and two audiovisual conditions including either facial or tongue movements of a speaker (AV_F, AV_T). The auditory signal intensities were normalized using a common maximal amplitude criterion, and each movie was 80 frames long (1333 msec). To limit possible effects of predictability, variability was introduced with different acoustic consonantal onsets (mean = 450 msec, SD = 193 msec), acoustic durations (mean = 514 msec, SD = 139 msec), visuo-facial onsets (mean = 250 msec, SD = 149 msec), and visuo-lingual onsets (mean = 276 msec, SD = 252 msec), while keeping temporal congruency between auditory and visual signals in audiovisual conditions.

Procedure

Behavioral Experiment

Before the fMRI session, participants were first presented with a subset of the recorded speech stimuli, with short explanations about the tongue movements during the production of /pa/, /ta/, and /ka/ syllables and how these movements are imaged by the ultrasound system. They then underwent a three-alternative forced-choice identification task, having been instructed to categorize as quickly as possible each perceived syllable with their right hand. Participants sat in front of a computer monitor at a distance of approximately 50 cm. The acoustic stimuli were presented at a comfortable sound level through headphones, with the same sound level set for all participants. The Presentation software (Neurobehavioral Systems, Albany, CA) was used to control the stimulus presentation and to record key responses. The experiment consisted of 60 trials presented in a randomized sequence, with 12 trials related to each modality of presentation (A, V_F, V_T, AV_F, and AV_T). The intertrial interval was of 3 sec, and the response key designation was fully counterbalanced across participants. Importantly, participants did not receive any feedback regarding their performance.

fMRI Experiment

Immediately after the behavioral experiment, the fMRI session consisted of one anatomical scan and one functional

run. During the functional run, participants were instructed to attentively listen to and/or watch speech stimuli related to /pa/, /ta/, and /ka/ syllables presented in five different modalities (A, V_F, V_T, AV_F, and AV_T). All stimuli were presented in silent interscanning periods because of sparse sampling acquisition, with the time interval between each stimulus onset and the midpoint of the following functional scan acquisition being set at 5 sec (see below). There were 144 trials, with an 8-sec intertrial interval, consisting of 24 trials for each modality of presentation (with each syllable presented two times) and 24 trials related to a resting condition without any sensory stimulation.

Data Acquisition

Magnetic resonance images were acquired with a 3-T whole-body MR scanner (Philips Achieva TX). Participants lay in the scanner with head movements minimized with a standard birdcage 32-channel head coil and foam cushions. Visual stimuli were presented using the Presentation software (Neurobehavioral Systems, Albany, CA) and displayed on a screen situated behind the scanner via a mirror placed above the participant's eyes. Auditory stimuli were presented through the MR-confon audio system (www.mr-confon.de).

A high-resolution T1-weighted whole-brain structural image was acquired for each participant before the functional run (magnetization prepared rapid gradient echo, sagittal volume of 256 × 224 × 176 mm^3 with a 1-mm isotropic resolution, inversion time = 900 msec, two segments, segment repetition time = 2500 msec, segment duration = 1795 msec, repetition time [TR]/echo time = 16/5 msec with 35% partial echo, flip angle = 30°).

Functional images were obtained in a subsequent functional run using a T2*-weighted EPI sequence with whole-brain coverage (TR = 8 sec, acquisition time = 3000 msec, echo time = 30 msec, flip angle = 90°). Each functional scan was composed of 53 axial slices parallel to the AC–PC plane acquired in noninterleaved order (72 × 72 matrix, field of view = 216 mm, 3 × 3 mm^2 in-plane resolution with a slice thickness of 3 mm without gap). To reduce acoustic noise, a sparse sampling acquisition was used (Birn, Bandettini, Cox, & Shaker, 1999; Hall et al., 1999). This acquisition technique is based on neurophysiological properties of the slowly rising hemodynamic response, which is estimated to occur with a 4- to 6-sec delay in case of speech perception (Grabski et al., 2013; Zaehle et al., 2007). In this study, functional scanning therefore occurred only during a fraction of the TR, alternating with silent interscanning periods, where stimuli were presented. All conditions were presented in a pseudorandom sequence. In addition, three "dummy" scans at the beginning of the functional run were added to allow for equilibration of the MRI signal and were removed from the analyses.

Data Analyses

Behavioral Analysis

For each participant and modality, the percentage of correct responses and median RTs (from the onset of the acoustic syllables) were computed. For each dependent variable, a repeated-measures ANOVA was performed with the modality (A, V_F, V_T, AV_F, and AV_T) as the within-participant variable. For both analyses, the significance level was set at $p = .05$ and Greenhouse–Geisser corrected (for violation of the sphericity assumption) when appropriate. When required, post hoc analyses were conducted with Newman–Keuls tests.

fMRI Analysis

fMRI data were analyzed using the SPM8 software package (Wellcome Department of Imaging Neuroscience, Institute of Neurology, London, United Kingdom) running on MATLAB (The MathWorks, Natick, MA). Brain-activated regions were labeled using the SPM Anatomy toolbox (Eickhoff et al., 2005) and, if a brain region was not assigned or not specified in the SPM Anatomy toolbox, using the Talairach Daemon software (Lancaster et al., 2000). For visualization, activation maps were superimposed on a standard brain template using the MRICRON software (www.sph.sc.edu/comd/rorden/mricron/).

Data preprocessing steps for each participant included rigid realignment of functional images, coregistration of the structural image to the mean functional image, segmentation and normalization of the structural image to common subject space using the groupwise DARTEL registration method implemented in SPM8, warping of all realigned functional images using deformation flow fields generated from the normalization step, transformation into the Montreal Neurological Institute (MNI) space, and spatial smoothing using an 8-mm FWHM Gaussian kernel.

For individual analyses, neural activations related to the perceptual conditions were analyzed using a general linear model, including five regressors of interest (A, V_F, V_T, AV_F, and AV_T) and the six realignment parameters, with the silent trials forming an implicit baseline. The BOLD response for each event was modeled using a single-bin finite impulse response basis function spanning the time of acquisition (3 sec). Before estimation, a high-pass filtering with a cutoff period of 128 sec was applied. Beta weights associated with the modeled finite impulse responses were then computed to fit the observed BOLD signal time course in each voxel for each condition. Individual statistical maps were calculated for each perceptual condition with the related baseline and subsequently used for group statistics.

To draw population-based inferences, a second-level random effects group analysis was carried out with the modality (A, V_F, V_T, AV_F, and AV_T) as the within-participant variable and the participants treated as a random factor.

First, for each modality, brain activity compared with the resting baseline was evaluated. Second, to determine common neural activity across modalities, several conjunction analyses were performed (i.e., $V_F \cap V_T$, $AV_F \cap AV_T$, $A \cap V_F \cap AV_F$, $A \cap V_T \cap AV_T$, $A \cap V_F \cap V_T \cap AV_F \cap AV_T$). Third, activity differences between visual conditions and between audiovisual conditions were evaluated (i.e., $V_F > V_T$, $V_T > V_F$, $AV_F > AV_T$, $AV_T > AV_F$). Fourth, to determine possible correlations between perceptual responses observed in the behavioral experiment and BOLD responses, covariate analyses were performed on the whole brain between neural activity in visual and audiovisual modalities (i.e., V_F, AV_F, V_T, AV_T) and visual identification scores as well as RTs related to visuo-lingual and visuo-facial speech movements (V_F, V_T). In addition, brain regions showing higher or lower audiovisual responses compared with unimodal auditory and visual responses were identified using the max criterion test (i.e., $[AV_F > A] \cap [AV_F > V_F]$, $[AV_F < A] \cap [AV_F < V_F]$, $[AV_T > A] \cap [AV_T > V_T]$, $[AV_T < A] \cap [AV_T < V_T]$; see Stevenson et al., 2014). Modality, conjunction, and correlation contrasts were calculated with the significance level set at $p < .05$, family-wise error (FWE) corrected at the voxel level with a cluster extent of at least 20 voxels. All other contrasts were calculated with a significance level set at $p < .001$ uncorrected at the voxel level with a cluster extent of at least 20 voxels.

RESULTS

Behavioral Results

Overall, the mean proportion of correct responses was 82%. The main effect of modality was significant ($F(4, 52) = 37.79$, $p < .001$), with more correct responses in the A, AV_F, and AV_T conditions than in the V_F condition and in V_F compared with V_T conditions (on average, A = 98%, AV_F = 98%, AV_T = 95%, V_F = 70%, V_T = 49%; all mentioned comparisons significant). The ANOVA on RTs demonstrated a significant effect of the modality ($F(4, 52) = 36.25$, $p < .001$), with faster RTs in AV_F than in V_F, A, AV_T, and V_T conditions and slower RTs in V_T than in the other conditions (on average, AV_F = 722 msec, V_F = 774 msec, A = 812 msec, AV_T = 913 msec, V_T = 1241 msec; all mentioned comparisons significant).

Importantly, despite slower RTs and lower recognition scores for visuo-lingual stimuli compared with visuo-facial

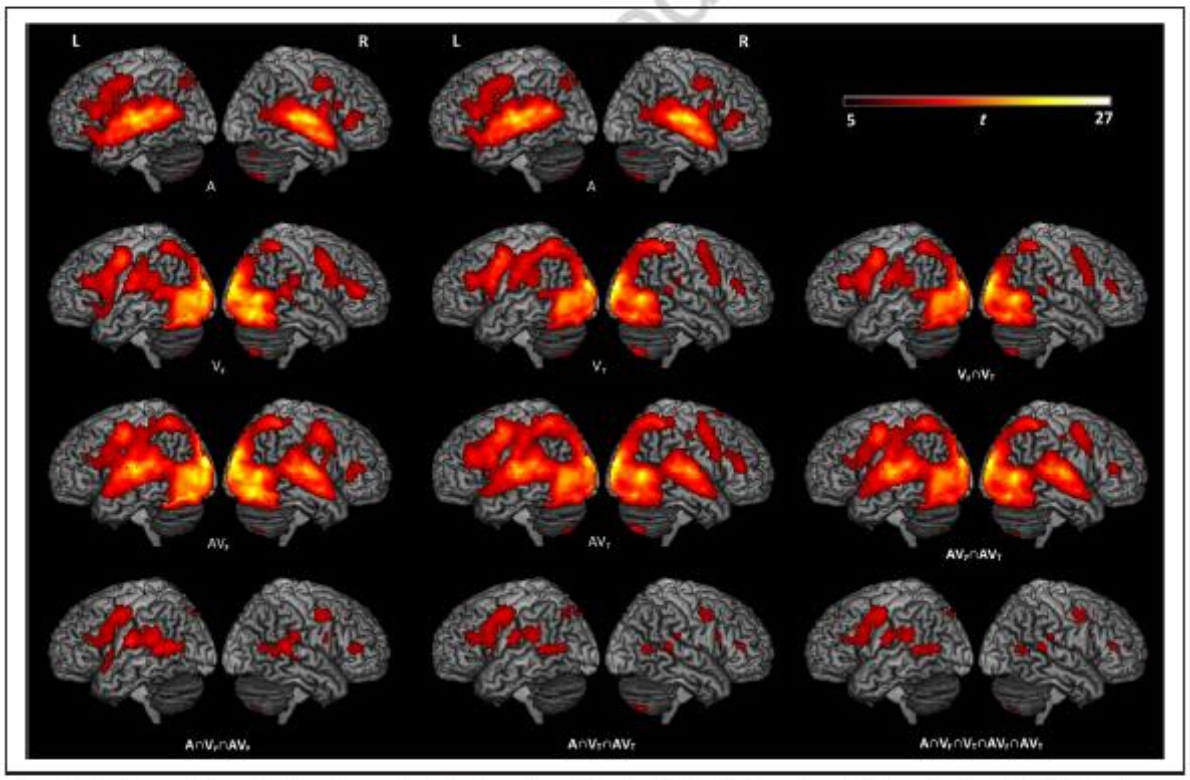

Figure 2. Surface rendering of brain regions activated in the auditory (A), visuo-facial (V_F), visuo-lingual (V_T), audio-visuo-facial (AV_F), and audio-visuo-lingual (AV_T) conditions and showing overlapping activity between lip-related conditions (conjunction $A \cap V_F \cap AV_F$), tongue-related conditions (conjunction $A \cap V_T \cap AV_T$), visual conditions (conjunction $V_F \cap V_T$), and audiovisual conditions (conjunction $AV_F \cap AV_T$) and between all modalities (conjunction $A \cap V_F \cap V_T \cap AV_F \cap AV_T$; $p < .05$, FWE corrected; cluster extent threshold of 20 voxels).

stimuli (and to the other conditions), recognition scores for visuo-lingual stimuli remained above chance level (i.e., 49% vs. 33%). Interestingly, at the syllable level, individual differences were observed between facial and tongue visual recognition (V_F: /pa/ 100%, /ta/ 64%, /ka/ 45%; V_T: /pa/ 50%, /ta/ 50%, /ka/ 46%; no statistical analyses were performed because of the small number of trials for each syllable). These differences suggest different categorization processes because of the nature of the stimuli.

Table 1. Maximum Activation Peak Summary and Contrast Estimates of Brain Regions Showing Overlapping Activity between All Conditions (Conjunction A ∩ V_F ∩ V_T ∩ AV_F ∩ AV_T; $p < .05$, FWE Corrected, Cluster Extent Threshold of 20 Voxels)

			MNI Coordinates				*Contrast Estimates*				
Regions	*BA*	*H*	*x*	*y*	*z*	*t*	*A*	V_F	V_T	AV_F	AV_T
Auditory Cortex											
STG	22	L	−50	−44	8	8.16	0.12	0.11	0.09	0.12	0.14
Middle temporal gyrus	39	L	−58	−56	6	7.23	0.12	0.13	0.10	0.11	0.13
Middle temporal gyrus	39	R	58	−62	8	6.63	0.08	0.14	0.13	0.15	0.13
STG	22	R	54	−60	12	6.39	0.10	0.14	0.11	0.14	0.14
Heschl's gyrus	42	R	56	−38	12	7.70	0.15	0.14	0.12	0.17	0.17
Parietal Cortex											
Parietal operculum (OP4)	40/43	L	−64	−14	16	8.32	0.13	0.10	0.07	0.13	0.13
Parietal operculum (OP1)	40/43	L	−58	−18	22	8.14	0.09	0.11	0.09	0.11	0.11
Inferior parietal lobule	40	L	−60	−34	20	7.77	0.19	0.14	0.12	0.21	0.20
Inferior parietal lobule	40	R	66	−28	22	7.03	0.14	0.13	0.10	0.19	0.17
Motor Cortex											
Primary motor cortex	4	L	−54	−6	46	9.08	0.14	0.20	0.18	0.21	0.22
Premotor cortex	6	L	−52	2	44	8.60	0.14	0.18	0.20	0.18	0.21
Insula	13	L	−36	10	24	8.50	0.12	0.15	0.13	0.13	0.15
Middle frontal gyrus	9	L	−44	12	28	8.26	0.19	0.23	0.23	0.19	0.25
pFC											
Inferior frontal gyrus (pars triangularis)	45	L	−52	30	24	6.49	0.08	0.12	0.13	0.08	0.13
Inferior frontal gyrus (pars opercularis)	44	L	−58	8	32	6.69	0.09	0.12	0.15	0.10	0.11
Inferior frontal gyrus (pars triangularis)	45	R	54	36	12	7.07	0.09	0.12	0.10	0.10	0.11
Superior frontal gyrus	6	L	−6	4	60	8.91	0.12	0.13	0.13	0.12	0.12
Middle frontal gyrus	9	R	56	2	44	7.20	0.11	0.13	0.10	0.16	0.12
Middle frontal gyrus	9	R	36	8	24	7.16	0.09	0.11	0.09	0.09	0.11
Other Regions											
Associative visual cortex	V5	L	−54	−66	10	7.87	0.10	0.15	0.12	0.16	0.14
Precuneus	7	L	−8	−78	46	8.46	0.27	0.36	0.35	0.39	0.41
Cerebellum (VIIb)		R	16	−74	−50	8.05	0.06	0.06	0.08	0.06	0.08
Anterior cingulate gyrus	32	L	−4	16	42	T6.41	0.12	0.16	0.18	0.12	0.14

fMRI Results: Modality and Conjunction Analyses

Brain activity compared with the resting baseline in each modality (A, V_F, V_T, AV_F, and AV_T) as well as conjunction analyses (i.e., $V_F \cap V_T$, $AV_F \cap AV_T$, $A \cap V_F \cap AV_F$, $A \cap V_T \cap AV_T$, $A \cap V_F \cap V_T \cap AV_F \cap AV_T$) are displayed in Figure 2. Globally, bilateral activity of auditory regions (including primary, secondary, and associative areas in the STG and extending to the middle temporal gyrus) as well as strong premotor activations (extending to the inferior frontal gyrus and left primary motor cortex) were observed in A condition (see Figure 2, Condition A). In both V_F and V_T conditions, visual (bilateral primary and associative regions, including V5), auditory (pSTS and pSTG), and motor (bilateral primary motor and premotor cortices as well as inferior frontal gyri) activities were observed (see Figure 2, Conditions V_F and V_T as well as conjunction $V_T \cap V_F$). Activities in AV_F and AV_T conditions were mainly found in primary and associative auditory and visual regions and in motor and frontal cortices (see Figure 2, Conditions AV_F and AV_T as well as conjunction $AV_F \cap AV_T$).

Importantly, common activations in all five conditions (see Table 1 and Figure 2, conjunction $A \cap V_F \cap V_T \cap AV_F \cap AV_T$) were observed in the pSTS, bilaterally extending to the adjacent posterior middle temporal gyrus and left V5. Additional auditory activity was also observed bilaterally in the posterior temporal gyrus, extending to the right secondary auditory cortex, the parietal operculum, and the antero-ventral part of the inferior parietal lobule. Interestingly, strong premotor activity was also observed, mainly in the left hemisphere, and also including activity in the opercular part of the left inferior frontal gyrus, the triangular part of the inferior frontal gyrus, the left anterior IC, and the left primary motor cortex. Finally, additional activity was also observed in the ACC, the left precuneus, and the right cerebellum (Lobule VII).

In summary, apart from sensory-specific activity in auditory and visual conditions, our results demonstrate a shared neural network involved in all conditions, mainly including multisensory activity around the pSTS and the pSTG extending to adjacent inferior parietal regions as well as the premotor cortex extending to inferior frontal regions.

fMRI Results: Modality Differences

$V_F > V_T$

Several auditory regions were more activated during visuo-facial than during visuo-lingual perception, with stronger bilateral activation of the posterior temporal gyrus/sulcus, extending to the middle temporal gyrus. Stronger activation of the left anterior temporal gyrus (temporopolar area) and the right primary auditory cortex was also observed. Large parts of the primary and associative visual areas were also more activated (V1, V2, V3, and V4), extending to the fusiform gyrus. In addition, stronger frontal activity was observed in the right pars triangularis and middle frontal gyrus, the left pars orbitalis, and the left anterior IC. Finally, stronger additional activity was also observed in the right BG in the lentiform nucleus and the left precuneus (see Figure 3 and Table 2).

$V_T > V_F$

Bilateral premotor dorsal regions were more activated during visuo-lingual perception than during visuo-facial perception. Interestingly, stronger activity was observed in the primary somatosensory cortices, extending to the adjacent parts of the dorsal inferior parietal lobule and intraparietal sulcus. Stronger bilateral visual activity was also observed, including primary and associative visual areas (V1 and V2). Finally, stronger additional activity was also observed in the right precuneus, the posterior cingulate cortex, and the middle part of the right parahippocampal gyrus.

$AV_F > AV_T$

Audio-visuo-facial stimuli, compared with audio-visuo-lingual stimuli, induced stronger bilateral activation of the

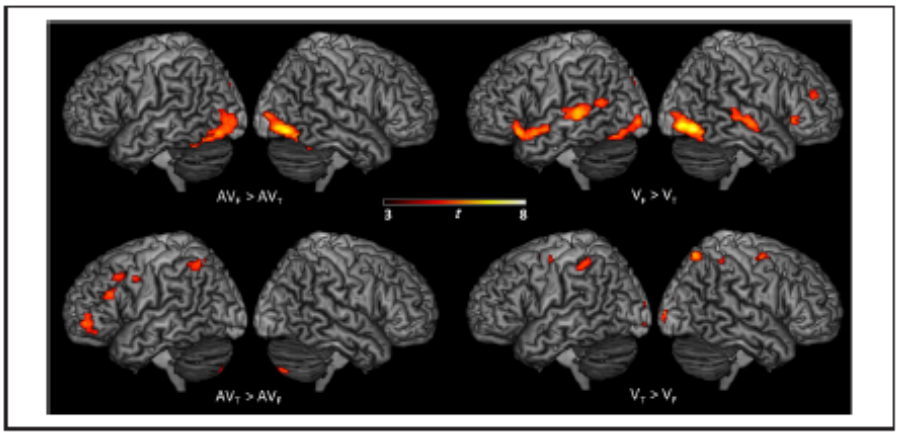

Figure 3. Surface rendering of brain regions activated showing significant change in activity between visual conditions related to lip and tongue movements ($V_F > V_T$ and $V_T > V_F$) and audiovisual conditions related to lip and tongue movements ($AV_F > AV_T$ and $AV_T > AV_F$; $p < .001$ uncorrected; cluster extend threshold of 20 voxels).

Table 2. Maximum Activation Peaks and Contrast Estimates of Brain Regions Showing Significant Change in Activity between Visuo-Facial and Visuo-Lingual Conditions (A: $V_F > V_T$; B: $V_T > V_F$; $p < .001$ Uncorrected, Cluster Extend Threshold of 20 Voxels) and between Audio-Visuo-Facial and Audio-Visuo-Lingual Conditions (C: $AV_F > AV_T$; D: $AV_T > AV_F$; $p < .001$ Uncorrected, Cluster Extend Threshold of 20 Voxels)

			MNI Coordinates				*Contrast Estimates*				
Regions	*BA*	*H*	*x*	*y*	*z*	*t*	*A*	V_F	V_T	AV_F	AV_T
A. VF > VT											
Auditory cortex											
STG	22	L	−64	−34	4	5.38	0.17	0.05	−0.02	0.10	0.12
STG	22	L	−60	−56	14	4.28	0.14	0.09	0.04	0.09	0.09
Heschl's gyrus	42	R	62	−32	8	3.85	0.20	0.06	0.01	0.16	0.15
STG	22	R	44	−22	−6	5.02	0.08	0.03	−0.02	0.05	0.05
Middle temporal gyrus	21	R	60	−14	−6	4.60	0.21	0.04	−0.04	0.14	0.14
Middle temporal gyrus	21	L	−54	6	−16	4.82	0.11	0.03	−0.03	0.08	0.05
Temporopolar area	38	L	−50	−2	−12	4.39	0.10	0.02	−0.03	0.07	0.07
Frontal cortex											
Inferior frontal gyrus (pars orbitalis)	47	L	−40	22	−14	5.29	0.10	0.06	−0.02	0.07	0.05
Inferior frontal gyrus (pars triangularis)	45	R	54	24	−2	4.26	0.06	0.01	−0.04	0.03	0.01
Middle frontal gyrus	10	R	38	44	20	3.82	0.04	0.03	−0.02	0.01	0.02
Insula	13	L	−38	0	−6	3.72	0.08	0.06	0.00	0.08	0.06
Visual cortex											
Primary visual cortex (V1)	17	L	−6	−76	8	5.93	0.04	0.41	0.23	0.47	0.24
Associative visual cortex (V2)	18	L	−30	−92	−4	4.65	0.00	0.23	0.16	0.23	0.16
Associative visual cortex (V3)	19	L	−10	−76	−4	5.37	0.01	0.27	0.13	0.32	0.14
Associative visual cortex (V4)	19	L	−34	−78	−14	4.30	0.01	0.29	0.20	0.29	0.19
Fusiform gyrus	37	L	−28	−72	−16	4.42	0.03	0.41	0.30	0.44	0.30
Primary visual cortex (V1)	17	R	10	−70	12	6.38	0.07	0.41	0.24	0.48	0.25
Associative visual cortex (V2)	18	R	22	−60	8	6.89	0.09	0.28	0.15	0.33	0.19
Associative visual cortex (V3)	19	R	34	−92	4	3.72	0.02	0.24	0.18	0.24	0.16
Associative visual cortex (V4)	19	R	40	−72	−10	7.31	−0.01	0.22	0.13	0.23	0.13
Fusiform gyrus	37	R	38	−50	−16	5.12	0.02	0.23	0.15	0.23	0.14
Other regions											
Lentiform nucleus		R	30	−20	−4	5.46	0.03	0.02	−0.03	0.02	0.03
Precuneus	7	L	−2	−82	36	4.65	0.13	0.39	0.19	0.42	0.24
B. VT > VF											
Motor regions											
Premotor cortex	6	R	26	−4	56	4.61	0.02	0.04	0.10	0.07	0.09
Premotor cortex	6	L	−24	−6	54	4.14	0.04	0.05	0.13	0.05	0.14
Parietal lobule											
Inferior parietal lobule	40	L	−44	−40	50	4.03	0.07	0.10	0.20	0.15	0.22
Primary somatosensory cortex	2	L	−40	−42	52	3.82	0.06	0.10	0.20	0.15	0.23

Table 2. (*continued*)

Regions	*BA*	*H*	*MNI Coordinates* *x*	*y*	*z*	*t*	*Contrast Estimates* *A*	V_F	V_T	AV_F	AV_T
Intraparietal sulcus		R	30	−40	40	4.32	0.01	0.04	0.10	0.06	0.07
Primary somatosensory cortex	2	R	34	−42	50	3.65	0.02	0.11	0.19	0.15	0.18
Primary somatosensory cortex	3	R	32	−32	42	3.54	0.01	0.05	0.09	0.07	0.08
Superior parietal lobule	7	R	24	−68	58	5.20	0.08	0.09	0.23	0.17	0.23
Visual regions											
Primary visual cortex (V1)	17	L	0	−94	0	5.30	0.03	0.42	0.59	0.47	0.59
Associative visual cortex (V2)	18	L	−4	−98	10	4.89	−0.01	0.22	0.31	0.25	0.30
Primary visual cortex (V1)	17	R	14	−94	4	5.86	0.03	0.36	0.49	0.38	0.47
Associative visual cortex (V2)	18	R	8	−86	−10	4.61	0.04	0.31	0.40	0.34	0.39
Other regions											
Posterior cingulate cortex	31	L	−16	−60	22	5.05	0.04	0.02	0.11	0.04	0.06
Posterior cingulate cortex	31	R	18	−58	22	3.80	0.02	0.03	0.12	0.05	0.06
Parahippocampal gyrus	36	R	26	−38	−16	4.33	−0.03	0.00	0.09	0.05	0.04
C. AVF > AVT											
Visual cortex											
Associative visual cortex (V2)	18	L	−2	−74	10	6.76	0.05	0.41	0.24	0.46	0.24
Primary visual cortex (V1)	17	L	−8	−76	10	6.65	0.05	0.38	0.22	0.43	0.22
Associative visual cortex (V3)	19	L	−12	−88	34	4.13	0.07	0.32	0.24	0.35	0.25
Superior parietal lobule (cuneus)	7	L	−2	−84	36	3.82	0.12	0.38	0.21	0.41	0.24
Primary visual cortex (V1)	17	R	10	−72	12	7.89	0.06	0.41	0.25	0.49	0.25
Associative visual cortex (V2)	18	R	20	−62	8	7.07	0.09	0.28	0.16	0.34	0.19
Associative visual cortex (V4)	19	R	40	−72	−10	6.54	−0.01	0.22	0.13	0.23	0.13
Other regions											
Amygdala		L	−18	−6	−14	4.26	0.08	0.10	0.05	0.10	0.02
Amygdala		R	22	−4	−14	3.90	0.06	0.05	0.02	0.09	0.02
Posterior cingulate cortex	31	R	16	−34	42	3.73	0.00	0.00	−0.01	0.04	−0.03
Frontopolar area (Fp2)	10	R	4	54	−10	3.95	0.05	−0.03	−0.07	0.03	−0.07
Temporopolar area	38	R	32	4	−20	4.23	0.13	0.05	0.03	0.14	0.05
D. AVT > AVF											
Parietal cortex											
Inferior parietal lobule	40	L	−44	−62	54	3.91	0.02	−0.02	−0.01	−0.04	0.04
pFC											
Premotor cortex	6	L	−42	0	36	3.71	0.10	0.17	0.18	0.10	0.19
Middle frontal gyrus	9	L	−38	2	36	3.78	0.11	0.15	0.17	0.10	0.18
Middle frontal gyrus	8	L	−52	16	42	3.63	0.11	0.03	0.02	0.01	0.10
Dorsolateral pFC	46	L	−46	26	24	4.25	0.07	0.10	0.11	0.05	0.12

Table 2. (*continued*)

			MNI Coordinates				Contrast Estimates				
Regions	*BA*	*H*	*x*	*y*	*z*	*t*	*A*	V_F	V_T	AV_F	AV_T
Dorsolateral pFC	10	L	−36	50	−4	4.39	0.05	0.02	0.04	−0.01	0.05
Dorsolateral pFC	11	L	−36	46	−6	4.30	0.04	0.01	0.02	−0.01	0.04
Other regions											
Primary visual cortex (V1)	17	L	−8	−100	2	4.25	0.02	0.26	0.33	0.24	0.33
Cerebellum (VIIb)		R	20	−76	−48	4.35	0.05	0.06	0.08	0.03	0.09
ACC	32	L	−22	42	4	3.69	0.00	−0.02	−0.01	−0.02	0.01

primary and associative visual areas (V1, V2, V3, and V4). Stronger activity was also observed in the amygdala and the right posterior cingulate gyrus as well as in the right temporopolar and frontopolar areas.

$AV_T > AV_F$

Audio-visuo-lingual stimuli, compared with audio-visuo-facial stimuli, induced stronger activation of the left premotor cortex, extending to the adjacent middle and inferior frontal gyri, and the left dorsal inferior parietal lobule, extending to the intraparietal sulcus. Stronger additional activity was also observed in the left dorsolateral pFC, the left primary visual cortex, the right cerebellum (Lobule VII), and the left ACC.

To summarize, seeing tongue-related stimuli globally induced stronger motor and somatosensory activity, whereas auditory and visual cortices were globally more activated during lip-related stimuli presentation.

fMRI Results: Correlation between Visual Recognition Scores and Neural Activity

For tongue-related stimuli, the covariance analysis between visual recognition scores in the behavioral experiment and BOLD activity observed in V_T and AV_T conditions in the fMRI experiment demonstrated a significant correlation in the left dorsal part of the premotor cortex (see Figure 4 and Table 3).

For lip-related stimuli, a significant correlation was observed between visual recognition scores and neural responses in the V_F condition in the right primary, secondary, and associative (MT/V5) visual regions and in the right fusiform gyrus. Similarly, a significant correlation in the AV_F condition was observed in the bilateral associative visual cortex, in the left fusiform gyrus, in the lingual gyrus, in the left cerebellum, and in the parahippocampal gyrus.

To summarize, a correlation between visual recognition scores and neural activity was observed in the left premotor cortex for tongue-related stimuli and in visual regions for lip-related stimuli.

fMRI Results: Correlation between Visual RTs and Neural Activity

For both lip- and tongue-related stimuli, the covariance analysis between RTs observed for unimodal visual stimuli in the behavioral experiment and BOLD activity observed in visual and audiovisual conditions in the fMRI

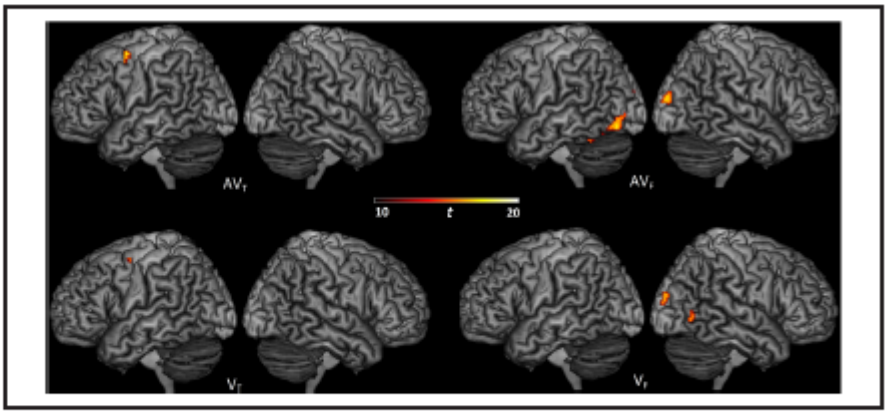

Figure 4. Surface rendering of brain regions activated showing correlation between visual recognition scores and neural activity in the audio-visuo-facial (AV_F), audio-visuo-lingual (AV_T), visuo-facial (V_F), and visuo-lingual (V_T) conditions ($p < .05$, FWE corrected; cluster extent threshold of 20 voxels).

Table 3. Maximum Activation Peaks Showing Correlation between Visual Recognition Scores and Neural Activity in the (A) Visuo-Lingual (V_T), (B) Audio-Visuo-Lingual (AV_T), (C) Visuo-Facial (V_F), and (D) Audio-Visuo-Facial Conditions (AV_F; $p < .05$, FWE Corrected, Cluster Extent Threshold of 20 Voxels)

			MNI Coordinates			
Regions	*BA*	*H*	*x*	*y*	*z*	*t*
A. V_T						
Premotor cortex	6	L	−34	−4	54	16.65
B. AV_T						
Premotor cortex	6	L	−34	0	54	16.34
C. V_F						
Visual cortex						
Associative visual cortex (MT/V5)	19	R	44	−64	0	11.21
Primary visual cortex (V1)	17	R	22	−60	2	12.21
Associative visual cortex (V2)	18	R	22	−90	20	12.75
Fusiform gyrus	37	R	52	−68	−2	11.82
D. AV_F						
Visual cortex						
Fusiform gyrus	37	L	−36	−50	−22	19.54
Associative visual cortex (V3)	19	L	−34	−76	−12	17.19
Associative visual cortex (V2)	18	R	22	−92	14	14.57
Associative visual cortex (V2)	18	L	−8	−88	22	14.96
Associative visual cortex (V3)	19	R	22	−66	−10	12.03
Lingual gyrus	18	R	8	−74	−8	13.14
Other regions						
Culmen		L	−14	−48	−6	18.82
Declive		L	−30	−58	−16	13.54
Parahippocampal gyrus	19	L	−20	−56	−10	13.93

experiment demonstrated a significant correlation in visual regions (including the primary and associative visual brain areas and the fusiform gyrus). Other correlational activity was found in the superior parietal lobule and adjacent intraparietal sulcus for V_T, V_F, and AV_F conditions as well as in the left premotor cortex for V_F (see Figure 5 and Table 4).

To summarize, a correlation between RTs and neural activity was mainly observed in visual and superior parietal regions for both tongue- and lip-related stimuli.

fMRI Results: Different Audiovisual Neural Responses Compared with Auditory and Visual Modalities

Higher neural responses were only found for audio-visuo-facial stimuli (see Figure 6, condition $[AV_F > A] \cap [AV_F > V]$) around the bilateral secondary visual areas, the right cerebellum, and the parahippocampal gyrus and in the left granular retrosplenial cortex (see Figure 6 and Table 5).

DISCUSSION

Four main results emerged from this fMRI study. First, the neural networks involved in visuo-lingual and visuo-facial perception strongly overlap and share similar sensorimotor brain areas. This suggests comparable visual processing of lingual and labial movements, both crucial for the realization of speech sounds. Second, further analyses demonstrate stronger motor and somatosensory activations during visuo-lingual perception and stronger activation of auditory and visual cortices during visuo-facial perception. This result suggests more important somatosensory–motor internal simulation of the presented syllables for visuo-lingual speech stimuli that in daily life are clearly audible but not visible, whereas visible and audible visuo-facial speech stimuli seem to strongly rely on well-known sensory representations. Third, behavioral results confirm that both visuo-lingual and visuo-facial speech stimuli were correctly recognized, although to a lower extent and slower for visuo-lingual stimuli. Complementing these findings, activity in the left premotor cortex and in visual brain areas was found to correlate with visual recognition scores observed for visuo-lingual and visuo-facial speech stimuli, respectively, whereas visual activity correlated with RTs for both stimuli. Altogether, these results suggest that visual processing of audible but not visible movements induce motor and visual mental simulation of the perceived speech actions to facilitate recognition and/or learn the association between auditory and visual signals.

Syllable Recognition

The recognition scores replicated a number of well-known effects in auditory, visual, and audiovisual speech perception. As expected, perceptual recognition scores show a ceiling effect for auditory and audiovisual modalities. Also consistent with previous studies on unimodal and multimodal speech perception, visual-only syllables

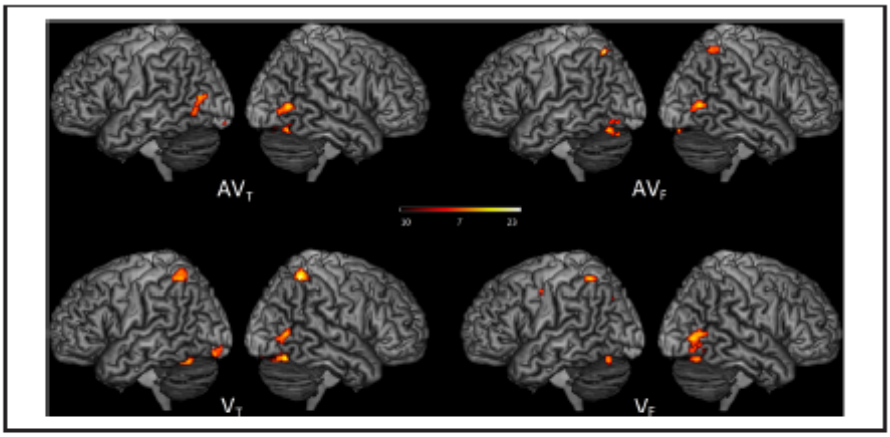

Figure 5. Surface rendering of brain regions activated showing correlation between visual RTs and neural activity in the audio-visuo-facial (AV_F), audio-visuo-lingual (AV_T), visuo-facial (V_F), and visuo-lingual (V_T) conditions ($p < .05$, FWE corrected; cluster extent threshold of 20 voxels).

were less well recognized, especially in the case of tongue movements. In addition, in line with previous studies (Katz & Mehta, 2015; d'Ausilio et al., 2014; Badin et al., 2010), despite lower recognition scores compared with visuo-facial stimuli (and to the other conditions), the recognition of visuo-lingual stimuli remained above chance level.

Regarding RTs, faster recognition was observed when visual information was added to the auditory signal, a result suggesting a temporal advantage of vision on the auditory signal during individual syllable recognition. This effect only happened for familiar visuo-facial speech movements but not for visuo-lingual movements. Contrary to this result, d'Ausilio et al. (2014) found faster RTs for audio-visuo-lingual stimuli when comparing the perception of congruent audio-visuo-lingual syllables with an auditory-only condition with visual noise. The difference between the two studies likely comes from experimental parameters. First, d'Ausilio and colleagues improved the visual recognition of the tongue shape by adding a red line on the tongue surface. In addition, they used more trials, possibly leading to a stronger learning effect for visual tongue movements. Finally, our RTs were calculated from the acoustic onset of the presented consonant, not from the onset of the visual movement, with a clear difference of visual anticipation between labial (strong) and lingual (low) movements. Surprisingly, in our study, audio-visuo-lingual syllables were identified even slower than auditory-only stimuli. This suggests that the sight of tongue movements disrupted and slowed down the final decision processes, even when adding the corresponding auditory signal.

Visuo-lingual and Visuo-facial Speech Stimuli Share a Common Sensorimotor Network

The fMRI results first demonstrate for visuo-facial and visuo-lingual stimuli common overlapping activity between auditory, visual, and audiovisual modalities in the pSTS, extending to the adjacent posterior middle temporal gyrus and left V5. These results appear in line with previous studies indicating a key role of this region in speech processing, biological motion perception (including face perception), and audiovisual integration (e.g., Beauchamp, 2005; Beauchamp, Argall, et al., 2004; Beauchamp, Lee, et al., 2004; Calvert et al., 1997, 2000). Additional auditory activity was also observed bilaterally in the posterior temporal gyrus, extending to the right secondary auditory cortex, the parietal operculum, and the antero-ventral part of the inferior parietal lobule.

In addition, strong premotor activity was also observed, mainly in the left hemisphere, and also including activity in the opercular part of the left inferior frontal gyrus, the triangular part of the inferior frontal gyrus, the left anterior IC, and the left primary motor cortex. These motor and premotor activations are in accordance with previous studies on auditory, visual, and audiovisual speech perception showing a key role of motor regions in speech processing (e.g., Grabski et al., 2013; d'Ausilio et al., 2009, 2011; Sato et al., 2009, 2010; Möttönen & Watkins, 2009; Meister et al., 2007; Skipper et al., 2005, 2007; Pekkola et al., 2006; Pulvermuller et al., 2006; Wilson & Iacoboni, 2006; Ojanen et al., 2005; Callan et al., 2003, 2004; Watkins & Paus, 2004; Wilson et al., 2004; Calvert & Campbell, 2003; Jones & Callan, 2003; Watkins et al., 2003; Campbell et al., 2001; Calvert et al., 2000). It is worthwhile noting that, in this study, participants were only asked to attentively listen to and/or watch speech stimuli. Given the strong motor activity observed in all modalities, it appears quite likely that participants were therefore engaged to some extent in conscious subvocal sensorimotor simulation or covert rehearsal of the presented syllables. This strategy might have occurred especially because of the difficulty to decode visuo-lingual ultrasound images. However, it cannot be concluded whether this subvocal rehearsal strategy was related to some phonetic decision/recognition

Table 4. Maximum Activation Peaks Showing Correlation between Visual RT and Neural Activity in the (A) Visuo-Lingual (V_T), (B) Audio-Visuo-Lingual (AV_T), (C) Visuo-Facial (V_F), and (D) Audio-Visuo-Facial Conditions (AV_F; $p < .05$, FWE Corrected, Cluster Extent Threshold of 20 Voxels)

			MNI Coordinates			
Regions	*BA*	*H*	*x*	*y*	*z*	*t*
A. V_T						
Visual cortex						
Fusiform gyrus	37	R	32	−66	−20	17.40
Associative visual cortex (V3)	19	R	50	−64	4	15.08
Associative visual cortex (V3)	19	L	−26	−88	−14	14.16
Fusiform gyrus	37	R	58	−64	4	15.27
Parietal lobule						
Intraparietal sulcus	7/40	L	−30	−56	54	19.58
Intraparietal sulcus	7/40	R	32	−52	56	22.61
Superior parietal lobule	7	L	−30	−52	52	17.42
Other regions						
Cerebellum	Lobule VI	R	22	−76	−20	13.00
Cerebellum	Lobule VI	L	−28	−58	−22	16.90
B. AV_T						
Visual cortex						
Associative visual cortex (V3)	19	R	48	−64	2	16.75
Associative visual cortex (V3)	19	L	−12	−94	−14	13.13
Associative visual cortex (V2)	18	R	10	−88	−12	12.54
Primary visual cortex (V1)	17	R	6	−88	−10	12.26
Fusiform gyrus	37	R	48	−72	−2	13.08
Middle temporal gyrus	39	L	−38	−72	12	16.63
Fusiform gyrus	37	L	−42	−66	−4	12.55
C. V_F						
Visual cortex						
Associative visual cortex (V2)	18	R	38	−60	−6	17.84
Fusiform gyrus	37	R	50	−72	0	16.79
Parietal lobule						
Superior parietal lobule	7	L	−34	−52	56	17.12
Motor region						
Premotor cortex	6	L	−38	−4	44	12.46
Other region						
Cerebellum	Lobule VIIa	L	−30	−70	−22	14.40
D. AV_F						
Visual cortex						
Associative visual cortex (V3)	19	L	−34	−70	−10	13.70
Associative visual cortex (V3)	19	R	52	−66	6	18.53
Associative visual cortex (V2)	18	R	10	−84	−16	20.70
Primary visual cortex (V1)	17	R	6	−58	4	13.55
Parietal lobule						
Superior parietal lobule	7	L	−24	−64	56	22.77
Superior parietal lobule	7	R	32	−56	58	14.82
Other regions						
Cerebellum	Lobule VI	R	4	−76	−12	11.40
Cerebellum	Lobule VI	L	−34	−70	−20	14.86
Cerebellum	Lobule VIIa	R	22	−84	−20	11.04
Culmen		R	8	−48	−4	14.79

processes or, rather, to an associative learning strategy between the auditory and visual signals. Indeed, the poor temporal resolution of fMRI obviously collapsed the different timings of neural activation corresponding to the "genuine" response in the perceptual/recognition process and the "fake" response caused by such possible mental motor rehearsal, making it difficult to conclude which components are observed.

Neural Specificity of Visuo-lingual and Visuo-facial Processing

Using a less conservative statistical threshold, a direct comparison of audiovisual and visual conditions related

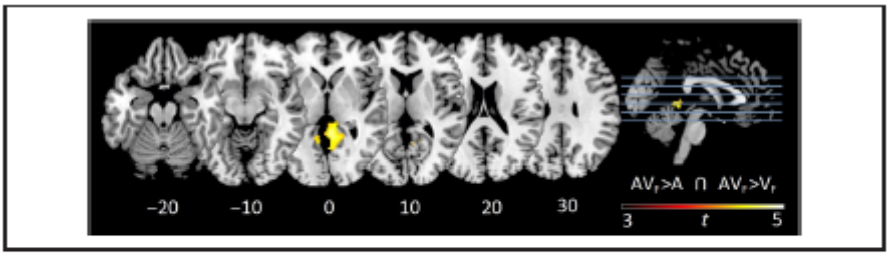

Figure 6. Axial views of brain regions showing higher neural responses (condition [AV_F > A] ∩ [AV_F > V]) in the audio-visuo-facial condition; p < .001 uncorrected, cluster extent threshold of 20 voxels).

to facial or lingual stimuli demonstrates stronger activation of the premotor regions and the primary somatosensory cortices during the observation of tongue movements. Because tongue movements are not usually visible and participants were not experienced with visuo-lingual ultrasound images, this result could be explained by a more important somatosensory–motor covert simulation of tongue movements and the use of both motor and proprioceptive knowledge, to better achieve a phonetic decoding of the presented visuo-lingual stimuli or to learn the association between the two signals. Apart from covert simulation, another explanation could be related to the unusual nature of the lingual stimuli that might imply increased difficulty and high-level categorization processes in the premotor cortex (Venezia, Saberi, Chubb, & Hickok, 2012; Sato et al., 2011).

These somatosensory–motor activations appear however reduced for lip movements. This is likely due to the fact that visuo-facial speech stimuli are perceived in daily life, with their processing being more automatized and requiring less motor simulation. In contrast, in both visual and audiovisual conditions related to lip movements, stronger visual activity was however observed, extending to a large part of primary and associative visual areas. This result might come from low-level features (contrast, luminance, and motion energy), the facial nature as well as stronger visual experience for facial stimuli. In line with previous studies, our results also showed stronger activity within the auditory cortex during lip reading condition than in the visuo-lingual condition. It was indeed demonstrated that syllables' visual cues are sufficient to activate auditory cortical sites, normally engaged during the perception of heard speech, in the absence of auditory speech sound (Campbell et al., 2001; Calvert et al., 1997). This result suggests a direct matching between the visible articulatory movements and auditory representation of the perceived syllables/phonemes. These stronger visual and auditory activations during facial perception could be the result of projections between auditory and visual regions—possibly mediated by the STS. Indeed, studies have demonstrated direct functional and anatomical pathway between primary sensory areas in nonhuman (Cappe & Barone, 2005) and human (Eckert et al., 2008; Watkins,

Table 5. Maximum Activation Peaks and Contrast Estimates of Brain Regions Showing Higher Neural Responses in the Audio-Visuo-Facial Condition (p < .001 Uncorrected, Cluster Extent Threshold of 20 Voxels)

			MNI Coordinates				*Contrast Estimates*				
Regions	*BA*	*H*	*x*	*y*	*z*	*t*	*A*	V_F	V_T	AV_F	AV_T
Visual Cortex											
Associative visual cortex (V2)	18	R	8	−60	−2	4.12	0.06	0.14	0.11	0.27	0.10
Associative visual cortex (V2)	18	L	−10	−54	0	3.85	0.07	0.26	0.18	0.37	0.19
Cerebellum											
Cerebellum (I)		R	4	−44	−2	4.76	0.13	0.23	0.21	0.46	0.27
Other Regions											
Parahippocampal gyrus	30	R	10	−52	4	4.58	0.08	0.20	0.17	0.32	0.16
Parahippocampal gyrus	30	L	−16	−52	2	3.51	0.04	0.18	0.14	0.29	0.13
Granular retrosplenial cortex	29	L	−14	−52	6	3.72	0.02	0.10	0.09	0.20	0.08

Shams, Tanaka, Haynes, & Rees, 2006) cerebral cortex. From that view, lower activation of the visual cortex during the sight of tongue movements could also be explained because such movements are not likely to directly excite the auditory cortex because of their unusual characteristics.

Correlation between Behavioral Performance and Neural Activity

Interestingly, activities in the left premotor cortex and in visual brain areas were found to correlate with visual recognition scores observed for visuo-lingual and visuo-facial speech stimuli, respectively. Hence, the more these areas were activated, the better were the visual recognition scores. These results appear consistent with those observed from the direct comparison between visuo-lingual and visuo-facial movements. As previously noted, given the poor temporal resolution of fMRI, it is however impossible to determine whether motor simulation is related to some recognition/decision processes or rather to some associative learning effect.

Another result is that activity in visual and superior parietal brain areas correlated with RTs for both visuo-facial and visuo-lingual stimuli. Given that these brain regions are known to play a role in visual imagery, this later finding might indicate the use of a visual imagery strategy by the participants to learn the association between auditory and visual signals.

Integration between Auditory and Visual Signals

As previously noted, fMRI studies have demonstrated the existence of specific multisensory brain areas involved in the integration process of auditory and visual signals. More specifically, when compared with auditory and visual unimodal modalities, the observation of audio-visual stimuli was found to induce supra-additive responses in pSTS/pSTG (Beauchamp, 2005; Beauchamp, Argall, et al., 2004; Beauchamp, Lee, et al., 2004; Calvert et al., 2000) as well as subadditive responses in Broca's area (Calvert et al., 2000). Beauchamp (2005) determined two minimal criteria to select brain regions involved in audio-visual speech integration: The region must be activated during auditory, visual, and audiovisual modalities and must display supra-additive audiovisual response. In this study, higher neural responses using the max criterion test ([AV > A] ∩ [AV > V]) were only found for audio-visuo-facial stimuli around the bilateral secondary visual areas, the right cerebellum, and the parahippocampal gyrus and in the left granular retrosplenial cortex. Although a pSTS/pSTG activation was observed for all conditions, no higher response was found for this region supposed to be a specific brain area involved in the integration process. Although we do not have a clear explanation for this null result, one possibility is that the strong sensorimotor activity observed in all modalities, including the pSTS/pSTG, might have changed the classical audiovisual integration network.

Concluding Remarks

Taken together, our results provide new evidence for an action–perception functional coupling in speech processing. According to a recent neurobiological and perceptuo-motor model of multisensory speech perception by Skipper and colleagues (2007), apart from sensory processing, motor activity during speech perception might partly constrain phonetic interpretation of the sensory inputs through the internal generation of candidate articulatory categorizations and, in return, auditory and somatosensory predictions. In this study, because of the lack of visual knowledge in the processing of the generally hidden tongue movements, a larger motor recruitment could have been necessary to infer appropriate motor speech representations to correctly decode the perceived syllables. This process would have been guided by the participant's expertise in speech production, enabling to transfer procedural motor knowledge into a better understanding of such unfamiliar visual stimuli. One alternative explanation is that motor activity does not directly reflect some phonetic decision processes but rather a learning effect between auditory and visual signals.

Visual and motor familiarities have already been compared in the course of action recognition, and previous studies have shown that the involvement of the motor system during action observation strongly relies on motor learning (e.g., Calvo-Merino et al., 2005, 2006). In line with previous behavioral studies (Katz & Mehta, 2015; d'Ausilio et al., 2014; Badin et al., 2010), the present data demonstrate that, even if participants have no visual familiarity with one given human action, they are nevertheless able to recognize this action because of their motor knowledge and past auditory and somatosensory experience. This is in line with the assumption of sensory–motor transfer mechanisms at hand in the visual perception of audible but invisible tongue actions. The situation experienced by the participants of the present experiment is to a certain extent similar to the one experienced by newborns and 3-month-old infants, in the classical experiments on facial imitation by Meltzoff and Moore (1977, 1983). They have shown astonishing capacities to replicate to a certain extent a facial movement they have never seen done by a caregiver. These abilities are interpreted by the authors in reference to the link between proprioceptive and motor information feeding newborns with information about their own unseen movements in relation with the visual representation of the perceived movement of the caregiver and enabling the required action matching. Despite the correlational approach used in this study, our results suggest that, even if we have no visual but auditory and somatosensory experiences of an action, the connection between our motor abilities and the visual

incoming signal exists and enables adequate processing and performance.

UNCITED REFERENCES

Grabski, Schwartz, et al., 2013
Grabski, Tremblay, Gracco, Girin, & Sato, 2013

Acknowledgments

This study was supported by research grants from CNRS (Centre National de la Recherche Scientifique), from Agence Nationale de la Recherche (ANR SPIM, "Imitation in Speech: From Sensorimotor Integration to the Dynamics of Conversational Interaction"), and from the European Research Council (FP7/2007-2013 Grant agreement no. 339152, "Speech Unit(e)s"). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding agencies. We thank Jean-Luc Schwartz for helpful discussions.

Reprint requests should be sent to Avril Treille, GIPSA-lab, UMR 5216, Université Stendhal, 1180, Avenue Centrale, BP25, 38031 Grenoble Cedex 9, France, or via e-mail: avril.treille@gipsa-lab.grenoble-inp.fr.

REFERENCES

Aziz-Zadeh, L., Iacoboni, M., Zaidel, E., Wilson, S., & Mazziotta, J. (2004). Left hemisphere motor facilitation in response to manual action sounds. *European Journal of Neuroscience, 19,* 2609–2612.

Badin, P., Tarabalka, Y., Elisei, F., & Bailly, G. (2010). Can you "read" tongue movements? Evaluation of the contribution of tongue display to speech understanding. *Speech Communication, 52,* 493–503.

Beardsworth, T., & Buckner, T. (1981). The ability to recognize oneself from a video recording of one's movements without seeing one's body. *Bulletin of the Psychonomic Society, 18,* 19–22.

Beauchamp, M. S. (2005). Statistical criteria in fMRI studies of multisensory integration. *Neuroinformatics, 3,* 93–114.

Beauchamp, M. S., Argall, B. D., Bodurka, J., Duyn, J. H., & Martin, A. (2004). Unraveling multisensory integration: Patchy organization within human STS multisensory cortex. *Nature Neuroscience, 7,* 1190–1192.

Beauchamp, M. S., Lee, K. E., Argall, B. D., & Martin, A. (2004). Integration of auditory and visual informations about objects in superior temporal sulcus. *Neuron, 41,* 809–823.

Birn, R. M., Bandettini, P. A., Cox, R. W., & Shaker, R. (1999). Event-related fMRI of tasks involving brief motion. *Human Brain Mapping, 7,* 106–114.

Buccino, G., Lui, F., Canessa, N., Patteri, I., Lagravinese, G., Benuzzi, F., et al. (2004). Neural circuit involved in the recognition of actions performed by nonconspecifics: An fMRI study. *Journal of Cognitive Neurosciences, 16,* 114–126.

Callan, D. E., Jones, J. A., Munhall, K. G., Callan, A. M., Kroos, C., & Vatikiotis-Bateson, E. (2003). Neural processes underlying perceptual enhancement by visual speech gestures. *NeuroReport, 14,* 2213–2217.

Callan, D. E., Jones, J. A., Munhall, K. G., Callan, A. M., Kroos, C., & Vatikiotis-Bateson, E. (2004). Multisensory integration sites identified by perception of spatial wavelet filtered visual speech gesture information. *Journal of Cognitive Neuroscience, 16,* 805–816.

Calvert, G. A., Bullmore, E., Brammer, M. J., Campbell, R., Iversen, S. D., Woodruff, P., et al. (1997). Silent lip reading activates the auditory cortex. *Science, 276,* 593–596.

Calvert, G. A., & Campbell, R. (2003). Reading speech from still and moving faces: The neural substrates of visible speech. *Journal of Cognitive Neuroscience, 15,* 57–70.

Calvert, G. A., Campbell, R., & Brammer, M. J. (2000). Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. *Current Biology, 10,* 649–657.

Calvo-Merino, B., Glaser, D. E., Grèzes, J., Passingham, R. E., & Haggard, P. (2005). Action observation and acquired motor skills: An fMRI study with expert dancers. *Cerebral Cortex, 15,* 1243–1249.

Calvo-Merino, B., Grèzes, J., Glaser, D. E., Passingham, R. E., & Haggard, P. (2006). Seeing or doing? Influence of visual and motor familiarity in action observation. *Current Biology, 16,* 1905–1910.

Campbell, R., MacSweeney, M., Surguladze, S., Calvert, G., McGuire, P., Suckling, J., et al. (2001). Cortical substrates for the perception of face actions: An fMRI study of the specificity of activation for seen speech and for meaningless lower-face acts (gurning). *Cognitive Brain Research, 12,* 233–243.

Cappe, C., & Barone, P. (2005). Heteromodal connections supporting multisensory integration at low levels of cortical processing in the monkey. *European Journal of Neuroscience, 22,* 2886–2902.

d'Ausilio, A., Bartoli, E., Maffongelli, L., Berry, J. J., & Fadiga, L. (2014). Vision of tongue movements bias auditory speech perception. *Neuropsychologia, 63,* 85–91.

d'Ausilio, A., Bufalari, I., Salmas, P., & Fadiga, L. (2011). The role of the motor system in discriminating degraded speech sounds. *Cortex, 48,* 882–887.

d'Ausilio, A., Pulvermüller, F., Salmas, P., Bufalari, I., Begliomini, C., & Fadiga, L. (2009). The motor somatotopy of speech perception. *Current Biology, 19,* 381–385.

Di Pellegrino, G., Fadiga, L., Fogassi, L., Gallese, V., & Rizzolatti, G. (1992). Understanding motor events: A neurophysiological study. *Experimental Brain Research, 91,* 176–180.

Eckert, M. A., Kamdar, N. V., Chang, C. E., Beckmann, C. F., Greicius, M. D., & Menon, V. (2008). A crossmodal system linking primary auditory and visual cortices: Evidence from intrinsic fMRI connectivity analysis. *Human Brain Mapping, 29,* 848–885.

Eickhoff, S. B., Stephan, K. E., Mohlberg, H., Grefkes, C., Fink, G. R., Amunts, K., et al. (2005). A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data. *Neuroimage, 25,* 1325–1335.

Ferrari, P. F., Gallese, V., Rizzolatti, G., & Fogassi, L. (2003). Mirror neurons responding to the observation of ingestive and communicative mouth actions in the monkey ventral premotor cortex. *European Journal of Neuroscience, 17,* 1703–1714.

Fogassi, L., Ferrari, P. F., Gesierich, B., Rozzi, S., Chersi, F., & Rizzolatti, G. (2005). Parietal lobe: From action organization to intention understanding. *Science, 308,* 662–667.

Gallese, V., Fadiga, L., Fogassi, L., & Rizzolatti, G. (1996). Action recognition in the premotor cortex. *Brain, 119,* 593–609.

Grabski, K., Schwartz, J. L., Lamalle, L., Vilain, C., Vallée, N., Baciu, M., et al. (2013). Shared and distinct neural correlates of vowel perception and production. *Journal of Neurolinguistics, 26,* 384–408.

Grabski, K., Tremblay, P., Gracco, V., Girin, L., & Sato, M. (2013). A mediating role of the auditory dorsal pathway

in selective adaptation to speech: A state-dependent transcranial magnetic stimulation study. *Brain Research, 1515,* 55–65.
Hall, D. A., Haggard, M. P., Akeroyd, M. A., Palmer, A. R., Summerfield, A. Q., Elliott, M. R., et al. (1999). Sparse temporal sampling in auditory fMRI. *Human Brain Mapping, 7,* 213–223.
Haueisen, J., & Knösche, T. R. (2001). Involuntary motor activity in pianists evoked by music perception. *Journal of Cognitive Neuroscience, 13,* 786–792.
Howard, R. J., Brammer, M., Wright, I., Woodruff, P. W., Bullmore, E. T., & Zeki, S. (1996). A direct demonstration of functional specialization within motion-related visual and auditory cortex of the human brain. *Current Biology, 6,* 1015–1019.
Hueber, T., Chollet, G., Denby, B., & Stone, M. (2008). Acquisition of ultrasound, video and acoustic speech data for a silent-speech interface application. In *Proceedings of International Seminar on Speech Production (Strasbourg, France)* (pp. 365–369).
Johansson, R. (1973). Visual perception of biological motion and a model for its analysis. *Perception & Psychophysics, 14,* 201–211.
Jones, J., & Callan, D. E. (2003). Brain activity during audio-visual speech perception: An fMRI study of the McGurk effect. *NeuroReport, 14,* 1129–1133.
Katz, W. F., & Mehta, S. (2015). Visual feedback of tongue movements for novel speech sound learning. *Frontiers in Human Neuroscience, 9,* 612.
Keysers, C., Kholer, E., Umilta, M. A., Fogassi, L., Gallese, V., & Rizzolatti, G. (2003). Audiovisual mirror neurons and action recognition. *Experimental Brain Research, 153,* 628–636.
Kohler, E., Keysers, C., Umilta, M. A., Fogassi, L., Gallese, V., & Rizzolatti, G. (2002). Hearing sounds, understanding actions: Action representation in mirror neurons. *Science, 297,* 846–848.
Lahav, A., Saltzman, E., & Schlaug, G. (2007). Action representation of sound: Audiomotor recognition network while listening to newly acquired actions. *Journal of Neuroscience, 27,* 3008–3014.
Lancaster, J. L., Woldorff, M. G., Parsons, L. M., Liotti, M., Freitas, C. S., Rainey, L., et al. (2000). Automated Talairach atlas labels for functional brain mapping. *Human Brain Mapping, 10,* 120–131.
Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition, 21,* 1–36.
Loula, F., Prasad, S., Harber, K., & Shiffrar, M. (2005). Recognizing people from their movements. *Journal of Experimental Psychology: Human Perception and Performance, 31,* 210–220.
Meister, I. G., Wilson, S. M., Deblieck, C., Wu, A. D., & Iacoboni, M. (2007). The essential role of premotor cortex in speech perception. *Current Biology, 17,* 1692–1696.
Meltzoff, A. N., & Moore, M. K. (1977). Imitation of facial and manual gestures by human neonates. *Science, 198,* 75–78.
Meltzoff, A. N., & Moore, M. K. (1983). Newborn infants imitate adult facial gestures. *Child Development, 54,* 702–709.
Möttönen, R., & Watkins, K. E. (2009). Motor representations of articulators contribute to categorical perception of speech sounds. *Journal of Neuroscience, 29,* 9819–9825.
Ojanen, V., Möttönen, R., Pekkola, J., Jääskeläinen, I. P., Joensuu, R., Autti, T., et al. (2005). Processing of audio-visual speech in Broca's area. *Neuroimage, 25,* 333–338.
Oldfield, R. C. (1971). The assessment and analysis of handedness: The Edinburgh inventory. *Neuropsychologia, 9,* 97–114.
Pekkola, J., Laasonen, M., Ojanen, V., Autti, T., Jaaskelainen, I. P., Kujala, T., et al. (2006). Perception of matching and conflicting audio-visual speech in dyslexic and fluent readers: An fMRI study at 3T. *Neuroimage, 29,* 797–807.
Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences, 36,* 329–347.
Pizzamiglio, L., Aprile, T., Spitoni, G., Pitzalis, S., Bates, E., D'Amico, S., et al. (2005). Separate neural systems for processing action- or non-action related sounds. *Neuroimage, 24,* 852–861.
Prather, J. F., Peters, S., Nowicki, S., & Mooney, R. (2008). Precise auditory-vocal mirroring in neurons for learned vocal communication. *Nature, 451,* 305–310.
Pulvermuller, F., Huss, M., Kherif, F., Moscosodel Prado Martin, F., Hauk, O., & Shtyrov, Y. (2006). Motor cortex maps articulatory features of speech sounds. *Proceedings of the National Academy of Sciences, U.S.A., 103,* 7865–7870.
Rizzolatti, G., & Craighero, L. (2004). The mirror-neuron system. *Annual Review of Neuroscience, 27,* 169–192.
Rizzolatti, G., Fadiga, L., Gallese, V., & Fogassi, L. (1996). Premotor cortex and the recognition of motor actions. *Cognitive Brain Research, 3,* 131–142.
Rizzolatti, G., Fogassi, L., & Gallese, V. (2001). Neurophysiological mechanisms underlying the understanding and imitation of action. *Nature Review Neuroscience, 2,* 661–670.
Sato, M., Buccino, G., Gentilucci, M., & Cattaneo, L. (2010). On the tip of the tongue: Modulation of the primary motor cortex during audio-visual speech perception. *Speech Communication, 52,* 533–541.
Sato, M., Grabski, K., Glenberg, A., Brisebois, A., Basirat, A., Ménard, L., et al. (2011). Articulatory bias in speech categorization: Evidence from use-induced motor plasticity. *Cortex, 47,* 1001–1003.
Sato, M., Tremblay, P., & Gracco, V. (2009). A mediating role of the premotor cortex in phoneme segmentation. *Brain and Language, 111,* 1–7.
Saygin, A. P. (2007). Superior temporal and premotor brain areas necessary for biological motion perception. *Brain, 130,* 2452–2461.
Schwartz, J. L., Ménard, L., Basirat, A., & Sato, M. (2012). The Perception for Action Control Theory (PACT): A perceptuo-motor theory of speech perception. *Journal of Neurolinguistics, 25,* 336–354.
Skipper, J., Van Wassenhove, V., Nussman, H., & Small, S. L. (2007). Hearing lips and seeing voices: How cortical areas supporting speech production meditate audio-visual speech perception. *Cerebral Cortex, 17,* 2387–2399.
Skipper, J. I., Nusbaum, H. C., & Small, S. L. (2005). Listening to talking faces: Motor cortical activation during speech perception. *Neuroimage, 25,* 76–89.
Stevenson, R. A., Ghose, D., Krueger Fister, J., Sarko, D. K., Altieri, N. A., Nidiffer, A. R., et al. (2014). Identifying and quantifying multisensory integration: A tutorial review. *Brain Topography, 27,* 707–730.
Tai, Y. F., Scherfler, C., Brooks, D. J., Sawamoto, N., & Castiello, U. (2004). The human premotor cortex is "mirror" only for biological actions. *Current Biology, 14,* 117–120.
Venezia, J. H., Saberi, K., Chubb, C., & Hickok, G. (2012). Response bias modulates the speech motor system during syllable discrimination. *Frontiers in Psychology, 3,* 157.
Viviani, P., & Stucchi, N. (1992). Biological movements look uniform: Evidence of motor perceptual interactions. *Journal*

of Experimental Psychology: Human Perception and Performance, 18, 603–623.
Watkins, K. E., & Paus, T. (2004). Modulation of motor excitability during speech perception: The role of Broca's area. *Journal of Cognitive Neuroscience, 16,* 978–987.
Watkins, K. E., Strafella, A. P., & Paus, T. (2003). Seeing and hearing speech excites the motor system involved in speech production. *Neuropsychologia, 41,* 989–994.
Watkins, S., Shams, L., Tanaka, S., Haynes, J. D., & Rees, G. (2006). Sound alters activity in human V1 in association with illusory visual perception. *Neuroimage, 31,* 1247–1256.
Wilson, S., & Iacoboni, M. (2006). Neural responses to non-native phonemes varying in producibility: Evidence for the sensorimotor nature of speech perception. *Neuroimage, 33,* 316–325.
Wilson, S. M., Saygin, A. P., Sereno, M. I., & Iacoboni, M. (2004). Listening to speech activates motor areas involved in speech production. *Nature Neuroscience, 7,* 701–702.
Zaehle, T., Schmidt, C. F., Meyer, M., Baumann, S., Baltes, C., Boesiger, P., et al. (2007). Comparison of "silent" clustered and sparse temporal fMRI acquisitions in tonal and speech perception tasks. *Neuroimage, 37,* 1195–1204.

PARTIE EXPÉRIMENTALE - C

ETUDE EEG SUR LA PERCEPTION AUDIO-VISUELLE DE SES PROPRES MOUVEMENTS DE PAROLE

Dans ce chapitre, nous présenterons, sous la forme d'un article scientifique soumis à la revue *Experimental Brain Research* (février 2017), une étude en EEG sur la perception audio-visuelle de nos propres productions de parole (relative à l'état de l'art proposé dans la section D-3 de la partie théorique de cette thèse). Cette expérience a été réalisée au sein du laboratoire Gipsa-lab de Grenoble, sous la direction de Marc Sato et Coriandre Vilain et avec la collaboration de Sonia Kandel (Laboratoire de Psychologie et Neurocognition, Université Grenoble Alpes)

Treille, A., Vilain, C., Kandel, S. & Sato, M. (soumis). Seeing our own voice: an Electrophysiology study of audiovisual self speech perception. Experimental Brain Research.

Nous avons vu que la perception de nos propres actions pouvait être facilitée par rapport à celles d'un inconnu. Cependant, les résultats d'études précédentes apparaissent plus mitigés quant à une possible amélioration de la perception de stimuli de parole lorsque nous entendons notre propre voix et/ou regardons nos propres gestes articulatoires. De plus, les mécanismes neuronaux précoces qui sous-tendent ces processus n'ont à ce jour jamais été testés. Cette étude en EEG avait pour but d'examiner l'impact de nos connaissances de soi lors de la perception de syllabes auditives (A), visuelles (V) et audio-visuelles.

Nous nous sommes donc intéressés à la modulation des potentiels évoqués auditifs N1 et P2 lors de la présentation de stimuli relatifs aux productions du participant (préalablement enregistrés) ou aux productions d'une personne inconnue lors d'une tâche d'identification syllabique (/pa/, /ta/ et /ka/). Une condition audio-visuelle incongruente a été ajoutée, composée du signal sonore d'une syllabe issu des productions du participant et du signal visuel de cette même syllabe mais issu des productions d'un locuteur inconnu. Cette condition avait pour but de déterminer, en cas d'un possible effet du soi, s'il provenait des informations auditives ou visuelles.

Lors de cette étude, nous avons retrouvé une diminution de l'amplitude de PEA P2 en condition bimodale (AV) par rapport à la somme des signaux unimodaux (A+V), en accord avec la littérature portant sur les mécanismes d'intégration audio-visuelle de la parole. De plus, nous avons également démontré une diminution de la latence du PEA N1 lorsque le participant percevait ses propres productions visuelles.

Pour conclure, nous avons pu confirmer l'existence de mécanismes d'intégration précoce des informations auditives et visuelles des syllabes présentées ainsi qu'une facilitation temporelle, non visible au niveau comportemental, mais présente sur les signaux EEG lors de la perception visuelle de nos propres productions de parole.

Electrophysiological evidence for a self processing advantage during audiovisual speech integration

Avril Treille[1,CA], Coriandre Vilain[1], Sonia Kandel[1] and Marc Sato[2]

[1] GIPSA-lab, Département Parole & Cognition, CNRS & Grenoble Université, France
[2]Laboratoire Parole &Langage, CNRS & Aix-Marseille Université, France

[CA]Corresponding author:
Avril Treille
Gipsa-lab, Université Stendhal
1180, avenue Centrale BP25
38031 GRENOBLE CEDEX 9
Email: avril.treille@gipsa-lab.grenoble-inp.fr
Phone: +33 (0)4 76 82 41 28

ABSTRACT

Previous electrophysiological studies have provided strong evidence for early multisensory integrative mechanisms during audiovisual speech perception. From these studies, one unanswered issue is whether hearing our own voice and seeing our own articulatory gestures facilitate speech perception, possibly through a better processing and integration of sensory inputs with our own sensory-motor knowledge. The present EEG study examined the impact of self-knowledge during the perception of auditory (A), visual (V) and audiovisual (AV) speech stimuli that were previously recorded from the participant or from a speaker he/she had never met. Audiovisual interactions were estimated by comparing N1 and P2 auditory evoked potentials during the bimodal condition (AV) with the sum of those observed in the unimodal conditions (A+V). In line with previous EEG studies, our results revealed an amplitude decrease of P2 auditory evoked potentials in AV compared to A+V conditions. Crucially, a temporal facilitation of N1 responses was observed during the visual perception of self speech movements compared to those of another speaker. This facilitation was negatively correlated with the saliency of visual stimuli. These results provide evidence for a temporal facilitation of the integration of auditory and visual speech signals when the visual situation involves our own speech gestures.

Keywords: Self recognition, speech perception, audiovisual integration, EEG.

INTRODUCTION

Lip-reading alone is not enough to understand an utterance. However, adding information on the speaker's face improves speech perception. Several studies indicate that visual speech information enhances intelligibility in noisy environments (Sumby & Pollack, 1954; Benoît, Mohamadi & Kandel, 1994), facilitates phoneme identification of non-native phonemes (Navarra &Soto-Faraco, 2005; Burfin et al., 2014) or even contributes to the comprehension of complex content (Reisberg, McLean & Goldfield, 1987). In addition, in laboratory experimental situations, visual incongruent information (/ga/) when added to an auditory syllable (/ba/) can generate a new percept (/da/) different from both the auditory and visual syllables. This perceptual illusion was first displayed by McGurk and MacDonald in 1976 and strikingly underlines the complementarity and intimate interaction between auditory and visual speech information. Interestingly, visual information is not the only way to facilitate auditory speech decoding. Behavioral studies on tactile and audio-tactile speech perception also demonstrate that perceiving orofacial gestures of the speaker through the hand (via the TADOMA method; see Alcorn, 1932)can facilitate syllable discrimination(Reed et al., 1985, 1992; Reed et al., 1982; Fowler & Dekle, 1991; Gick et al., 2008; Sato et al., 2010; Treille et al., 2014a, 2014b).

At the brain level, electro-encephalographic (EEG) and magneto-encephalographic (MEG) studies demonstrate that N1/M1 and P2 auditory evoked potentials are attenuated and speeded up when an auditory syllable is combined with visual or tactile information from the speaker's face (Klucharev et al., 2003; Besle et al., 2004; van Wassenhove et al., 2005; Stekelenburg & Vroomen, 2007;Arnal et al., 2009; Pilling, 2010; Vroomen & Stekelenburg, 2010; Frtusova, Winneke & Phillips, 2013; Kaganovich & Schumaker, 2014;Treille et al., 2014a, 2014b; Baart et al., 2014; Baart & Samuel, 2015). This temporal facilitation and amplitude suppression of N1/M1 and P2 auditory evoked potentials is thought to reflect early multisensory integrative mechanisms through visual predictions of the incoming auditory events. However, the speech specific nature of these effects

remains controversial. Indeed, Stekelenburg and Vroomen (2007) and Vroomen and Stekelenburg (2010) observed similar N1 latency and amplitude decreases during the observation of biological transitive (shock of a spoon against a cup) and intransitive (handclapping) non-speech actions, and even during the observation of non-biological actions(a pure tone synchronized with a deformation of a rectangle, or a collision of moving disks).These studies suggested that N1 and P2modulations would reflect different aspects of audiovisual integration mechanisms (van Wassenhove et al., 2005; Arnal et al., 2009; Baart et al., 2014). There would be a nonspeech-specific stage in audiovisual integration that processes the early arrival of visual information. This would be reflected byN1 latency and amplitude modulations. A subsequent speech-specific featural phonetic stage would be reflected in P2 modulations.

Neuroimaging studies further demonstrate the existence of specific brain areas playing a key role in the audiovisual integration of speech. In particular, audiovisual speech perception has an impact on the activity of unisensory visual and auditory regions (the visual motion-sensitive cortex, V5/MT, and the Heschl's gyrus) as well as multisensory regions (the posterior part of the left superior temporal gyrus/sulcus, pSTS/pSTG), when compared to auditory and visual unimodal conditions(Calvert, Campbell and Brammer, 2000; Callan et al., 2003, 2004; Skipper et al., 2005, 2007). Interestingly, the premotor cortex-that is involved in speech production and is part of the dorsal stream (Hickok & Poeppel, 2007)- might also play a role in audiovisual speech integration mechanisms. Indeed, previous studies on audiovisual speech perception demonstrated stronger activation of this premotor region during the presentation of bimodal speech stimuli compared to auditory and visual only conditions (Campbell et al., 2001; Calvert & Campbell, 2003; Watkins, Strafella & Paus, 2003; Watkins & Paus, 2004; Skipper et al., 2005, 2007; Sato et al., 2010). This occurred during the presentation of incongruent stimuli compared to congruent ones (Jones &Callan, 2003; Ojanen et al., 2005; Pekkola et al., 2006) and also in the case of degraded visual or auditory speech signals (Callan et al., 2003, 2004). Although the debate is still open, the latter support the idea that the motor knowledge we use to produce speech sounds might constrain phonetic decoding of the sensory inputs. This comforts, to a certain extent, the motor and sensorimotor theories of speech perception and language comprehension (Liberman & Mattingly, 1985; Skipper et al., 2007; Schwartz et al., 2012; Pickering & Garrod, 2013) and support the long-standing proposal that perception and action are two closely linked processes.

From these studies on audiovisual speech perception, one intriguing question is whether hearing our own voice and seeing our own articulatory gestures facilitate speech perception, possibly through a better processing and integration of sensory inputs with our own sensory-motor knowledge. From this question, a few behavioral studies have provided contrasted results. Tye-Murray and colleagues (2012, 2014) demonstrated that, during sentence lip-reading, participants recognize better their visual productions than those of others. In contrast, Aruffo and Shore (2012) found a self-auditory but not a self-visual advantage during the presentation of incongruent audiovisual speech stimuli. Other behavioral studies attempted to show a self-processing effect during audiovisual syllable perception, but the results were not concluding (Schwartz and Savariaux, 2001).

The present study examined whether self-information processing constitutes an advantage during audiovisual speech integration. We used EEG to examine N1 and P2 auditory evoked potentials during the perception of auditory and/or visual speech stimuli that were previously recorded from the participant (self) and a speaker he/she had never met (other). For each participant, eight conditions were tested, consisting on four distinct modalities: an auditory modality (A_{self}, A_{other}), a visual modality (V_{self}, V_{other}), an audiovisual modality ($A_{self}V_{self}$, $A_{other}V_{other}$) and an audiovisual modality with incongruent speakers in which the acoustic and visual signals were produced by the participant and the other speaker respectively ($A_{self}V_{other}$, $A_{other}V_{self}$).The audiovisual modality with incongruent speakers was designed to determine whether a possible self-effect comes from auditory or visual information. Using an additive model, we tested whether N1/P2 auditory evoked potentials were attenuated and speeded up during audiovisual conditions compared to the sum of those observed in unimodal conditions, and whether these effects were modulated by a self-processing advantage.

METHOD

Participants

Eighteen healthy adults participated in the study (12 females; mean age 23, SD ± 5 years). All the participants were right-handed native French speakers, had normal or corrected-to-normal vision and reported no history of speaking, hearing or motor disorders. They gave written consent for their participation in the study. They were compensated for the time spent in the study. The study received approval by the Grenoble Alpes University Ethical Committee (CERNI, N°2013-12-24-33).

Stimuli

Recording–We recorded 10 utterances of /apa/, /ata/ and /aka/ sequences of each participant in a soundproof room. Previous research on audiovisual speech perception has shown that these sequences correspond to a gradient of visuo-labial saliency: the unvoiced bilabial /p/ stop consonant is more salient visually than unvoiced alveolar stop consonant /t/ and in turn stop consonant velar /k/ unvoiced (e.g., van Wassenhove et al., 2005 for an EEG study). Moreover, these stop consonants have precise acoustics onsets, which is crucial for the EEG analyses we intended to carry out (see below). Then, we selected four utterances of each sequence for each participant on the basis of visual and acoustical durations (using Adobe Premiere, Adobe Systems, and Praat software; Boersma & Weenink, 2013).

Stimulus preparation–The movies were created on the basis of 30 frames (1200 ms) before the acoustic burst and 5 frames (200 ms) after it, for a total duration of 1400 ms for all the stimuli. Prior to generating movies, we extracted the acoustic signal and erased the first vowel /a/ so that all the audio signals began with a 1200 ms silence. Then, we merged the audio and video signals in four different types of movies:

-Auditory modality (A): the movie consisted of a fixed image of the last frame before the acoustic onset during the initial vowel /a/ dubbed on the acoustic signal.

-Visual modality (V): The movie consisted of the visual input without the sound.

-Audiovisual modality (AV): The movie consisted of both the auditory and visual signals.

-Audiovisual modality with incongruent speakers (AV incongruent speakers): The movie consisted of the acoustic signal of the speaker dubbed with the visual signal of the same syllable but produced by another participant (see below for the matching method).

Participant pair matching – Because of possible idiosyncrasy or production differences between participants that might cause facilitation of visual or auditory stimuli recognition, each participant was associated to an unknown participant of the same gender.

Our experiment therefore consisted of 9 pairs of participants. To each participant we presented both her/his own productions and those of her/his unknown partner (see Figure 1).For each participant, eight conditions were tested, consisting on four distinct modalities applied either on the participant her/himself (self) or the unknown speaker(other): an auditory modality (A_{self}, A_{other}), a visual modality (V_{self}, V_{other}), an audiovisual modality ($A_{self}V_{self}$, $A_{other}V_{other}$) and an audiovisual modality with incongruent speakers in which the acoustic and visual signals were produced by the participant and the other speaker ($A_{self}V_{other}$, $A_{other}V_{self}$). The audiovisual modality with incongruent speakers was designed to determine whether a possible self-effect comes from auditory or visual information. With this procedure, a total of 864 stimuli were created (18 speakers x 4 modalities x 3 syllables x 4 utterances).

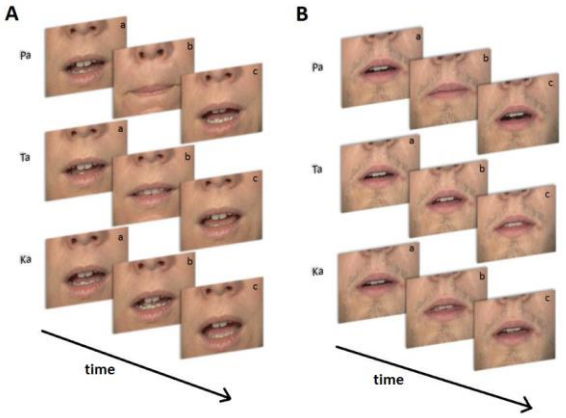

Figure 1: Examples of the visual stimuli for two participants (A,B). Each utterance begins with the mouth open (a); is followed by the stop consonant (b); and ends with the second /a/ vowel (c).

Experimental procedure

The participants sat in front of a computer monitor at a distance of approximately 50 cm. The acoustic stimuli were presented at a comfortable sound level through loudspeakers, with the same sound level set for all participants. The software *Presentation* (Neurobehavioral Systems, Albany, CA) controlled stimulus presentation and recorded the participants' responses. The participants were instructed to identify the syllable presented by the

movies by pressing a key on the keyboard with their left hand. It was a three-alternative /pa/, /ta/ and /ka/ forced-choice identification task. In order to dissociate sensory/perceptual responses from motor responses on EEG data, a brief single audio beep was delivered 600 ms after the end of each stimulus. The participants had to respond after this audio beep. The experiment consisted of576 trials presented in a pseudo-randomized sequence, with 24 trials in each condition (4 modalities (A, V, AV, AV with incongruent speakers) x 2 speakers (self and other) x 3 syllables (/pa/, /ta/ and /ka/) x 24 trials). The inter-trial interval was set at3 s and the response key designation was fully counterbalanced across participants.

EEG acquisition and processing

EEG data were recorded continuously from 64 scalp electrodes (Electro-Cap International, INC, according to the international 10-20 system) using the Biosemi Active Two AD-box EEG system operating at a 256 Hz sampling rate. Two additional electrodes served as reference (Common Mode Sens [CMS] active electrode) and ground (Driven Right Leg [DRL] passive electrode). One other external reference electrode was set at the top of the nose. Horizontal (HEOG) and vertical (VEOG) eye movements were recorded using an electro-oculogram with electrodes positioned at the outer canthus of each eye as well as above and below the right eye. Before the experiment, the impedance of all electrodes was adjusted to get low offset voltages and stable DC.

EEG data were processed using the EEGLAB toolbox (Delorme and Makeig, 2004) running on Matlab (Mathworks, Natick, MA, USA). Since N1/P2 auditory evoked potentials have maximal response over central sites on the scalp (Scherg and Von Cramon, 1986; Näätänen and Picon, 1987), EEG data preprocessing and analyses were conducted on 6 representative fronto-central electrodes (F3, Fz, F4, C3, Cz, C4). This is in line with previous EEG studies on audiovisual speech perception and auditory evoked potentials (e.g. Pilling, 2010; Stekelenburg and Vroomen, 2007; van Wassenhove et al., 2005; Vroomen and Stekelenburg, 2010). A topographic analysis conducted on all the participants and 64 electrodes demonstrated a maximal response of N1/P2 auditory evoked potentials on fronto-central electrodes (see Figure 2). This confirmed the reliability of our selection of fronto-central electrodes. EEG data were first off-line re-referenced to the nose recording and band-pass filtered using a two-way least-square FIR filtering (1-20 Hz). Data were then segmented into 1000 ms epochs including a 100 ms pre-stimulus baseline (from -500 ms to -400 ms relative to the acoustic syllable onset). Epochs with an amplitude change exceeding ± 60 uV at any channel (including HEOG and VEOG channels) were rejected (on average, less than 6%). For each participant and condition (A_{self}, A_{other}, V_{self}, V_{other}, $A_{self}V_{self}$, $A_{other}V_{other}$, $A_{self}V_{other}$, $A_{other}V_{self}$), the data were averaged on the 6 electrodes. Then the maximal amplitude and peak latency of auditory N1 and P2 evoked responses were determined on the EEG waveform using a fixed window (N1: 70-150 ms; P2: 150-250 ms).

Data analyses

Behavioral analyses

The percentage of correct responses was determined for each participant, syllable and modality. We conducted a three-way repeated-measures ANOVAs with speaker type (self *vs.* other), modality (A, V, AV, AV with incongruent speakers) and the syllable (/pa/, /ta/ and /ka/) as within-participants variables.

EEG analyses

Audiovisual integration –To test audiovisual speech integration, we used an additive model, with EEG responses in the bimodal conditions (AV) compared to the sum of auditory and visual EEG responses (A+V).We conducted three-way repeated-measures ANOVAs on N1/P2 amplitudes and latencies with signal type (bimodal ***vs.*** sum), auditory speaker (self *vs*. other) and visual speaker (self, other or none) as within-participants factors.

Correlation between accuracy and EEG signals– To test the relation between the perceptual visual saliency and degree of integration observed on the EEG signals, we conducted Pearson correlation analyses. The analyses concerned the relation between visual accuracy and the modulations of either N1/P2 amplitude or latency. They were related to the difference between the bimodal conditions and the sum of unimodal conditions (e.g., EEG responses on $[A_{self}V_{self} – (A_{self} + V_{self})]$and $[A_{other}V_{self} – (A_{other} + V_{self})]$correlated with V_{self} scores, or EEG responses on $[A_{self}V_{other} – (A_{self} + V_{other})]$and $[A_{other}V_{other} – (A_{other} + V_{other})]$correlated with V_{other} scores).

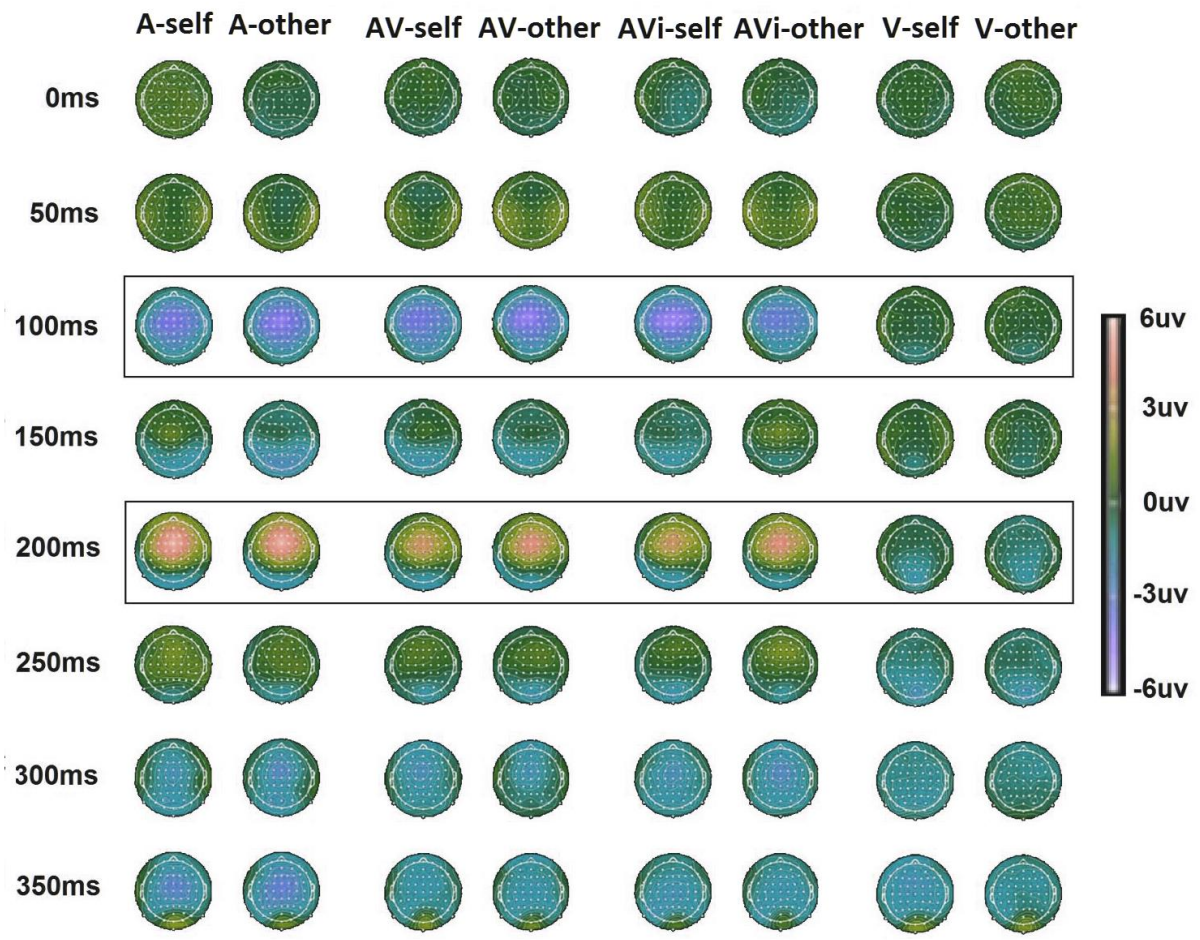

Figure 2: Topographic analysis conducted on all the participants and electrodes demonstrating a maximal response of N1/P2 auditory evoked potentials on fronto-central electrodes.

RESULTS

Accuracy

Overall, the mean proportion of correct responses was 94% (see Figure 3). The analyses revealed a main effect of presentation modality ($F(3,51)=67.6$; $p<.0001$). The percentages of correct responses for the visual stimuli (83%) were lower than for auditory (A: 98%) and audiovisual stimuli (AV: 99%; AVi: 98%). In addition, consonant saliency also yielded a main effect ($F(2,34)=23.3$; $p<.0001$). The /pa/syllables were identified better (98%) than the /ta/ (92%) and in turn /ka/ (93%) ones. Finally, the interaction between the presentation modality and the syllable was reliable ($F(6,102)=24.1$; $p<.0001$). There was an effect of syllable saliency in the visual modality (V-/pa/: 99%; V-/ta/: 75%; V-/ka/: 74%).

EEG results

Amplitude–None of the effects reached significance for N1 amplitude. There was a main effect of signal type for P2 amplitude ($F(1,16)=6.9$; $p<.02$; see Figure 4). The amplitude was smaller for the bimodal conditions (3.8 µV) than the sum of the auditory and visual signals (4.7 µV).

Latency– Regarding the analyses on N1 latency, there was a significant main effect of the visual speaker ($F(1,16)=8.2$; $p<.02$; see Figure 4). There was a temporal facilitation during the perception of visual-self speech movements (107 ms) compared to visual-other speech movements (113 ms). No significant effects were found for P2 latency.

Correlation between behavioral scores and EEG signals

Amplitude - No significant correlation was found between EEG signals related to AV integration and the visual saliency of syllables for both N1 and P2 amplitude (N1: self : $r=.09$; $F(1,32)=0.2$; $p<.63$; other: $r=.24$; $F(1,32)=2.0$; $p<.16$; P2: self: $r=.22$; $F(1,32)=1.6$; $p<.22$; other: $r=.18$; $F(1,32)=1.1$; $p<.30$; see Figure 5).

Latency - N1 latency difference between AV and A+V EEG responses related to the visual-self syllables was negatively correlated with the visual recognition scores (V-self: r=.41; F(1,32)=6.5; p<.02). No significant correlation was observed for the visual syllables from an unknown speaker (V-other: r=.01; F(1,32)=0; p<.94). Finally, no significant correlation was observed between P2 latency data related to the degree of integration of self and other visual information and visual accuracy (V-self: r=.11; F(1,32)=0.32; p<.54; V-other: r=.29; F(1,32)=2.95; p<.10; see Figure 5).

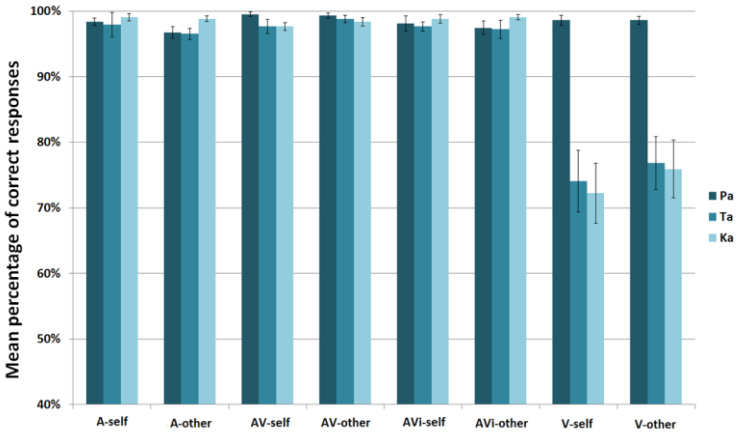

Figure 3: Mean percentage of correct responses observed for each speaker type, presentation modality and each syllable.

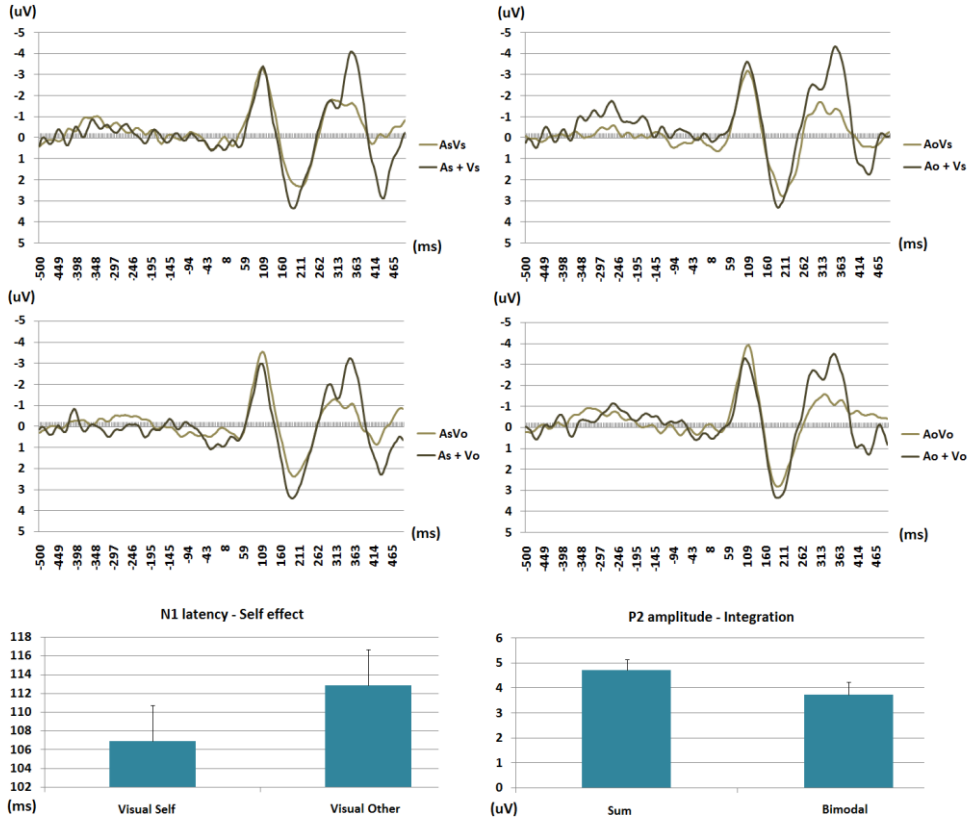

Figure 4: Top: Averaged event-related potentials on fronto-central electrodes related to the audiovisual conditions (AV) and the sum of unimodal conditions (A+V) according to the auditory and the visual speakers (s: self; o: other). Bottom-left: Latency reduction of N1 observed in audiovisual conditions for self compared to other visual movements. Bottom-right: Amplitude reduction of P2 observed for AV compared to A+V.

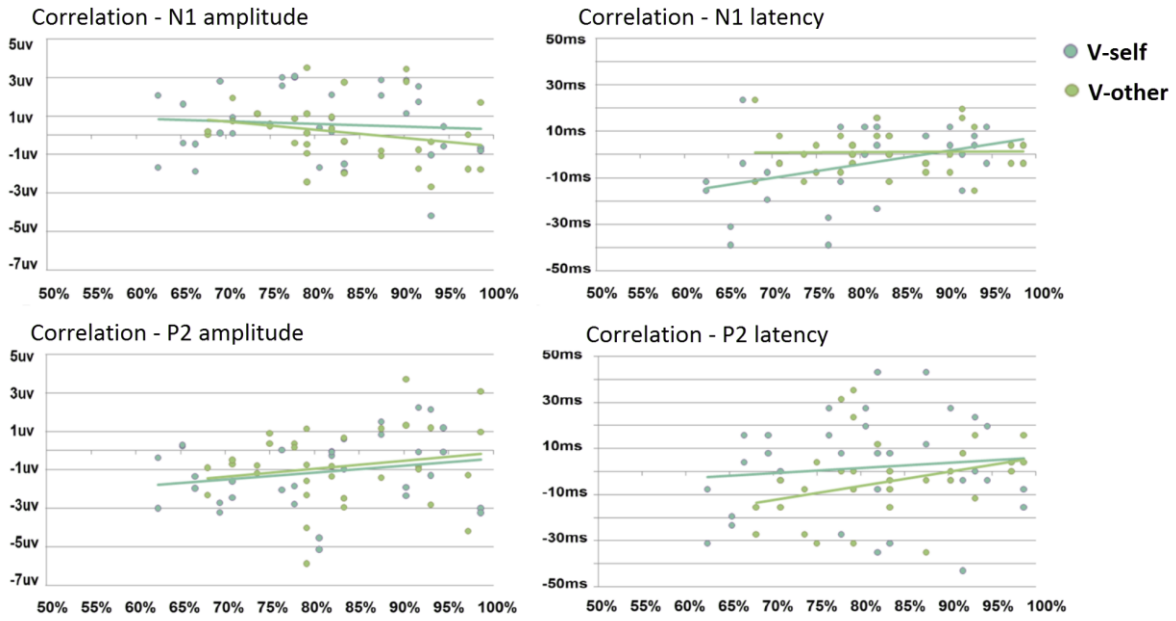

Figure 5: Correlation between the visual recognition scores for self and other visual movements (x-axis) and the difference in amplitude and latency of N1 and P2 auditory evoked potentials between AV and A+V (y-axis).

DISCUSSION

The present EEG study investigated a possible self-processing advantage during speech perception, and its related impact on audiovisual integration mechanisms. Two main results were observed. First and in line with previous EEG studies on audiovisual speech integration, we observed an amplitude decrease on P2 auditory evoked potentials during the bimodal presentation compared to the sum of auditory and visual unimodal responses. Crucially, during audiovisual speech integration, a temporal facilitation related to self lip movements was observed on N1 auditory evoked potentials, a facilitation that appears negatively correlated with the saliency of visual stimuli.

Previous studies on audiovisual speech integration demonstrated that bimodal presentations produce a decrease in N1 and/or P2 latency and amplitudes (Besle et al., 2004; van Wassenhove et al., 2005; Stekelenburg and Vroomen, 2007; Pilling, 2010; Baart et al., 2014; Treille et al., 2014a, 2014b) and latency (van Wassenhove et al., 2005; Stekelenburg et Vroomen, 2007; Baart et al., 2014; Treille et al., 2014a; see also Arnal et al., 2009 for similar results with MEG) when compared to auditory responses or to the sum of auditory and visual responses. These modulations of the N1/P2 responses are thought to reflect specific stages of audiovisual speech integration. N1 latency and amplitude modulations would reflect a non speech-specific stage while P2 latency shifts or amplitude decreases would rather be speech-specific and related to a featural phonetic stage. Using an additive model, our results revealed a P2 amplitude decrease during the bimodal presentation compared to the sum of the unimodal auditory and visual conditions. In line with previous studies (van Wassenhove et al., 2005; Stekelenburg and Vroomen, 2007; Pilling, 2010; Baart et al., 2014; Treille et al., 2014b), this result suggests that visual speech information affects ongoing auditory activity and further demonstrates the integration of auditory and visual speech signals. However, there were no differences on P2 latency, nor N1 amplitude and latency. This contrasts with previous studies reporting latency shifts of auditory evoked responses and/or N1 amplitude decreases in the bimodal condition. Some aspects of the present experimental procedure might explain these differences. A first important point is related to the stimulus variability. In our experiment we presented four tokens of three syllables produced by two speakers. The above mentioned studies only presented one token of each presented syllable (van Wassenhove et al., 2005; Stekelenburg and Vroomen, 2007; Arnal et al., 2009; Baart et al., 2014) and/or a more limited number of syllables (i.e., one or two; Stekelenburg and Vroomen, 2007; Pilling, 2010; Vroomen and Stekelenburg, 2010; Baart et al., 2014; Treille et al., 2014a). In the present EEG experiment, the higher stimulus variability might have decreased eventual habituation/learning effects. This might have limited latency shifts on auditory evoked potentials. From that view, a recent meta-analysis suggests that variability across EEG/MEG studies on audiovisual speech integration may potentially be driven by many experimental, procedural, and methodological

differences, such as the number and quality of stimuli, the sound intensity, the inter-trial interval, the task, the degree of selective attention, the preprocessing and the analysis of the data (Baart, 2016).

It is noteworthy that our behavioral results did not reveal any visual, auditory or audiovisual self-processing advantage. This contrasts with a behavioral study conducted by Tye-Murray and colleagues (2012). They showed that we lip-read more accurately sentences produced by ourselves than by other speakers. For the authors, these results provide support to the common coding theory (Prinz, 1997; Hommel et al., 2001), which posits that producing and perceiving share the same representations of motor plans. Because of this perceptuo-motor coupling, observing one's own action activates these motor plans to a greater extent than observing someone else's action. A reason for this divergence could reside on stimulus length. In the present study we used syllables whereas Tye-Murray et al used sentences. The use of short CV syllables therefore limited the quantity of visual information and facilitated correct responses (mean 94%). Our results appear consistent however with the study by Aruffo and colleagues (2012) who did not find any visual self-processing advantage with participants presented with incongruent audiovisual syllables (McGurk stimuli), although self-voice appeared to weaken the illusion effect.

The major contribution of our EEG study is that it provides evidence fora visual self-processing advantage on N1 latency during audiovisual speech perception. More specifically, a temporal facilitation of audiovisual speech processing was observed when participants watched their own productions compared to those of another speaker. This facilitation was negatively correlated with the saliency of visual self-stimuli. This suggests that the visual self-processing effect is linked to specific visual speech features of the presented syllables, like the place of articulation of the consonants (with their acoustic bursts here used as onsets for EEG analyses). Interestingly, this effect seems to be largely driven by visually "ambiguous" syllables (see Figure 5). Although this correlational result precludes any causal inferences, a plausible explanation could be that the difficulty to decode our own speech gestures would increase the degree of audiovisual integration and temporally facilitate auditory process.

In conclusion, the present EEG study provides the first electrophysiological evidence for a self-processing advantage during audiovisual speech integration. The observed temporal facilitation of N1 responses during the visual perception of self speech movements compared to those of another speaker suggest that perceiving our own articulatory gestures speed up auditory speech perception, possibly through a better processing and integration of sensory inputs with our own sensory-motor knowledge.

ACKNOWLEDGEMENTS

This study was supported by research funds from the European Research Council (FP7/2007-2013 Grant Agreement no. 339152).

REFERENCES

Alcorn, S. (1932).The Tadoma method. Volta Review, 34:195-198

Arnal, L.H., Morillon, B., Kell, C.A. & Giraud, A.L. (2009).Dual neural routing of visual facilitation in speech processing. The Journal of Neuroscience, 29(43):13445-13453

Aruffo, C., & Shore, D.I. (2012). Can you McGurk yourself? Self-face and self-voice in audiovisual speech. Psychonomic Bulletin & Review, 19:66–72

Baart, M., Stekelenburg, J. J. & Vroomen, J. (2014). Electrophysiological evidence for speech-specific audiovisual integration. *Neuropsychologia*, 65:115–211

Baart, M., & Samuel, A. G. (2015).Turning a blind eye to the lexicon: ERPs show no cross-talk between lip-read and lexical context during speech sound processing. Journal of Memory and Language, 85:42–59

Baart, M. (2016). Quantifying lip-read induced supression and facilitation of the auditory N1 and P2 reveals peak enhancements and delays. Psychophysiology, 53(9):1295-306

Benoît,C., Mohamadi, T. & Kandel, S. (1994). Effects of phonetic context on audio-visual intelligibility of French speech in noise. Journal Speech Hearing Research, 37:1195–1203

Besle, J., Fort, A., Delpuech, C. &Giard, M.H. (2004). Bimodal speech: early suppressive visual effects in human auditory cortex. European journal of Neuroscience, 20:2225-2234

Boersma, P. & Weenink,D.(2013). Praat: doing phonetics by computer. Computer program, Version 5.3.42, retrieved 2 March 2013 from ⟨http://www.praat.org/⟩

Burfin, S., Pascalis, O., Ruiz Tada, E., Costa, A., Savariaux, C. & Kandel S. (2014). Bilingualism affects the audio-visual processing of non-native phonemes. Frontiers in Psychology (Research Topic "New advances on the perception and production of non-native speech sounds" – Section Language Sciences), 5: 1179

Callan, D.E., Jones, J.A., Munhall, K.G., Callan, A.M., Kroos, C. & Vatikiotis-Bateson, E. (2003).Neural processes underlying perceptual enhancement by visual speech gestures. Neuro Report, 14:2213-2217

Callan, D.E., Jones, J.A., Munhall, K.G., Callan, A.M., Kroos, C. & Vatikiotis-Bateson, E. (2004). Multisensory integration sites identified by perception of spatial wavelet filtered visual speech gesture information. Journal of Cognitive Neuroscience, 16:805-816

Calvert, G.A. & Campbell, R. (2003). Reading speech from still and moving faces: the neural substrates of visible speech. Journal of Cognitive Neuroscience, 15:57–70

Calvert, G.A., Campbell, R. & Brammer, M.J. (2000). Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. Current Biology, 10(11):649-657

Campbell, R., MacSweeney, M., Surguladze, S., Calvert, G., McGuire, P., Suckling, J., Brammer, M.J. & David, A.S. (2001). Cortical substrates for the perception of face actions: an fMRI study of the specificity of activation for seen speech and for meaningless lower-face acts (gurning).Cognitive Brain Research, 12:233-243

D'Ausilio A., Pulvermüller F., Salmas P., Bufalari I., Begliomini C., Fadiga L. (2009) The motor somatotopy of speech perception. Current Biology, 19:381-385

D'Ausilio A., Bufalari I., Salmas P., Fadiga L., (2012) The role of the motor system in discriminating degraded speech sounds. Cortex, 48(7):882–887

Fowler, C. & Dekle, D. (1991).Listening with eye and hand: crossmodal contributions to speech perception. Journal of Experimental Psychology- Human Perception and Performance, 17:816–828

Frtusova, J. B., Winneke, A. H., & Phillips, N. A. (2013). ERP evidence that auditory–visual speech facilitates working memory in younger and older adults. Psychology and Aging, 28(2), 481–494

Gick, B., Jóhannsdóttir, K.M., Gibraiel, D. & Mühlbauer, M. (2008).Tactile enhancement of auditory and visual speech perception in untrained perceivers. Journal of Acoustical Society of America, 123:72-76

Grabski, K., Tremblay, P., Gracco, V., Girin, L. & Sato, M. (2013). A mediating role of the auditory dorsal pathway in selective adaptation to speech: a state-dependent transcranial magnetic stimulation study. Brain Research, 1515: 55-65

Hickok, G. & Poeppel, D. (2007).The cortical organization of speech processing. Nature Review Neurosciences, 8:393-402

Hommel, B., Musseler, J., Aschersleben, G. & Prinz, W. (2001).The theory of event oding (TEC): A framework for perception and action planning. Behavioral and Brain Sciences, 24:849–878

Jones, J.A. & Callan, D.E (2003). Brain activity during audiovisual speech perception: an fMRI study of the McGurk effect. Neuro report, 14:1129–1133

Kaganovich, N., & Schumaker, J. (2014). Audiovisual integration for speech during mid-childhood: Electrophysiological evidence. Brain and Language, 139:36–48

Klucharev, V., Möttönen, R. & Sams, M. (2003). Electrophysiological indicators of phonetic and non-phonetic multisensory interactions during audiovisual speech perception. Brain Res. Cogn. Brain Res., 18:65-75

Liberman, A.M. & Mattingly, I.G (1985). The motor theory of speech perception revised. Cognition, 21:1-36

McGurk, H. & MacDonald, J. (1976). Hearing lips and seeing voices. Nature, 264:746-748

Meister, I.G., Wilson, S.M., Deblieck, C., Wu, A.D., & Iacoboni, M. (2007).The essential role of premotor cortex in speech perception. Current Biology, 17(19):1692–1696

Möttönen, R. & Watkins, K.E.(2009). Motor representations of articulators contribute to categorical perception of speech sounds. J. Neuroscience, 29(31):9819-9825

Näätänen, R. & Picton, T.W. (1987). The N1 wave of the human electric and magnetic response to sound: a review and an analysis of the component structure. Psychophysiology, 24:375–425

Ojanen, V., Möttönen, R., Pekkola, J., Jääskeläinen, I.P., Joensuu, R.,Autti, T. & Sams, M. (2005).Processing of audiovisual speech in Broca'sarea.NeuroImage,25:333–338

Oldfield, R.C. (1971). The Assessment and analysis of handedness: The Edinburgh inventory. Neuropsychologia, 9:97-113

Pekkola, J., Laasonen, M., Ojanen, V., Autti, T., Jääskeläinen, L.P., Kujala, T. & Sams, M. (2006).Perception of matching and conflicting audiovisual speech in dyslexic and fluent readers: an fMRI study at 3T.NeuroImage, 29(3):797–807

Pickering, M.J. & Garrod, S. (2013). An integrated theory of language production and comprehension. Behav. Brain Sci., 36:329–347

Pilling, M. (2010).Auditory event-related potentials (ERPs) in audiovisual speech perception. Journal of Speech, Language, and Hearing Research, 52(4):1073-1081

Prinz, W. (1997).Perception and action planning. European Journal of Cognitive Psychology, 9:129–154

Pulvermuller, F., Huss, M., Kherif, F., Moscosodel Prado Martin, F., Hauk, O., & Shtyrov, Y. (2006). Motor cortex maps articulatory features of speech sounds. Proceedings of the National Academy of Sciences of the United States of America, 103(20):7865–7870

Reed, C.M., Rabinowitz, W.M., Durlach, N.I., Braida, L.D., Conway-Fithian, S. & Schultz, M.C. (1985). Research on the Tadoma method of speech communication. J Acoust Soc. Am., 77(1):247-257

Reed, C.-M., Rabinowitz, W.-M., Durlach, N.-I., Delhorne, L.-A., Braida, L.-D., Pemberton, J.-C., Mulcahey, B.-D. & Washington, D.-L. (1992). Analytic study of the Tadoma method: Improving performance through the use of supplementary tactual displays. Journal of Speech and hearing Research, 35:450-465

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 LipReading, 97-114

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

Sato, M., Tremblay, P. & Gracco, V. (2009).A mediating role of the premotor cortex in phoneme segmentation. Brain and Language, 111(1):1-7

Saygin, A.P. (2007). Superior temporal and premotor brain areas necessary for biological motion perception. Brain, 130:2452-2461

Scherg, M., and VonCramon, D. (1986). Evoked dipole source potentials of the human auditory cortex. Electroencephalogr. Clin. Neurol., 65:344–360

Schwartz, J.L. & Savariaux C. (2001).Is it Easier to Lipread One's Own Speech Gestures Than Those of Somebody Else? It Seems Not! Auditory-Visual Speech Processing, Aalborg, Denmark, 18-23

Schwartz, J.L., Ménard, L., Basirat, A. & Sato, M. (2012).The Perception for Action Control Theory (PACT): a perceptuo-motor theory of speech perception. Journal of Neurolinguistics, 25(5):336-354

Skipper, J., Van Wassenhove, V, Nussman, H. & Small, S. (2007). Hearing lips and seeing voices: How cortical areas supporting speech production meditate audiovisual speech perception. Cerebral Cortex, 17:2387-2399

Skipper, J.I., Nusbaum, H.C. & Small, S.L. (2005).Listening to talking faces: motor cortical activation during speech perception. NeuroImage, 25:76–89

Stekelenburg, J.J. & Vroomen, J. (2007). Neural correlates of multisensory integration of ecologically valid audiovisual events. Journal of Cognitive Neuroscience, 19:1964–1973

Sumby, W.H. & Pollack, I. (1954).Visual contribution to speech intelligibility in noise. The Journal of the Acoustical Society of America, 26:212-215

Treille, A., Cordeboeuf, C., Vilain, C. & Sato, M. (2014a). Haptic and visual information speed up the neural processing of auditory speech in live dyadic interactions. Neuropsychologia, 57:71-77

Treille, A., Vilain, C. & Sato, M. (2014b). The sound of your lips: electrophysiological cross-modal interactions during hand-to-face and face-to-face speech perception. Frontiers in Psychology, 5(420):1-9

Tye-Murray, N., Spehar B., Myerson, J., Hale, S. & Sommers, M.S. (2012).Reading your own lips: Common coding theory and visual speech perception. Psychonomic Bulletin & Review, 20:115-119

Tye-Murray N., Hale S., Spehar B., Myerson J., & Sommers M. (2014). Lipreading in school-age children: The roles of age, hearing status, and cognitive ability. Journal of Speech, Language, and Hearing Research, 57:556–565

van Wassenhove, V., Grant, K.W. & Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. Proceedings of the National Academy of Sciences U.S.A., 102:1181-1186

Vroomen, J. & Stekelenburg, J.J. (2010). Visual anticipatory information modulates multisensory interactions of artificial audiovisual stimuli. Journal of Cognitive Neuroscience, 22:1583-1596

Watkins, K.E., Strafella, A.P. & Paus, T. (2003). Seeing and hearing speech excites the motor system involved in speech

production. Neuropsychologia, 41:989-994

Watkins, K.E. & Paus, T. (2004). Modulation of motor excitability during speech perception: the role of Broca's area. Journal of Cognitive Neuroscience, 16(6):978-987

Wilson, S. & Iacoboni, M. (2006).Neural responses to non-native phonemes varying in producibility: Evidence for the sensorimotor nature of speech perception. NeuroImage, 33:316-325

Wilson, S.M., Saygin, A.P., Sereno, M.I. & Iacoboni, M., (2004). Listening to speech activates motor areas involved in speech production. Nat.Neurosci., 7:701–7

PARTIE EXPÉRIMENTALE - D

ETUDE TMS SUR LA PERCEPTION MULTISENSORIELLE DE LA PAROLE AU COURS DU VIEILLISSEMENT

Dans ce chapitre, nous présenterons, sous la forme d'un article scientifique soumis à la revue *Brain and Language* (novembre 2016), une étude par stimulation magnétique transcrânienne (TMS) de la perception multisensorielle de la parole au cours du vieillissement (relative à l'état de l'art proposé dans la section D-4 de la partie théorique de cette thèse). Cette expérience a été réalisée dans le cadre d'un stage de trois mois à l'étranger effectué d'octobre à décembre 2015 dans le Laboratoire des Neurosciences de la Parole et de l'Audition de l'Université Laval (Québec). Elle a été réalisée sous la supervision de Pascale Tremblay (directrice du laboratoire d'accueil) et de Marc Sato, Coriandre Vilain et Jean-Luc Schwartz.

Treille, A., Sato, M., Schwartz, J-L., Vilain, C. & Tremblay, P (soumis). On the lateralization of the premotor cortex in multimodal speech perception as a function of age: a rTMS study. Brain and Language.

Nous avons vu précédemment que les modèles neurobiologiques du langage étaient en faveur d'une latéralisation gauche de la voie dorsale audio-motrice durant la perception de la parole. Cependant, le rôle spécifique du cortex prémoteur ventral (PMv) droit et gauche dans les processus d'intégration multisensorielle de la parole reste relativement peu étudié. D'autre part, la littérature nous a également montré que les personnes âgées avaient un gain audio-visuel supérieur aux jeunes adultes malgré un déficit sensoriel naturel. Un recrutement plus important des régions motrices pourrait être une piste pour expliquer ce gain. Le but de cette expérience était de clarifier le rôle du PMv droit et gauche lors d'une tâche de perception de parole uni- et multisensorielle chez des adultes âgés de 18 à 78 ans.

Pour cela, nous avons utilisé une technique de stimulation magnétique trancrânienne (TMS) répétée, pour induire une inhibition des régions prémotrices gauche ou droite. Cette séance de stimulation était immédiatement suivie d'une tâche d'identification syllabique dans cinq conditions de présentation : auditive seule, audio-visuelle, visuelle seule, tactile seule et audio-tactile. Pour la perception tactile nous nous sommes inspirés de la méthode Tadoma que nous avions utilisée dans deux de nos précédentes études (voir la section A de la partie expérimentale de cette thèse).

Trois résultats principaux ont été obtenus. Tout d'abord, indépendamment de l'âge des participants, une facilitation et une amélioration de la reconnaissance de la parole auditive ont été observées lors de l'ajout des informations visuelles ou tactiles au signal acoustique, suggérant une préservation des mécanismes d'intégration au cours du vieillissement. D'autre part, nos résultats ont également montré une réduction de l'asymétrie

hémisphérique liée à l'âge lors d'une stimulation inhibitrice du PMv durant la perception auditive seule. Cela suggère l'existence de mécanismes compensatoires moteurs pour pallier le déclin naturel des traitements de la parole. Finalement, une interaction entre l'ordre de stimulation (PMv droit en premier ou PMv gauche en premier) et la région cible (PMv droit ou PMv gauche) a été observée, suggérant une implication du PMv gauche durant la phase d'apprentissage associée à la reconnaissance de la syllabe.

Pris ensemble, ces résultats démontrent que les mécanismes d'intégration sont maintenus, au moins en partie, avec l'âge et ce malgré un déclin de l'acuité auditive inévitable. Ce maintien serait notamment possible grâce au recrutement du PMv droit comme mécanisme compensatoire pour pallier cette perte sensorielle.

On the lateralization of the premotor cortex in multimodal speech perception as a function of age: a rTMS study

Avril Treille[1], Marc Sato[2], Jean-Luc Schwartz[1], Coriandre Vilain[1] and Pascale Tremblay[3,4]

[1.] GIPSA-Lab, Département Parole & Cognition, CNRS & Grenoble Université, France
[2.] Laboratoire Parole & Langage, CNRS & Aix-Marseille Université, France
[3.] Centre de Recherche de l'Institut Universitaire en santé mentale de Québec, Québec, Canada.
[4.] Département de Réadaptation, Université Laval, Québec, Canada

Corresponding author:
Pascale Tremblay, Ph.D.
Département de réadaptation, Université Laval
1050 Avenue de la Médecine, Québec (QC)
Office 4462
CANADA, G1V 0A6
Email:
Pascale.Tremblay@fmed.ulaval.ca
Phone: +001 418 663-5000 ext. 4738
Fax: +001 418 656-5476

ABSTRACT

Although neurobiological models of language argue for a left lateralization of the audio-motor dorsal pathway during speech perception, the specific role of the right and left premotor ventral (PMv) areas in multisensory speech integration processes remains largely unknown. The goal of this study was to clarify the role of these areas during uni- and multisensory speech perception in cognitively healthy adults varying in age. To this aim, an inhibitory offline transcranial magnetic stimulation approach was used in combination with a multimodal syllable identification task. Three main results emerged from this study. First, independent of the age of the participant, better and faster recognition of auditory speech stimuli was observed when visual or tactile information was added to the acoustic speech signal. Second, an age-related reduction in hemispheric asymmetry on the effect of TMS on the PMv during auditory syllable perception was found, suggesting a compensatory motor mechanism to support decline of speech processing. Third, an interaction between the order of stimulation and the target region was observed, suggesting an involvement of the left PMv during the learning phase associated with syllable recognition. Taken together, these results demonstrate that multisensory integration mechanisms are, at least in part, maintained with age despite a decline in auditory acuity, through the recruitment of the right PMv as a possible compensatory mechanism to support a declining peripheral auditory system.

Keywords: Aging, Multisensory Integration, Transcranial magnetic stimulation, syllable discrimination

1- INTRODUCTION

The human brain is continuously bombarded with large amounts of sensory information from various sources, some of which are quickly and effortlessly combined into single objects, events or understandable coherent percepts. Speech perception is a unique example of such multisensory processing that interfaces with the linguistic system. During natural conversation, we simultaneously perceive speech via sounds (i.e., a speaker's voice) and visual cues (i.e., a speaker's articulatory movements and facial expressions). Although humans are proficient in extracting phonetic features from the acoustic signal alone, they are, to a lesser extent, capable to read lip movements when the auditory signal is degraded or not present, and interactions between auditory and visual modalities are beneficial during speech perception. One of the most well-known illustrations of audio-visual interactions is the McGurk illusion [1]. The illusion occurs when the auditory signal of one syllable (/ba/) is paired with the visual component of another syllable (/ga/), leading to the perception of a third syllable (/da/), demonstrating the interaction between auditory and visual information in speech processing. The complementarity of auditory and visual information and their redundancy are also known to improve the intelligibility of a degraded acoustic signal [2, 3], to facilitate the comprehension of semantically complex discourses [4], to enhance the perception of strongly accented speech as well as non-native-languages [5]. In addition to the visual and auditory modalities, speech can also be perceived by the hand, with orofacial speech gestures felt and monitored from

manual tactile contact with the speaker's face. Previous behavioral studies have shown that adults without sensory impairment, as well as blind and deaf-blind adults are capable to discriminate syllables based on tactile information alone [6, 7, 8, 9]. Moreover, few studies also demonstrated the influence of tactile information on auditory speech perception, especially when the signal is noisy or ambiguous [7, 10, 11]. Several magneto- and electro-encephalographic studies have shown that both visual and tactile speech inputs modulate activity early in the auditory cortex [8, 9, 12, 13, 14, 15, 16, 17, 18, 19]. Taken together, these results support the existence of multisensory integration mechanisms during speech perception, not only between the well-known auditory and visual modalities but also between the auditory and less-used tactile modalities.

Recently, neurobiological models of speech perception [20, 21] have highlighted a possible role for the motor system not only during speech production but also during the process of matching motor and auditory speech representations. According to these models, the speech perception network includes auditory, temporo-parietal and frontal regions and is divided into two different streams: a ventral and a dorsal stream. The ventral stream connects several temporal and prefrontal areas and is thought to map auditory and phonological representations onto lexical conceptual representations. The dorsal stream, or ''auditory–motor integration'' pathway, runs dorso-caudally from the posterior part of the superior temporal gyrus to the ventral premotor cortex (PMv), through the ventral inferior parietal lobule, in order to map auditory and phonological representations to articulatory motor representations. Interestingly, in addition to the posterior superior temporal sulcus and gyrus known to play a crucial role in multisensory speech integration [22, 23, 24, 25, 26, 27] the PMv also appear to be involved during audio-visual speech interactions [7, 28, 29]. This hypothesis is supported by brain-imaging studies showing stronger motor activity during audio-visual compared to auditory speech perception [25, 26], and during the perception of incongruent compared to congruent audio-visual conditions [30, 31] or to unimodal conditions [22]. Yet, irrespective of modality, the role of the motor system in speech perception is still debated. Because speech perception and production are closely linked in everyday life situations (e.g., during social interactions), it remains unclear whether the motor activity that is observed during tasks requiring auditory-motor mapping is causally related to speech processing, or whether it is simply epiphenomenal [20, 26, 32, 33, 34]. Another unresolved question is related to the lateralization of the contribution of the motor system to speech perception. While several models suggest a left lateralization of speech functions [20, 35], some studies have reported bilateral PMv activation [e.g. 36, 37], and most TMS studies have focused only on the left PMv [e.g. 33, 38, 39], leaving the question of a potential hemispheric specialization unanswered.

While it has been suggested that the implication of the PMv in speech perception may be stronger when the sounds are masked or distorted compared to clear speech perception [40, 41, 42] the role of the PMv in situations of permanent sensory degradation is unknown. The study of normal aging provides an opportunity to test whether permanent degraded auditory abilities are associated with stronger recruitment of this region [34]. According to this hypothesis, the well-known decline in the peripheral hearing system associated with aging [43, 44, 45, 46] could be compensated to a certain extent by relying more strongly on preserved motor representations stored in this region, exploiting a possible compensatory ability of the motor system to help maintain performance [35]. This audio-motor comparison would partly constrain phonetic interpretation of the sensory inputs through the internal generation of candidate articulatory categorizations [26, 21, 47, 48]. Regarding multisensory speech perception, it has been shown that older adults benefit significantly from additional visual cues compared to an auditory signal alone [49], are more sensitive than younger adults to visual information during the perception of McGurk stimuli [50] and show a larger gain of audio-visual speech integration [51, 52]. It is therefore possible that older adults use multisensory integration as another compensatory mechanism against declining auditory abilities, whereby the motor system would help disambiguate the incoming sensory speech inputs.

Participants' Characteristics

	Age		Education (in years)		MoCA		Acuities		
							Hearing	Vision	Tactile
N	Mean ± SD	Range	Mean ± SD	Range	Mean ± SD	Range	Mean ± SD	Mean ± SD	Mean ± SD
24	46 ± 19	19-78	17 ± 3	11-23	28 ± 1,5	25-30	- 8,7 ± 8,1	0,9 ± 0,34	3,1 ± 0,7

Table 1: Participants' characteristics: Number of participants, age (in years), education (in years), MoCA (Montreal Cognitive Assessment) scores and sensory acuities (Hearing, vision and tactile).

The goal of this study was to clarify the role of the left and right PMv during uni- and multi-sensory speech perception in healthy adults varying in age. To this aim, an inhibitory offline transcranial magnetic stimulation (TMS) approach was used in combination with a multimodal syllable identification task under five modalities of presentation: auditory, visual, tactile, audio-visual and audio-tactile. First, given evidence of a lateralization of the dorsal pathway during speech perception [20, 39; for meta-analyses: 53, 54], we expected the effect of TMS on syllable recognition to be stronger when applied on the left PMv compared to the right PMv. Second, given the known differences in speech perception between young and older adults, we expected the effect of TMS to be stronger in older adults, reflecting stronger reliance on the PMv to compensate for a less efficient auditory system. Finally, due to the unfamiliar aspect of the tactile modality and previous studies showing stronger motor activity during visual and audio-visual compared to auditory modalities, we expected the effect of TMS to be stronger in these conditions.

2- METHODS

Participants

A total of 26 healthy adults (16 females; 19 to 78 years; mean 46± 19 years) participated in the study. Participants were volunteers recruited by means of emails and poster distributed in Québec City including but not limited to Université Laval's campus. One participant was excluded due to an incidental finding (presence of a lesion in the orbitofrontal cortex), leaving 25 participants in the study. The mean number of years of education in the sample was 17± 3 years (11-23 years). All participants were native French speakers and were right-handed as assessed by the Edinburgh Handedness Inventory [55]. All participants had normal hearing and normal or corrected-to-normal vision and reported no history of language disorders. Participants were screened for neurological, psychiatric, and other medical conditions and contraindications to MRI and TMS [56, 57] before their arrival at the laboratory (phone interview) and again upon arrival. All participants scored normal or above normal (≤26/30) on the Montreal Cognitive Assessment (mean 28/30± 1.5; [58]). Participants' characteristics are reported in Table 1. Written informed consent was obtained for all participants and they were compensated for the time spent in the study. The protocol was approved by the Institutional Ethical Committee of the Institut Universitaire en Santé Mentale de Québec (#393-2015).

Sensory acuity assessment

Hearing: To ensure that participants had normal hearing, pure tone audiometry was performed using a clinical audiometer (AC40, Interacoustic) for each ear separately, at the following frequencies: .5, 1, 2 kHz. For each participant, a standard pure tone average (PTA: average of threshold at .5, 1 and 2 kHz) was computed for the left and right ear. PTAs are used in clinical settings as a measure of hearing loss for speech because most speech sounds fall within this range [59]. All participants had normal hearing. The result of the hearing assessment is provided in Table 1.

Vision: To ensure that participants had comparable normal or corrected-to-normal vision, a standard Snellen test was used to measure visual acuity at 6 meters, with eleven lines of block letters decreasing in size. Participants were asked to cover one eye and to read aloud the letters one row at a time, beginning at the top. The smallest row that can be read accurately indicates the visual acuity in that specific eye. The procedure was repeated for the other eye. The participants who have corrected vision keep their glasses or contact lens during the assessment. All participants had normal vision. The results of the vision assessment are provided in Table 1.

Tactile sensitivity: To ensure that participants had comparable tactile sensitivity in the right hand, tactile sensitivity of the right hand thumb tip was measured with a standard two-point testing procedure. The thumb was used because this finger was used to perceive lip movements during the tactile perception condition of the main experiment (see section 3.2). A plastic disc with 8 labeled fixed 2 point intervals ranging from 1 to 8mm was used (Discriminator, Jamar). Participants were asked to close their eyes and the disc was placed randomly either 1 or 2 points

perpendicular to the skin. We began with the largest interval (8mm) and decreased the interval until 2mm, alternating with 1 or 2 points. Participants were asked to report how many points they felt. The first 2 points reported as a single point is the tactile threshold. The results of the tactile assessments are provided in Table 1.

The analysis of the correlation between sensory measures and age revealed that only hearing acuity decreased with age ($r = -0.693$, $p<.001$) while visual ($r = -0.388$, $p<.06$) and tactile ($r = 0.324$, $p<.12$) scores were not correlated with age.

Experimental procedure

The experiment entailed two visits on two different days. During the first visit, participants were screened for MRI and TMS counter-indications and then underwent structural magnetic resonance imaging (MRI). Two participants already had an MRI saved in the lab's participant databank (Banque de données sur l'Audition et la Communication Humaine "BACH", approved by our local research ethics committee, project #369-2014); for those participants, the study entailed only one visit. The second and main visit took place at the Institut Universitaire en Santé Mentale de Québec, in a double-walled sound attenuated room. The visit was divided into sessions during which participants underwent rTMS of left or right vPM, each followed by a short forced-choice identification task (see section 3.2)

1.1. Transcranial magnetic stimulation

1.1.1. MRI acquisition and co-registration

A high-resolution T1-weighted MRI scan was obtained for all participants on a Philip Achieva 3T MRI scanner (Philips Healthcare, Best, The Netherlands; matrix 256mm x 256 mm, 181 slices, 1mm x 1mm x 1mm, no gap). Once obtained, the anatomical MRI was incorporated into Brainsight TMS (Rogue Resolutions Ltd, Montreal, Canada) to guide coil placement. For each participant, an MRI-to-head co-registration was performed. The position of four anatomical landmarks (tip of the nose, bridge of the nose, inferior edge of the tragus of left and right ears), previously identified on participant's MRI, was assessed using an infrared tracking system (Polaris, Northern Digital, Waterloo, Canada). Upon successful co-registration, infrared tracking was used to monitor the position of the coil with respect to the participant's brain.

1.1.2. Resting motor threshold (RMT)

Stimulation was performed with a70mm Double Air Film coil connected to a Magstim Rapid2 biphasic stimulator (Magstim company, Dyfed, UK). The Air Film coil allows for quiet, temperature-regulated stimulation using managed ambient airflow and integrated temperature-regulated fan technology.

For the determination of each subject's resting motor threshold (RMT), the coil was placed over the participant's left motor cortex hand area (M1) within the precentral gyrus with the coil held tangentially to the skull, and the handle pointing posterior and down. Muscle activity was obtained from two disposable solid gel surface electrodes (EL504, Biopac Systems Inc; bipolar montage) placed over the first dorsal interosseous (FDI) muscle of the right hand and connected to the EMG module of Brainsight. Single pulses were delivered to M1, with the intensity of the stimulation adjusted until a muscle evoked potential (MEP) in the right hand was observed on the EMG in 5 out of 10 trials with an amplitude of at least 50 μV [60]. For 5 subjects, the RMT was obtained visually by one of the investigators in 5 out of 10 trials because of technical difficulties with the EMG recordings. The location of the stimulation was adjusted to locate the maximally excitable hand area.

1.1.3. rTMS stimulation

The intensity of stimulation was set at 115% of subjects' RTM, which ranged from 45-85 % of the output capacity of the stimulator, with a mean of 66 ± 10%. The stimulation sites were determined individually for each participant using Brainsight TMS software (Rogue Resolutions Ltd, Montreal, Canada). First, we identified the left inferior frontal sulcus (IFS) and the left precentral sulcus. The stimulation site was then set to the anterior part of precentral gyrus at the junction of the IFS and the precentral sulcus, corresponding to the ventral part of the premotor cortex (PMv). The mean coordinates, in MNI space were -52, 9, 26. The same procedure was applied to find the right PMv. The mean coordinates, in MNI space were 52, 11, 27. The location of these two stimulation sites is illustrated in Figure 1.

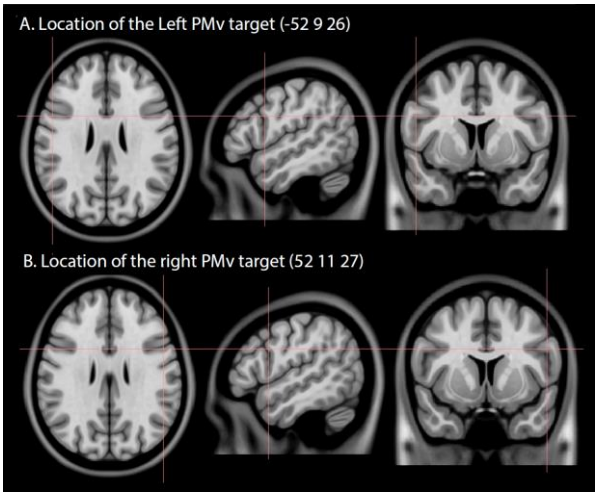

Figure 1: Mean location of the rTMS stimulation and associated MNI coordinates in the right and the left ventral premotor cortex (PMv), shown on the lateral surface of a Brainsight 3D rendered brain.

A conventional slow paradigm with an intensity level of 115% of RTM, that is, a low frequency-high intensity protocol [57] was used to inhibit activation in the PMv. Stimulation was applied in one offline train of 900 pulses delivered at the rate of 1 Hz for 15 min. External triggering of the Magstim device was achieved via Presentation software (Neurobehavioral Systems, Albany, CA, USA). Similar rTMS protocols have been shown to decrease the sensitivity of PMv during sentence processing or phonological task [33, 37]. Importantly, these parameters are well within the published Safety Guidelines for rTMS [57]. The two TMS sessions (right and left PMv) were separated by 1 hour and fully counterbalanced across participant to avoid order effects.

1.2. Forced-choice identification task

Following completion of each TMS session, participants completed a short (10 min) forced-choice identification task. During this task, participants were seated in a double-walled sound attenuated room at arm's length from a female experimenter (A.T.). They were told that they would be presented with /pa/, /ta/ or /ka/ syllables produced by the experimenter in 5 different sensory modalities: auditory, visual, audio-visual, tactile or audio-tactile. Participants were instructed to identify, as quickly as possible, each perceived syllable by pressing on one of three keys corresponding to /pa/, /ta/ or /ka/ on a response pad (Cedrus, RB-840) with their left hand. The production of the stimuli by the experimenter was visually triggered using a computer screen (DELL P2412Hb) placed behind the participant (but facing the experimenter) using Presentation software (Neurobehavioral Systems, Albany, CA, USA). The stimuli for the identification task were three simple consonant-vowel (CV) French syllables, /pa/, /ta/, and /ka/, which were selected based on previous studies on audio-visual and audio-tactile speech perception [8, 9]. These syllables were selected in order to ensure a gradient of visual and tactile recognition with the bilabial /p/ consonant known to be more visually and haptically salient than the alveolar /t/ and velar /k/ consonants. In the auditory modality (A), participants were asked to keep their eyes closed while listening to each syllable produced by the experimenter. In the audio-visual modality (AV), they were instructed to also look at the experimenter's face. In the audio-tactile modality (AT), they were asked to keep their eyes closed with their right hand placed on the experimenter's face (the thumb placed lightly and vertically against the experimenter's lips and the other fingers placed horizontally along the jaw line in order to help distinguishing both lip and jaw movements). This experimental procedure was adapted from the Tadoma method [6, 7, 61, 62] and similar to that previously used by Treille and colleagues [8, 9]. Finally, the visual-only (V) and the tactile-only (T) modalities were similar to the AV and AT modalities with the exception that the experimenter silently articulated each syllable providing participants with visual or tactile but no auditory feedback.

To increase task-difficulty and thus recruit the PMv more intensely, half the trials were presented in noise, while the other half was presented in quiet. White noise was delivered through a loudspeaker (Bi-amplified loudspeakers 80208, GENELEC, Lisalmi, Finland) located on the left side of the participant and oriented towards his/her ear. A -6dB signal-to-noise ratio (SNR) was calculated from the average intensity of 10 productions of the experimenter. The experimenter adjusted her voice intensity to maintain the -6 dB SNR using a sound level meter before the participant arrival. Each modality was presented in 15 trials blocks (i.e. 5 trials for each syllable /pa/, /ta/ and /ka/) presented in a randomized sequence. The inter-trial interval was 3 sec. A total of 150 trials were produced per session. The order of the modality of presentation and the response key designation were fully counterbalanced across participants but stay the same for both sessions. Presentation software (Neurobehavioral Systems, Albany, CA) was used to control the stimulus presentation and to record key responses. The experimenter' oral productions were recorded using a high quality Lavalier microphone (SHURE, MX150) in order to calculate reaction times (RT).

Data analyses

One participant was removed from the statistical analysis because of technical problem during the recording of one condition. Consequently, all

statistical analyses were performed on 24 participants.

We first examined the accuracy for the main task. The proportion of correct responses was determined for each participant and each condition (target regions of stimulation, noise/no noise environments, modalities). A four-way repeated-measure ANOVA was performed on the mean accuracy with Target Region (left PMv/right PMv), Noise (noise/no noise), and Modality (A, AV, AT, V, T) as the within-subjects categorical factors, Order of stimulation (left then right PMv, right then left PMv) as a categorical between-subjects factor, and Age as continuous quantitative between-subjects co-variable.

In order to calculate reaction times (RT), a semi-automatic procedure based on intensity and duration algorithm detection using Praat software [63] roughly identified each syllable's onset on the voice recordings. For all syllables, these onsets were further manually and precisely edited, based on waveform and spectrogram information related to the acoustic characteristics of voiced stop consonants. Because the experimenter silently produced the syllable in the V and T modalities, hence with no available temporal reference for RT, acoustical analyses were only performed for A, AV and AT modalities. RT was defined as the time from the consonant onset of each syllable produced by the experimenter to the onset of the subject's response. A four-way repeated-measure ANOVA was performed on median RT with Target Region (left PMv/right PMv), Noise (noise/no noise), and Modality (A, AV, AT) as the within-subjects categorical factors, Order of the stimulation (left then right PMv, right then left PMv) as a between-subjects categorical factor, and Age as a continuous quantitative between-subjects co-variable.

For all the analyses, the significance level was set at p=0.05 and Greenhouse-Geisser corrected (for violation of the sphericity assumption) when appropriate. To decompose the main effects and two-Way interactions, *post hoc tests* were performed. To decompose 3-way interactions, two-way repeated-measure ANOVAs or linear regressions were computed. For all ANOVAs, measures of effect sizes are provided in the form of partial eta squared (η_p^2), which are reported for all main effects and interactions. When comparing two means, we report effect sizes in the form of Cohen *d* statistics.

3- RESULTS

Accuracy

The percentages of correct responses for each condition are displayed in Figure 2.

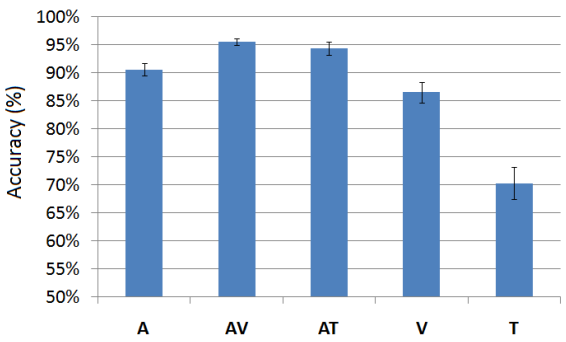

Figure 2: Mean accuracy (in percentage) for Auditory (A), Audio-Visual (AV), Audio-tactile (AT), Visual (V) and Tactile (T) conditions.

The ANOVA revealed a significant main effect of **Modality** ($F_{(4,84)}$=3.86, p<.02, η^2 = 0.15, Mean±SE: *A*: 91±1%; *AV*: 96±1%; *AT*: 94±1%; *V*: 87±2%; *T*: 70±3%). Overall, accuracy was lower in V and T modalities than in the bimodal conditions (V compare to AV: p<.0001, d=1.146; V compared to AT: p<.005, d=0.948; T compared to AV: p<.0001, d=1.556; T compared to AT: p<.0001, d=1.492). Accuracy was lower in T modality compared to A and V modalities but not in A compared to V (V compared to A: p<.484, d=0.546; T compared to A: p<.0001, d=1.38; T compared to V: p<.001, d=1.344). Moreover, AV stimuli were more correctly identified than A stimuli (p<.006, d=0,492) while AT scores were significantly different from A scores (p=.012, d=0.674) but not AV (p<.292, d=0.244).

The effect of **Order** was also significant ($F_{(1,21)}$=4.24, p=.05, η^2=0.17) with more correct responses when the left PMv was stimulated first than when the right PMv was stimulated first (left PMv first: 89±1%; right PMv first: 86±1%).

Finally, no significant main effects of Target Region, Noise and Age were observed for accuracy.

Two-way interactions were found between **Modality and Order** ($F_{(1,21)}$=2.45, p<.05, η^2=0.13) and between **Target Region and Order** ($F_{(1,21)}$=6.34, p<.02, η^2=0.23). To decompose the **Modality x Order interaction**, post hoc tests were conducted for each modality (A, AV, AT, V, T), which revealed a significant order effect only for V (p<.01, t: 2.51, d=0.924) and T (p<.04, t: 1.83, d=0.711), with less correct responses when the right PMv was stimulated first (V: left than right PMv: 91±1%, right than left PMv: 83%±2%; T: left than right PMv: 75± 2%, right than left PMv: 65±5%). To decompose the **Target region x Order** interaction, post hoc tests were conducted for each region (right PMv, left PMv), which revealed a significant order effect only for the right PMv (p<.08; t:2.605, d=0.951), with fewer correct responses when the right PMv was stimulated first (right PMv: left than right PMv:

90±2%; right than left PMv: 84%±2%; left PMv: left than right PMv: 88±1%; right than left PMv: 87%±2%).

In addition to the 2-way interactions, the ANOVA also revealed a significant 3-way interaction between **Target Region, Modality and Order** ($F_{(4,84)}$=3.55, p<.03, η^2=0.44). To decompose the 3-Way interaction, a series of 2-Way ANOVAs were conducted, one for each modality (A, AV, AT, V, T), with Order and Target Region as the within-subject factors. First, because age was controlled in the main analysis, the effect of age was regressed out from the dependent variables before running the 2-way ANOVAS. For A, AV and AT, the ANOVA revealed no significant main effect and no interaction. For V, the ANOVA revealed a main effect of order ($F_{(1,22)}$=6.40, p<.04, η^2=0.19, d=0,924) with less correct responses when the right PMv was stimulated first. For T, the ANOVA revealed a 2-way interaction between TMS and Order ($F_{(1,22)}$=, p<.05, η^2=0.17). To decompose this interaction, post hoc tests were conducted which revealed no significant difference between scores for the left PMv (t: 0.63, p<.27; d=0,234), but a significant difference was found for the right PMv (t: 2.93; p<.004; d=1,00) with more correct responses when the left PMv was stimulated first.

RT

RT for each condition are displayed in Figure 3.

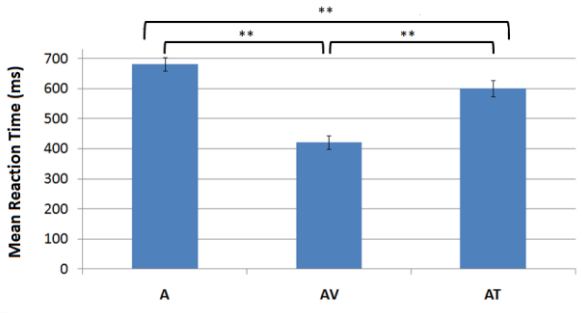

Figure 3: Mean RT for Auditory (A), Audio-Visual (AV) and Audio-Tactile (AT) conditions. Because the experimenter silently produced the syllable in the V and T modalities, RT could not be calculated for these conditions.

A significant main effect of **Age** was found on RT ($F_{(1,21)}$=9.30, p<.006, η^2=0.31) with increasing RT with older age (see Figure 4).

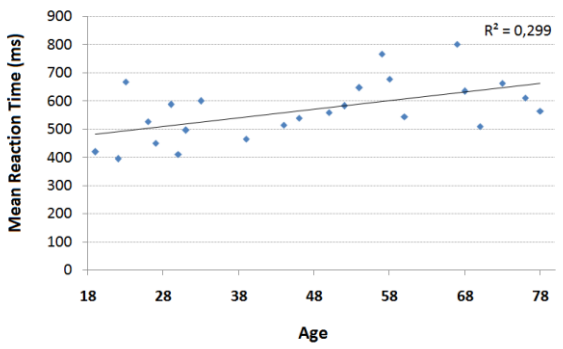

Figure 4: Significant positive relationships between mean RT and age.

The ANOVA also revealed a significant main effect of **Modality** ($F_{(2,42)}$=8.74, p<.001, η^2=0.294, Mean±SE: *A*:681±19ms; *AV*: 421±22ms; *AT*: 601±22ms). Post hoc tests revealed that responses were slower in A compared to AV (p<.0001, d=1.525) and AT (p<.0001, d=0.643), and in AT compared to AV (p<.0001, d=1.212).

Finally, no significant main effects of Target Region, Noise and Order were observed for RT.

A significant 2-way interaction between **Target Region and Order** was found ($F_{(1,21)}$=41.56, p<.0001, η^2=0.67; see Figure 5). To decompose this interaction, post hoc tests were conducted for each target region (left PMv and right PMv) which revealed no RT difference for the left hemisphere when it was stimulated in the first or in the second session (t: 0.83, p<.21, d=0.337; left PMv first: 599±114 ms, left PMv second: 562±107 ms), but lower RT for the right hemisphere when it was stimulated first (t: -1.78, p<. 05; d=0.696 right PMv first: 594±116 ms, right PMv second: 515±99 ms).

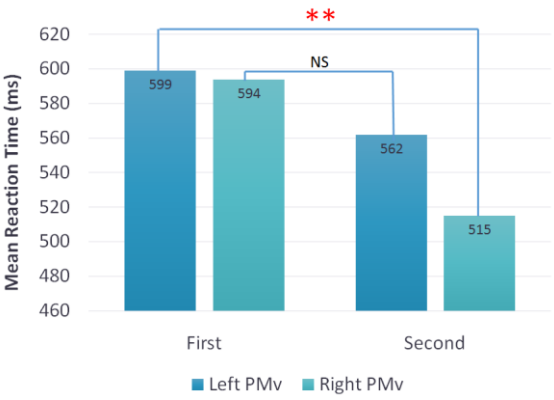

Figure 5: Interaction between Target region (left and right PMv) and Order of stimulation on RT: First refers to the first session of stimulation while second refers to the second session of stimulation.

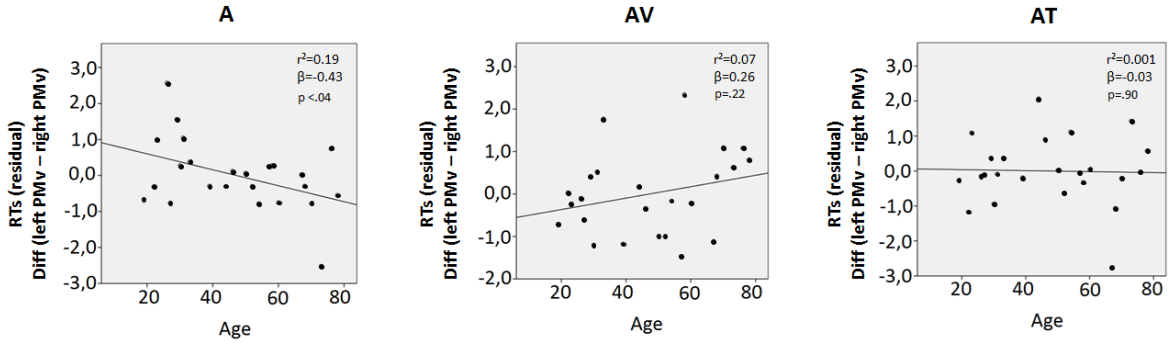

Figure 6: Residual Hemispheric difference score (RT left PMv – RT right PMv) as a function of Age for Auditory (A), Audio-Visual (AV) and Audio-Tactile (AT) conditions. A significant linear negative relationship was observed only in the Auditory condition.

Finally, a 3-way interaction was found between **Target Region, Modality and Age** ($F_{(2,42)}$=3.29, p<.047, η^2=0.13). To decompose this interaction, a hemispheric difference score was computed (RT left PMV – RT right PMv). The relationship between the RT and Age was assessed, separately for each modality, using simple linear regressions, after regressing out the effect of order (see Figure 6). For A, there was a significant linear negative relationship between the hemispheric difference score and age (r^2 =0.19, β=-0.43, p<.04). For AV and AT there was no significant relationship between the hemispheric difference score and age (AV: r^2= 0.07, β= 0.26, p=.22; AT: r^2: 0.001, β=-0.03, p=.90).

4- DISCUSSION

The goal of this study was to clarify the role of the left and right PMv during uni- and multisensory speech perception in cognitively healthy adults varying in age. To this aim, an inhibitory offline transcranial magnetic stimulation approach was used in combination with a multimodal syllable identification task. Three main results emerged from the present study.

First, better and faster auditory stimuli recognition was observed when visual and tactile information were added to the acoustic speech signal. These results are in line with previous behavioral studies on audio–visual and audio-tactile speech perception and support the notion that visual and tactile information can facilitate auditory speech processing. Importantly, the same gain in multisensory compared with auditory perception was observed whatever the age. Second, regarding reaction time and in line with the HAROLD model of neurocognitive aging [64], we found an age-related difference in hemispheric asymmetry in the PMv during auditory syllable perception – though not during audio-visual and audio-tactile perception. Third, a complex interaction effect between the order of stimulation and the target region was observed.

Before discussing these results, it is important to consider an important limitation of the present study. Contrary to what we expected, we found no significant effect of noise and interaction between noise and any factor. This null result may be explained by the experimental procedure that was used. Indeed, the syllable identification task implied live dyadic interactions between a listener and the experimenter. Although this procedure maximized the naturalness of the stimuli (live, unique), it was difficult to maintain the SNR constant. Noise levels might therefore not have been sufficient to make the behavioral task difficult enough to be disrupted by TMS. This interpretation is supported by an overall accuracy rate of 88 %. In addition, because of the use of online stimuli, it was impossible to merge the source of the speech provided by the experimenter with the noise sound being delivered through a loudspeaker. The separation of these two different auditory sources (voice and noise) may have limited the masking efficiency of the noise. Indeed, Füllgrabe [65] showed that spatially separated auditory sources are easier to recognize than co-localized sources for elderly adults. Given that previous repetitive and double-pulse TMS studies have shown a causal role for the PMv or the primary motor cortex during auditory syllable recognition task only when the stimuli were acoustically ambiguous or degraded syllables [32, 33, 38, 42, 66], the absence of noise effects along all the analysis of experimental results is likely to have reduced possible stimulation effects in the present study.

4.1 Multisensory perception in aging

One first important finding of this study is that participant's ability to combine visual or tactile information with auditory information appears to be preserved with age, at least during the performance of a simple syllable identification task. Interestingly, irrespective of the modality, a main effect of age was observed on reaction time but not on accuracy. That is, older participants were slower but not less accurate in identifying speech stimuli. Several studies have shown recognition difficulties in elderly adults

during syllable discrimination [45, 46] as well as during phoneme discrimination tasks [44], particularly in the presence of background noise or when the speech sounds are distorted [65, 67, 68, 69, 70]. The low level of noise and the separability of sounds sources may explain the absence of effect for accuracy. Perhaps even more importantly, performance was almost at ceiling level (on average, 88%), suggesting that identification task was easy even for older adults.

4.2 Change in hemispheric laterality

The HAROLD model (Hemispheric Asymmetry Reduction in OLDer adults), first introduced by Roberto Cabeza [64], suggests that, with age, neural activity in the prefrontal cortex (PFC) becomes less lateralized, and that homologous regions are recruited to compensate for the decline in the efficiency of neural processes during episodic memory. This phenomenon has been extensively described in the literature [71, 72, 73, 74, 75] Though this age-related change in laterality has been demonstrated mainly in the PFC, Berlingeri and colleagues [76] have shown a similar phenomenon in the superior temporal gyrus (STG) and temporal pole during semantic judgment of sentences, which suggests that hemispheric asymmetry reduction may be a general neurobiological aging mechanism. In the present experiment, it was assumed that the dorsal pathway, including the PMv – our region of interest - is relatively left lateralized [20]. Our first hypothesis was that rTMS on the left PMv, but not the right PMv, would have an impact on speech perception, due to the recruitment of the motor system to facilitate speech perception. Moreover, we expected this effect to become stronger with age to compensate for sensory/perceptual decline. In order to test the hemispheric asymmetry reduction hypothesis, we examined RT hemispheric difference as a function of age. The results demonstrate a difference in RT for left and right PMv. This suggests that stimulating the PMv has a different impact depending on the hemisphere. Importantly, this hemisphere difference changed with age, suggesting a stronger recruitment of the right PMv to compensate for the well-established age-related hearing decline. This result is in line with the hemispheric asymmetry reduction principle described in the HAROLD model and suggests (1) that the bi-lateralization of neural processes in aging occurs not only in frontal regions but also within the PMv, and (2) that these compensations may be cognitive in nature, or motor such as was found in the present study. Importantly, this hemispheric asymmetry change was observed only for the auditory condition. This is consistent with the results of the acuity tests: only auditory capabilities showed an age-related decline. When visual or tactile information was added, this effect disappeared. This suggests that adding sensory information from intact modalities can contribute to compensating for the age-related hearing decline and 2) the lack of age effect related to an inhibition of left PMv during audiovisual speech perception suggests a preservation of integration mechanisms in aging. We could have expected an age-related effect for audio-tactile stimuli after stimulation to the left PMv. The complex combination of normative hearing decline and an unfamiliar modality (tactile) could have made the audio-tactile stimuli more difficult. However, behavioral scores show a good identification of /pa/, /ta/ and /ka/ syllables in the tactile condition suggesting that audio-tactile integration is preserved in elderly adults.

4.3 Order of stimulation

An unexpected interaction between target region and order stimulation was observed on RT. In our experimental protocol, the two sessions of stimulation were separated by one hour to ensure that the effect of TMS did not cumulate from the first to the second session. Furthermore, the two sessions were identical with the exception of the target region. Hence, this Target by Order interaction might reflect a learning effect, whereby participants became faster during the second session. In the present study, no RT difference was found between the two hemispheres in the first session. This suggests that when participants perform the task for the first time, the left and right PMv are involved in a similar manner. When the right PMv was stimulated in the second session, we observed a reduction in RT compared to the first session related to the left PMv stimulation. This could be due to a classical learning effect, inducing faster RTs and no stimulation effect. However, when the left PMv was stimulated in second, RTs were not significantly different from the first right PMv stimulation. This nonsignificant result suggests that, despite a possible learning effect occurring after the first behavioral task, RTs are not facilitated when the left PMv was stimulated during the second session. This could indicate that, contrary to a right PMv inhibition, inhibiting the left PMv slowed task performance, which suggests that the left PMv is recruited after the task is learned but not the right PMv.

CONCLUSION

To conclude, in the present study we showed that multisensensory integration is preserved in aging, at least in the context of a simple speech perception task. Furthermore, in line with the HAROLD model of

neurocognitive aging, we found an age-related reduction in hemispheric asymmetry in the motor system during auditory syllable perception, suggesting an increased reliance on the motor system to compensate for a decline of auditory processes. Further studies are needed to replicate this finding in a larger group and in different experimental contexts. These results extend the domain of application of the HAROLD model from cognitive to sensorimotor processes and they also have important implications for models of neurobiological aging in general.

ACKNOWLEDGMENTS

This study was supported by research funds from the European Research Council to J-L.S. (FP7/2007-2013 Grant Agreement no. 339152), by a grant from the "Région Rhône-Alpes" to A.T., and by an infrastructure grant from the Canadian Foundation for Innovation (FCI; LOF #31408) to P.T., who also holds a Career Awards from the ‘‘Fonds de Recherche du Québec – Santé’’ (FRQS). MRI images were acquired through the “Consortium d’imagerie en neuroscience et santé mentale de Québec” (CINQ) via a platform support grant (#3456) from the Brain Canada Foundation to P.T. Thanks to C. Boudreau for her help with participant recruitment and testing, and to all participants.

REFERENCES

[1] McGurk, H. & MacDonald, J. (1976). Hearing lips and seeing voices. Nature, 264:746-748

[2] Sumby, W.H. & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. The Journal of the Acoustical Society of America, 26: 212-215

[3] Benoît C., Mohamadi T. & Kandel S. (1994). Effects of phonetic context on audiovisual intelligibility of French speech in noise. Journal Speech Hearing Research, 37:1195–1203

[4] 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 LipReading. 97-114

[5] Navarra, J. & Soto-Faraco, S. (2005). Hearing lips in a second language: visual articulatory information enables the perception of second language sounds. Psychological research, 71(1):4-12

[6] Reed, C.M., Durlach, N.I., Braida, L.D. & Schultz, M.C. (1982). A analytic study of the Tadoma Method: Identification of Consonants and Vowels by an Experienced Tadoma User. Journal of Speech and Hearing Research, 25:108-116

[7] Sato, M., Cavé, C., Ménard, L. & Brasseur, A. (2010). Auditory-tactile speech perception in congenitally blind and sighted adults. Neuropsychologia, 48(12): 3683-3686

[8] Treille, A., Cordeboeuf, C., Vilain, C. & Sato, M. (2014a). Haptic and visual information speed up the neural processing of auditory speech in live dyadic interactions. Neuropsychologia, 57: 71-77

[9] Treille, A., Vilain, C. & Sato, M. (2014b). The sound of your lips: electrophysiological cross-modal interactions during hand-to-face and face-to-face speech perception. Frontiers in Psychology, 5(420): 1-9

[10] Fowler, C. & Dekle, D. (1991). Listening with eye and hand: crossmodal contributions to speech perception. Journal of Experimental Psychology- Human Perception and Performance, 17:816–828.

[11] Gick, B., Jóhannsdóttir, K.M., Gibraiel, D. & Mühlbauer, M. (2008). Tactile enhancement of auditory and visual speech perception in untrained perceivers. Journal of Acoustical Society of America, 123:72-76

[12] Klucharev, V., Möttönen, R. & Sams, M. (2003). Electrophysiological indicators of phonetic and non-phonetic multisensory interactions during audiovisual speech perception. Brain Res. Cogn. Brain Res., 18:65-75

[13] Besle, J., Fort, A., Delpuech, C. & Giard, M.H. (2004). Bimodal speech: early suppressive visual effects in human auditory cortex. European journal of Neuroscience, 20:2225-2234.

[14] van Wassenhove, V., Grant, K.W. & Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. Proceedings of the National Academy of Sciences U.S.A., 102:1181-1186

[15] Stekelenburg, J.J. & Vroomen, J. (2007). Neural correlates of multisensory integration of ecologically valid audiovisual events. Journal of Cognitive Neuroscience, 19:1964–1973

[16] Arnal, L.H., Morillon, B., Kell, C.A. & Giraud, A.L. (2009). Dual neural routing of visual facilitation in speech processing. The Journal of Neuroscience, 29(43):13445-13453.

[17] Pilling, M. (2010). Auditory event-related potentials (ERPs) in audiovisual speech perception. Journal of Speech, Language, and Hearing Research, 52(4):1073-1081

[18] Vroomen, J. & Stekelenburg, J.J. (2010). Visual anticipatory information modulates multisensory interactions of artificial audiovisual stimuli. Journal of Cognitive Neuroscience, 22:1583-1596

[19] Baart, M., Stekelenburg, J. J. & Vroomen, J. (2014). Electrophysiological evidence for speech-specific audiovisual integration. Neuropsychologia, 65:115–211

[20] Hickok, G. & Poeppel, D. (2007). The cortical organization of speech processing. Nature Review Neurosciences, 8:393-402

[21] Rauschecker, J. & Scott, S. (2009). Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. Nature Neuroscience, 12:718-725

[22] Calvert, G.A., Campbell, R. & Brammer, M.J. (2000).Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. Current Biology, 10(11):649-657

[23] Callan, D.E., Jones, J.A., Munhall, K.G., Callan, A.M., Kroos, C. & Vatikiotis-Bateson, E. (2003). Neural processes underlying perceptual enhancement by visual speech gestures. NeuroReport, 14:2213-2217

[24] Callan, D.E., Jones, J.A., Munhall, K.G., Callan, A.M., Kroos, C. & Vatikiotis-Bateson, E. (2004). Multisensory integration sites identified by perception of spatial

wavelet filtered visual speech gesture information. Journal of Cognitive Neuroscience, 16:805-816

[25] Skipper, J.I., Nusbaum, H.C. & Small, S.L. (2005). Listening to talking faces: motor cortical activation during speech perception. NeuroImage, 25: 76–89

[26] Skipper, J., Van Wassenhove, V, Nussman, H. & Small, S. (2007). Hearing lips and seeing voices: How cortical areas supporting speech production meditate audiovisual speech perception. Cerebral Cortex, 17: 2387-2399

[27] Beauchamp, M.S., Nath, A.R. & Pasalar, S. (2010). fMRI-Guided Transcranial Magnetic Stimulation Reveals That the Superior Temporal Sulcus Is a Cortical Locus of the McGurk Effect. The journal of Cognitive neuroscience, 30(7):2414 –2417

[28] Campbell, R., MacSweeney, M., Surguladze, S., Calvert, G., McGuire, P., Suckling, J., Brammer, M.J. & David, A.S. (2001). Cortical substrates for the perception of face actions: an fMRI study of the specificity of activation for seen speech and for meaningless lower-face acts (gurning).Cognitive Brain Research, 12:233-243

[29] Calvert, G.A. & Campbell, R. (2003). Reading speech from still and moving faces: the neural substrates of visible speech. Journal of Cognitive Neuroscience, 15:57–70

[30] Ojanen, V., Möttönen, R., Pekkola, J., Jääskeläinen, I.P., Joensuu, R., Autti, T. & Sams, M. (2005). Processing of audiovisual speech in Broca's area. NeuroImage, 25:333-338

[31] Pekkola, J., Laasonen, M., Ojanen, V., Autti, T., Jaaskelainen, L.P., Kujala, T. & Sams, M. (2006). Perception of matching and conflicting audiovisual speech in dyslexic and fluent readers: an fMRI study at 3T. NeuroImage, 29(3):797-807

[32] D'Ausilio A., Pulvermüller F., Salmas P., Bufalari I., Begliomini C., Fadiga L. (2009) The motor somatotopy of speech perception. Current Biology, 19:381-385

[33] Sato, M., Tremblay, P. & Gracco, V. (2009). A mediating role of the premotor cortex in phoneme segmentation. Brain and Language, 111(1): 1-7

[34] Tremblay, P., & Small, S. L. (2010). From Language Comprehension to Action Understanding and Back Again. Cereb Cortex, 21(5):1166-1177

[35] Guenther, F.H. & Vladusich, T. (2012). A Neural Theory of Speech Acquisition and Production. J. Neurolinguistics, 25(5): 408-422

[36] Wilson, S.M., Saygin, A.P., Sereno, M.I., Iacoboni, M., 2004. Listening to speech activates motor areas involved in speech production. Nat. Neurosci. 7 (7), 701–702.

[37] Tremblay, P., Sato, M., & Small, S. L. (2012). TMS-induced modulation of action sentence priming in the ventral premotor cortex. Neuropsychologia, 50(2):319-326

[38] Meister, I.G., Wilson, S.M., Deblieck, C., Wu, A.D., & Iacoboni, M. (2007). The essential role of premotor cortex in speech perception. Current Biology, 17(19):1692–1696

[39] Murakami, T., Kell, C.A., Restle, J., Ugawa, Y. & Ziemann, U. (2015). Left Dorsal Speech Stream Components and their contribution to phonological Processing. Journal of Neuroscience. 35(4): 1411-1422

[40] Binder, J.R., Liebenthal, E., Possing, E.T., Medler, D.A. & Ward, B.D (2004). Neural correlates of sensory and decision processes in auditory object identification. Nature Neuroscience, 7:295-301

[41] Zekveld, A.A., Heslenfeld, D.J., Festen, J.M. & Schoonhoven, R. (2006). Top-down and bottom-up processes in speech comprehension. Neuroimage, 32:1826–1836

[42] D'Ausilio A., Bufalari I., Salmas P., Fadiga L., (2012) The role of the motor system in discriminating degraded speech sounds. Cortex, 48(7):882–887

[43] Smith, R.A. & Prather, W.F. (1971). Phoneme Discrimination in Older Persons Under Varying Signal-To-Noise Conditions. Journal of Speech, Language, and Hearing Research, 14(3): 630-638

[44] Gordon-Salant, S. (1986). Recognition of natural and time/intensity altered CVs by young and elderly subjects with normal hearing. The Journal of the Acoustical Society of America, 80(6):1599-1607

[45] Strouse, A., Ashmead, D.H., Ohde, R.N. & Grantham, D.W. (1998). Temporal processing in the aging auditory system. The Journal of the Acoustical Society of America, 104(4): 2385-2399

[46] Bilodeau-Mercure, M., Lortie, C. L., Sato, M., Guitton, M. J. & Tremblay, P. (2015). The neurobiology of speech perception decline in aging. Brain Struct Funct, 220(2):979-997

[47] Rauschecker, J. P. (2011). An expanded role for the dorsal auditory pathway in sensorimotor integration and control. Hear. Res. 271, 16–25

[48] Schwartz, J.L., Ménard, L., Basirat, A. & Sato, M. (2012). The Perception for Action Control Theory (PACT): a perceptuo-motor theory of speech perception. Journal of Neurolinguistics, 25(5):336-354

[49] Laurienti, P.J., Burdette, J.H., Maldjian, J.A. & Wallace, M.T. (2006). Enhanced multisensory integration in older adults. Neurobiology of aging, 27:1155-1163

[50] Cienkowski, K.M. & Carney, A.E. (2002). Auditory-visual speech perception and aging. Ear and Hearing, 23(5):439-449

[51] Tye-murray, N., Sommers, M., Spehar, B., Myerson, J. & Hale, S. (2010). Aging, audiovisual integration and the principle of inverse effectiveness. Ear and Hearing. 31(5):636-644

[52] Sekiyama, K., Soshi, T. & Sakamoto, S. (2014). Enhanced audiovisual integration with aging in speech perception: a heightened McGurk effect in older adults. Frontiers in psychology, 5(323)

[53] Vigneau, M., Beaucousin, V., Hervé, P.Y., Duffau, H., Crivello, F., Houdé, O., Mazoyer, B. & Tsourio-Mazoyer, N. (2006). Meta-analyzing left hemisphere language areas: Phonology, semantics, and sentence processing. NeuroImage, 30:1414-1432

[54] Vigneau, M., Beaucousin, V., Hervé, P.Y., Jobard, G., Petit, L., Crivello, F., Mellet, E., Zago, L., Mazoyer, B. & Tzourio-Mazoyer, N. (2011). What is right-hemisphere contribution to phonological, lexico-semantic, and sentence processing? Insights from a meta-analysis. NeuroImage, 54(1):577-593

[55] Oldfield, R.C. (1971). The Assessment and analysis of handedness: The Edinburgh inventory. Neuropsychologia, 9:97-113

[56] Wasserman, E.M. (1998). Risk and safety of repetitive transcranial magnetic stimulation: report and suggested guidelines from the International Workshop on the safety of repetitive transcranial magnetic stimulation, June 5-7, 1996. Electroencephalogr Clin Neurophysiol, 108(1):1-16

[57] Rossi, S., Hallet, M., Rossini, P.M. & Pascual-Leone, A. (2009). Safety, ethical considerations, and application guidelines for the ise of transcranial magnetic stimulation in clinical practice and research. Clin Neurophysiol., 120(12):2008-2039

[58] Nasreddine, Z.S., Phillips, N.A., Bédirian, V., Charbonneau, S., Whitehead, V., Collin, I., Cummings, J.L. & Chertkow, H. (2003). The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. J. Am Geriatr Soc, 53(4):695-699

[59] Stach, B. A. (2010). Clinical audiology: An Introduction (2nd ed.). Clifton Park, NY: Delmar.

[60] Rossini PM, Barker AT, Berardelli A, Caramia MD, Caruso G, Cracco RQ, et al. Noninvasive electrical and magnetic stimulation of the brain, spinal cord and roots: basic principles and procedures for routine clinical application. Report of an IFCN committee. Electroencephalogr Clin Neurophysiol. 1994; 91:79–92

[61] Alcorn, S. (1932). The Tadoma method. Volta Review, 34:195-198

[62] Reed, C.M., Rabinowitz, W.M., Durlach, N.I., Braida, L.D., Conway-Fithian, S. & Schultz, M.C. (1985). Research on the Tadoma method of speech communication. J Acoust Soc. Am., 77(1):247-257

[63] Boersma, P. & Weenink, D. (2013). Praat: doing phonetics by computer. Computer program, Version 5.3.42, retrieved 2 March 2013 from http://www.praat.org/.

[64] Cabeza (2002). Hemispheric Asymetry Reduction in Older Adults: The HAROLD Model. Psychology and aging. 17:85-100.

[65] Füllgrabe C. (2013). Age-dependent changes in temporal-fine-structure processing in the absence of peripheral hearing loss. Am. J. Audiol. 22:313–315

[66] Möttönen, R. & Watkins, K.E. (2009). Motor representations of articulators contribute to categorical perception of speech sounds. J. Neuroscience, 29(31):9819-9825

[67] Gelfand S. A., Piper N. & Silman S. (1985). Consonant recognition in quiet as a function of aging among normal hearing subjects. J. Acoust. Soc. Am. 78:1198–1206

[68] Gelfand S. A., Piper N. & Silman S. (1986). Consonant recognition in quiet and in noise with aging among normal hearing listeners. J. Acoust. Soc. Am. 80:1589–1598

[69] 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

[70] Stevenson, R.A., Nelms, C., Baum, S.H., Zurkovsky, L., Barense, M.D., Newhouse, P.A. & Wallace, M.T. (2015). Deficits in audiovisual speech perception in normal aging emerge at the level of whole-word recognition. Neurobiol Aging, 36(1): 283-291

[71] Grossman M., Cooke A., DeVita C., Chen W., Moore P., Detre J., et al. . (2002). Sentence processing strategies in healthy seniors with poor comprehension: an fMRI study. Brain Lang. 80, 296–313

[72] Stebbins, G.T., Carrillo, M.C., Dorman, J., Dirksen, C., Desond, J., Turner, D.A., Bennett, D. A., Wilson, R. S., Glover, G., & Gabrieli, D.E. (2002). Aging effects on memory encoding in the frontal lobes. Psychol. Aging, 17:44–55

[73] Reuter-Lorenz, P., Jonides, J., Smith, E. S., Hartley, A., Miller, A., Marshuetz, C., & Koeppe, R.A. (2000). Age differences in the frontal lateralization of verbal and spatial working memory revealed by PET. J. Cogn. Neurosci., 12:174–187

[74] Madden, D. J., Turkington, T. G., Provenzale, J. M., Denny, L. L., Hawk, T. C., Gottlob, L. R. & Coleman, R.E. (1999). Adult age differences in functional neuroanatomy of verbal recognition memory. Hum. Brain Mapp., 7:115–135

[75] Grady, C.L., Maisog, J.M., Horwitz, B., Ungerleider, L.G., Mentis, M.J., Salerno, J.A., Pietrini, P., Wagner, E. & Haxby, J.V. (1994) Age-related changes in cortical blood flow activation during visual processing of faces and location. Journal of Neuroscience, 14:1450–1462

[76] Berlingeri M., Danelli L., Bottini G., Sberna M., Paulesu E. (2013). Reassessing the HAROLD model: is the hemispheric asymmetry reduction in older adults a special case of compensatory-related utilisation of neural circuits? Exp. Brain Res. 224, 393–4

DISCUSSION GÉNÉRALE

Dans cette dernière partie, nous allons discuter des différents résultats obtenus dans nos cinq expériences au travers des trois grandes caractéristiques de la perception de la parole que nous avons évoquées dans la partie théorique : sa nature *multisensorielle*, *motrice* et *prédictive*. Les résultats ont tout d'abord été interprétés dans un cadre spécifique lié à l'expérience réalisée, c'est ce que nous avons montré dans chacun des articles présentés dans la partie expérimentale (partie A – perception audio-tactile de la parole ; partie B – perception audio-visuelle linguale de la parole ; partie C – perception audio-visuelle de nos propres productions de parole ; partie D – perception multisensorielle de la parole au cours du vieillissement). Nous allons à présent rappeler les résultats principaux de chacune des expériences réalisées, puis nous détaillerons ce que chacune de ces études peut apporter dans la compréhension des mécanismes qui sous-tendent la nature multisensorielle, motrice et finalement prédictive de la perception de la parole.

DISCUSSION GÉNÉRALE – A

RAPPEL DES PRINCIPAUX RÉSULTATS

1. Etudes EEG sur la perception audio-tactile de la parole (Etudes 1a et 1b)

Dans ces deux expériences, nous avons analysé les potentiels évoqués auditifs N1 et P2 lors de la présentation de syllabes (/pa/ *vs.* /ta/ (Etude 1a) ou /pa/ *vs.* /ta/ *vs.* /ka/ (Etude 1b)) unimodales (auditives (A), visuelles (V) ou tactiles (T)) et bimodales (audio-visuelles (AV) ou audio-tactiles (AT)). Ces études avaient pour but de tester les mécanismes d'intégration précoce d'informations issues de sources sensorielles connues (visuelles) ou non (tactiles) avec un signal auditif de parole et ce en fonction de la saillance perceptive tactile et visuelle du stimulus.

D'un point de vue comportemental, les résultats montrent une meilleure reconnaissance des stimuli auditifs, audio-visuels et audio-tactiles par rapport aux stimuli visuels et tactiles seuls. L'ajout de la modalité visuelle ou tactile n'a pas permis d'améliorer la perception auditive, en revanche, du point de vue des temps de réaction, les stimuli ont été perçus plus rapidement qu'en condition auditive seule. D'un point de vue électrophysiologique, nous avons pu observer une réduction de l'amplitude du PEA N1 en condition AV (Etudes 1a & 1b) par rapport à la condition A. Cette réduction d'amplitude du N1 a également été retrouvée en condition AT (Etude 1b) par rapport à la condition A mais uniquement pour la syllabe /pa/. Aucun effet n'a été observé sur l'amplitude du P2 (Etudes 1a & 1b). D'autre part une diminution de la latence des PEA N1 (Etude 1a) et P2 (Etude 1b) a été trouvée lors d'une présentation bimodale (AV et AH) par rapport à une condition unimodale (A). Finalement, aucune corrélation n'a pu être observée entre les modulations des potentiels évoqués auditifs observées et la saillance perceptive des syllabes présentées (Etude 1b).

2. Etude IRMf sur la perception audio-visuelle linguale de la parole (Etude 2)

Dans cette expérience en IRMf, nous nous sommes intéressés aux réseaux neuronaux activés lors de la perception de stimuli relatifs aux mouvements des lèvres ou de la langue. Des syllabes (/pa/, /ta/ et /ka/) auditives, visuelles et audio-visuelles ont été présentées aux participants lors d'une tâche de perception passive (sans réponses orale ou manuelle). Le signal visuel était relatif soit aux mouvements des lèvres (enregistrés par une caméra vidéo), soit aux mouvements de la langue (enregistrés par une sonde à ultrasons).

Tout d'abord, les résultats de l'expérience comportementale préalable à l'enregistrement IRMf ont montré une meilleure reconnaissance des stimuli auditifs et audio-visuels par rapport aux stimuli visuels seuls. Cependant, en termes de temps de réaction, seul l'ajout de l'information labiale a permis aux participants de percevoir plus rapidement les stimuli auditifs. D'autre part, les résultats neurofonctionnels ont montré un large réseau commun activé par toutes les conditions comprenant principalement le pSTS bilatéral, les régions auditives bilatérales et le cortex prémoteur gauche. Par comparaison des deux modalités audio-visuo-labiale et audio-visuo-linguale, une activation plus importante des régions

sensorielles auditives et visuelles a été observée pour les stimuli relatifs aux mouvements des lèvres tandis que la vue des mouvements de la langue a entraîné une activation plus importante des aires motrices et somatosensorielles. De plus, une corrélation a pu être mise en évidence entre les scores de reconnaissance visuelle et l'activité BOLD observée dans le cortex prémoteur gauche lors des conditions de présentations relatives à la langue tandis qu'une corrélation apparaissait entre les scores de reconnaissance visuelle et l'activité BOLD observée dans les régions visuelles lors des conditions de présentations relatives aux lèvres.

3. Etude EEG sur la perception de nos propres actions de parole (Etude 3)

Dans cette étude en EEG, nous avons analysé les potentiels évoqués auditifs N1 et P2 lors de la présentation de syllabes (/pa/, /ta/ et /ka/) auditives, visuelles et audio-visuelles. La moitié des stimuli était relative aux productions auditives et/ou visuelles du participant tandis que l'autre moitié était relative à celles d'un locuteur inconnu (apparié en genre). Afin de préciser la provenance visuelle ou auditive d'un éventuel effet du locuteur (soi ou autrui), nous avons ajouté une condition audio-visuelle incongruente composée du signal auditif du participant et du signal visuel de son binôme pour une même syllabe prononcée. Nous avons testé les mécanismes d'intégration en utilisant le modèle additif (AV ≠ A+V) et un possible effet du locuteur en nous intéressant aux productions auditives et visuelles propres au participant ou relatives au locuteur inconnu.

D'un point de vue comportemental, nos résultats ont montré une meilleure reconnaissance des stimuli auditifs et audio-visuels par rapport aux stimuli visuels seuls, mais sans effet du locuteur (soi *vs.* autrui). D'un point de vue électrophysiologique, nous avons pu observer une amplitude du PEA P2 plus réduite lors d'une condition bimodale (indépendamment du locuteur) par rapport à la somme des conditions unimodales (A+V). De plus, un effet du locuteur a pu être observé sur la latence du PEA N1 avec une facilitation temporelle des traitements auditifs lorsque le participant voyait ses propres gestes articulatoires par comparaison avec ceux d'un locuteur inconnu. Cette modulation de la latence liée aux productions visuelles du participant observée sur le PEA N1 apparait corrélée négativement avec les scores de reconnaissance visuelle.

4. Etude TMS sur la perception multisensorielle de la parole au cours du vieillissement (Etude 4)

Dans cette étude, nous avons voulu clarifier le rôle du cortex prémoteur ventral (PMv) droit et gauche lors de la perception uni- et multisensorielle de syllabes (/pa/, /ta/ et /ka/). Nous avons cherché à inhiber, à l'aide d'une procédure de stimulation magnétique transcrânienne répétée, le PMv gauche ou droit des participants, puis nous avons testé leurs capacités à identifier correctement et le plus rapidement possible les stimuli présentés dans plusieurs conditions : auditive seule (A), visuelle seule (V), audio-visuelle (AV), tactile seule (T) et audio-tactile (AT). Lors de cette étude, nous avons également testé le rôle du PMv dans les mécanismes de perception au cours du vieillissement.

Indépendamment du site de stimulation, nous avons observé une reconnaissance plus rapide et plus précise des syllabes lorsque les modalités visuelle et tactile étaient ajoutées au signal auditif. De plus, l'intégration audio-visuelle et audio-tactile est apparue stable au cours du vieillissement. D'autre part, nous avons montré une réduction de l'asymétrie hémisphérique du PMv liée à l'âge en condition auditive seule, le PMv droit étant plus

recruté au cours du vieillissement. Finalement, un dernier résultat montre un effet du PMv gauche durant la phase d'apprentissage de la tâche d'identification syllabique.

DISCUSSION GÉNÉRALE – B

LA NATURE *MULTISENSORIELLE* DE LA PERCEPTION DE LA PAROLE

Au travers des expériences réalisées durant cette thèse, nous avons montré que la parole n'était pas seulement auditive et visuelle. En effet, nous avons pu observer que nous étions capables de traiter des informations de parole issues de modalités peu utilisées ordinairement pour comprendre un message linguistique. Nous avons également montré qu'en plus de traiter indépendamment chacune de ces modalités, nous étions capables de les intégrer au signal auditif pour en faciliter le décodage, et ce, même au cours du vieillissement. Nos travaux ont donc permis d'élargir la notion de « multisensorialité de la parole » et d'approfondir nos connaissances des mécanismes cérébraux qui sous-tendent ces traitements spécifiques de la parole. Seuls les résultats comportementaux seront discutés dans cette partie sur la multisensorialité de la parole, les mécanismes cérébraux seront détaillés dans les parties suivantes.

1. La parole peut être perçue/traitée au travers de différentes modalités

Nous nous sommes intéressés aux traitements d'informations habituellement peu utilisées pour comprendre la parole, telles que la perception tactile des gestes articulatoires ou bien la perception visuelle des mouvements de la langue.

Nous avons dans un premier temps confirmé la bonne reconnaissance des stimuli lors des conditions de présentation auditive, avec des scores corrects d'identification supérieurs à 90% pour toutes nos études, que les stimuli soient prononcés en direct par l'expérimentatrice (Etudes 1a, 1b et 4) ou issus de vidéos combinées à une image fixe du locuteur (lèvre ou langue ; Etudes 2 et 3), qu'ils soient prononcés par un inconnu ou par le sujet lui-même (Etude 3), et ce aussi bien pour des participants adultes jeunes que séniors (étude 4).

Par la suite, les résultats relatifs à nos premières études sur la perception tactile de la parole (voir partie Expérimentale – B) ont montré que les participants étaient capables, sans aucune connaissance préalable de la méthode Tadoma, d'identifier les syllabes /pa/, /ta/ et /ka/ grâce à la récupération des informations tactiles relatives aux mouvements du conduit vocal (principalement les lèvres et la mâchoire ici). Bien que le taux de reconnaissance en perception tactile seule soit plus faible que celui en perception auditive, il est comparable à celui obtenu en perception visuelle seule et supérieur à 75%. D'autre part, la syllabe /pa/ connue pour être plus saillante visuellement que /ta/ et /ka/ a également été mieux reconnue de façon tactile. En effet, le geste articulatoire nécessaire pour produire une consonne bilabiale est très perceptible au toucher (des résultats similaires ont été obtenus par Sato et collègues, 2010).

Lors de notre étude sur la perception des mouvements de la langue (voir partie Expérimentale – C), les résultats ont montré que les participants étaient également capables d'identifier les syllabes /pa/, /ta/ et /ka/ uniquement sur la base des informations visuelles

relatives aux mouvements de la langue. Bien que le taux de reconnaissance soit plus faible que celui des stimuli visuels relatifs aux mouvements des lèvres (du fait d'une plus grande expérience visuelle), ils restent supérieurs à la chance et cohérents avec les précédentes études comportementales portant sur le même sujet (Katz et Mehta, 2015; d'Ausilio et al., 2014; Badin et al., 2010).

Pris ensemble, ces résultats démontrent que la parole n'est pas qu'auditive et/ou visuelle et suggèrent qu'en l'absence de connaissances et d'expériences relatives à une modalité donnée, nous sommes néanmoins capables d'utiliser de telles sources d'information inhabituelles. Il est possible de supposer que la référence aux compétences motrices joue un rôle dans ce processus : nous discuterons plus avant de ce point dans la partie C de la discussion.

2. Ces différentes modalités peuvent être intégrées pour améliorer la perception auditive

Nous avons également observé une intégration de ces différentes modalités avec le signal auditif de parole. Nos études montrent que, par la vision de mouvements de la langue ou la perception tactile des gestes d'un locuteur, l'ajout d'une modalité, connue ou non, au signal auditif permet d'améliorer la perception de la parole.

D'un point de vue comportemental, nous avons ainsi pu confirmer dans un premier temps une amélioration de la reconnaissance des stimuli (Etude, 4) ainsi qu'une facilitation temporelle (Etudes 1a, 2 et 4) du traitement auditif lors de l'ajout d'informations visuelles sur les gestes articulatoires disponibles (Reisberg et al., 1987). Des processus similaires ont également été observés en perception audio-tactile en condition d'écoute confortable chez des participants normo-entendants et normo-voyants n'ayant aucune connaissance de la méthode Tadoma (Etudes 1a, 1b et 4). Ces résultats sont cohérents avec les précédentes études sur la perception audio-tactile de la parole (Gick et al., 2008 ; Sato et al., 2010) et montrent qu'une intégration de ces deux modalités peut se faire malgré le peu de connaissance dont nous disposons sur l'une des deux modalités.

Du fait d'un effet plafond pour la condition auditive seule (scores de reconnaissance supérieurs à 98 %) dans quatre de nos cinq études, l'ajout d'une seconde modalité, qu'elle soit visuelle (Etudes 1a, 1b, 2 et 3) ou tactile (Etude 1a, 1b) n'a pas toujours permis d'améliorer la reconnaissance déjà quasi-parfaite des stimuli auditifs. Cependant, la présence d'informations sensorielles supplémentaires a accéléré la reconnaissance des syllabes auditives. Des temps de réactions plus rapide ont en effet pu être observés pour les conditions audio-tactiles (1a et 4) par rapport aux conditions auditives seules.

Un résultat non attendu a été observé lors de la présentation de stimuli audio-visuels relatifs aux mouvements de la langue. En effet, bien que nous ayons retrouvé un taux de réponses correctes supérieur à la chance lors de la perception visuelle de la langue, l'ajout de ces informations ne semble pas faciliter temporellement le traitement auditif. Ces résultats sont contradictoires avec des temps de réaction plus rapides pour la condition audio-visuo-linguale comparée à la condition auditive seule observés par d'Ausillio et collègues (2014). La différence entre nos résultats provient certainement en grande partie des paramètres expérimentaux utilisés. D'Ausillio et collègues ont en effet ajouté une ligne rouge sur l'image du contour de la langue, favorisant ainsi la reconnaissance visuelle de la langue. D'autre part, ils ont utilisé un nombre plus important d'essais, conduisant peut-être à un effet

d'apprentissage plus fort pour les mouvements visuels de la langue. Un dernier paramètre peut rentrer en compte dans ces différentes observations : le calcul des temps de réactions. Nous avons effectivement calculé nos temps de réaction à partir du début du signal acoustique et non à partir du début du mouvement visuel avec une anticipation visuelle très différente entre les mouvements labiaux (forte anticipation) et les mouvements linguaux (faible anticipation). De fait, nos résultats montrent une reconnaissance moins bonne et plus lente des stimuli audio-visuo-linguaux par rapport à la condition auditive seule. Cela suggère que la vue des mouvements de la langue vient perturber et ralentir le processus de décision final, même lorsque les signaux auditif et visuel sont congruents. Une expérience en EEG sur la perception audio-visuo-linguale est en cours d'analyse et pourrait permettre de nous éclairer sur les mécanismes d'intégration mis en œuvre pour percevoir ces stimuli et vérifier ainsi si la vue des mouvements de la langue pourrait ou non faciliter le traitement auditif de façon plus précoce.

Finalement, se percevoir soi-même ne semble pas améliorer l'intégration audio-visuelle des stimuli relatifs à nos propres productions. Contrairement à Tye-Murray et collègues (2012, 2014) et Aruffo et Shore (2012), nos résultats ne montrent pas de scores perceptif plus élevés lors de la vue de nos propres mouvements labiaux ni d'avantage auditif à percevoir notre propre signal auditif de parole. Cependant, un effet plafond des scores de reconnaissance auditive et audio-visuelle (>97%) de notre étude suggère que la tâche était peut-être trop facile pour permettre de faire ressortir un éventuel effet du soi. L'ajout d'un bruit blanc aurait pu permettre d'éviter cet effet plafond, cependant cet ajout aurait posé des difficultés pour l'analyse EEG. D'autre part, d'autres équipes (Schwartz et Savariaux, 2001 par exemple) ont également échoué à retrouver cet avantage comportemental à percevoir nos propres productions. Notre étude a toutefois permis de montrer un effet de soi dans les mécanismes d'intégration précoces liés à la vue de nos propres productions visuelles, nous en discuterons plus avant dans la partie D de cette discussion.

3. Conservation des mécanismes d'intégration multisensoriels avec l'âge

De façon très intéressante, notre quatrième étude nous a permis de montrer qu'en dépit d'une baisse générale des temps de réaction avec l'âge, les intégrations audio-visuelle et audio-tactile étaient conservées, avec une amélioration des traitements de la parole en présence d'un signal visuel (Laurienti et al., 2006) ou tactile cohérent avec le signal auditif.

Ces résultats sont en adéquation avec les précédentes études sur le vieillissement qui montrent tout d'abord que les personnes âgées ont des difficultés pour identifier des syllabes (Strouse et al., 1998 ; Bilodeau-Mercure et al., 2015) ou des phonèmes (Gordon-Salant, 1986) particulièrement lorsque du bruit est ajouté au signal auditif ou que le signal de parole est déformé (Gelfand et al., 1985, 1986 ; Sommers et al., 2005 ; Füllgrabe, 2013 ; Stevenson et al., 2015). Cependant une augmentation du gain audio-visuel chez les séniors a également été retrouvée (Tye-Murray et al., 2010 ; Sekiyama et al., 2014) faisant ainsi écho au principe d'efficacité inverse (« inverse effectiveness principle ») qui stipule que plus une des modalités est dégradée plus l'intégration multisensorielle est forte. Or nos mesures montrent que seule l'acuité auditive décroît avec l'âge de nos participants, suggérant ainsi que les effets liés à une perte auditive due à l'âge pourraient être limités grâce à l'apport d'informations supplémentaires provenant d'autres entrées sensorielles comme le visuel ou le tactile. Cela pourrait également suggérer une mise en place de mécanismes cognitifs supplémentaires, notamment un possible recrutement du système

moteur pour pallier le déclin des traitements auditifs, nous allons y revenir dans la partie suivante (partie C – 2).

DISCUSSION GÉNÉRALE – C

LA NATURE *SENSORI-MOTRICE* DE LA PERCEPTION DE LA PAROLE

Un des grands débats actuel est de mieux comprendre l'implication des aires motrices de la parole dans les mécanismes de compréhension. De manière globale, cette thèse a permis d'envisager une implication plus ou moins importante du système moteur dans toutes les modalités de présentation que nous avons testées. Nous avons effectivement proposé des stimuli mettant en avant les caractéristiques articulatoires de la parole, impliquant par exemple la perception visuelle des mouvements de la langue ou bien la perception tactile des gestes du conduit vocal. Du fait d'une possible utilisation accrue du système moteur lors de conditions de perception difficiles (Binder et al. 2004; Callan et al., 2004; Ojanen et al., 2005; Wilson et Iacoboni, 2006 ; Zekveld et al., 2006 ; Skipper et al., 2007 ; d'Ausilio et al., 2011), nous avons focalisé notre attention sur un sens peu utilisé pour percevoir de la parole (le tactile), un articulateur habituellement peu visible en lecture labiale (la langue), et une population vieillissante (impliquant notamment des sujets de plus de 60 et jusqu'à 78 ans). Nous avons également testé l'utilisation de nos connaissances/représentations motrices en présentant des stimuli relatifs aux propres productions du participant. Et nous nous sommes finalement intéressés aux réseaux neuronaux et/ou aux mécanismes d'intégration précoce de ces différentes modalités spécifiques, cherchant ainsi de façon indirecte une possible implication du système moteur lors de la perception multisensorielle de la parole.

A travers les expériences menées durant cette thèse, nous avons ainsi pu confirmer l'activation des régions motrices et somatosensorielles, particulièrement lors de la présentation de stimuli visuels dont nous avons une grande expérience motrice mais une expérience visuelle moindre. Nous avons en outre observé une modulation de ces régions (PMv notamment) liée selon nous à la compensation d'un déficit sensoriel lié à l'âge. Nos résultats nous ont également conforté dans l'idée d'une possible connexion entre les régions sensorielles et motrices du fait d'une facilitation temporelle précoce liée à la vue de nos propres gestes de parole. Finalement, de façon plus globale, nos résultats nous ont amené à envisager une utilisation du système moteur pour traiter une ou plusieurs modalités, qu'elles nous soient familières ou complètement nouvelles et inattendues, afin de faciliter le traitement de la parole.

1. Activation des régions motrices

Notre expérience en IRMf sur la perception de stimuli de parole relatifs aux mouvements des lèvres ou de la langue (Etude 2) a permis de montrer une activation d'un réseau commun à toutes nos modalités de présentation, composé de régions sensorielles, motrices et intégratives, en accord avec l'hypothèse d'un rôle fonctionnel du système moteur dans les mécanismes de perception de la parole. L'activation du PMv gauche que nous avons observée est aussi cohérente avec les précédentes études sur la perception auditive, visuelle et audio-visuelle de stimuli de parole (par exemple Calvert et Campbell, 2003; Callan et al., 2003, 2004; Möttonen et al., 2004; Wilson et al., 2004; Ojanen et al., 2005; Pekkola et al.,

2005; Skipper, Nusbaum et Small, 2005; Pulvermüller et al., 2006; Wilson et Iacoboni, 2006; Skipper et al., 2007; Callan et al., 2010; Tremblay et Small, 2011).

D'autre part, nous avons également pu observer une activation plus importante des régions motrices et somatosensorielles lors de la perception des stimuli visuels linguaux par rapport aux stimuli visuels labiaux. Compte tenu de la difficulté que représente la perception des mouvements de la langue et au vu des résultats comportementaux observés notamment lors de l'ajout de ces informations au signal auditif de parole, cette activation plus importante du système moteur parait cohérente avec les études montrant un recrutement fonctionnel plus important des régions motrices lors de tâches de perception rendues difficiles (Binder et al. 2004; Zekveld et al., 2006 ; Callan et al., 2004; Wilson et Iacoboni, 2006 ; Ojanen et al., 2005; Skipper et al., 2007). Nos résultats suggèrent ainsi que cette tâche d'identification de stimuli inhabituels pourrait avoir induit une simulation interne des mouvements de la langue de la part des participants afin d'améliorer la perception des stimuli. Cela pourrait expliquer notamment les temps de réactions plus lents observés lors de la perception audio-visuelle des mouvements linguaux.

2. Modulation du système moteur

Lors d'une stimulation inhibitrice du cortex prémoteur ventral (PMv) gauche et droit (Etude 4), nous avons constaté un changement dans la latéralisation hémisphérique de cette région motrice en fonction de l'âge des participants lors d'une tâche d'identification de syllabes. Dans cette étude, nous avons pris le parti de considérer la voie dorsale (ou audio-motrice, incluant le PMv) comme étant plutôt latéralisée à gauche chez le jeune adulte (Hickok et Poeppel, 2007). Or nos résultats montrent qu'au cours du vieillissement, le cerveau recruterait des régions supplémentaires, ici le PMv droit, pour compenser un possible déclin des traitements auditifs. Ce résultat fait écho au modèle HAROLD (Hemispheric Asymmetry Reduction in OLDer adults ; Cabeza, 2002) qui propose, en s'appuyant sur un large ensemble de données expérimentales, qu'il y ait une diminution de la latéralisation du cortex préfrontal, c'est-à-dire un recrutement plus bilatéral des régions correspondantes au cours du vieillissement, notamment lors de tâches de mémoire épisodique. Nos résultats permettent donc de généraliser ce principe de recrutement bilatéral à des régions corticales et dans des types de tâches encore inexplorés dans ce contexte, suggérant que cette tendance vers la bilatéralisation des traitements au cours du vieillissement pourrait également se retrouver au niveau du système moteur.

De façon intéressante, ce changement dans l'asymétrie hémisphérique n'est retrouvé que pour la condition auditive seule. Cet effet est cohérent avec nos résultats sur les tests d'acuité qui montrent que seules les capacités auditives des participants sont corrélées avec l'âge. Dès lors qu'une modalité non dégradée, qu'elle soit visuelle ou tactile, est ajoutée au signal auditif, l'effet disparait. Cela suggère que le système moteur aurait un rôle de soutien plutôt qu'un rôle majeur dans les mécanismes de perception et d'intégration de la parole. Les études précédentes de TMS inhibitrice sur la perception de la parole ont en effet déjà montré une perturbation des capacités du sujet lors de tâches simples d'identification syllabique mais uniquement lors de la présentation de stimuli auditifs bruités ou ambigus (Meister et al., 2007 ; Möttönen et Watkins, 2009 ; d'Ausilio et al., 2009, 2011, 2012 ; Sato et al., 2009).

3. Possible utilisation de nos représentations motrices

Alors que percevoir nos propres productions n'a conduit à aucun avantage ou désavantage perceptif d'un point de vue comportemental, contrairement à Tye-Murray et collègues (2012) ou Aruffo et Shore (2012), notre étude des signaux EEG (Etude 3) nous a permis de démontrer un avantage temporel du traitement auditif lors des mécanismes d'intégration précoce de la parole. Cet avantage serait lié à la vue des productions du participant et serait négativement corrélé à la saillance perceptive des stimuli.

Ainsi, sans fournir de preuves concrètes d'un rôle causal du système moteur lors des mécanismes de perception de notre propre parole audio-visuelle, cet effet précoce du soi suggèrerait fortement une comparaison entre les représentations motrices du participant et les signaux auditifs et visuels entrants. Cette comparaison serait d'autant plus importante que la saillance visuelle des stimuli est faible. Une correspondance quasi parfaite entre les signaux entrants et les représentations motrices du participant pourrait conduire à une sélection plus réduite des candidats phonétiques et ainsi accélérer les processus de traitement du signal auditif. Ces résultats viennent en appui de la théorie du codage commun (Prinz, 1997 ; Hommel et al., 2001) selon laquelle il existerait, à un niveau d'abstraction suffisant, un cadre commun de référence qui permettrait d'intégrer les informations sensorielles et les représentations motrices. Ils sont cohérents avec le modèle d'analyse par synthèse de Skipper et collègues (2007).

4. Utilisation de nos connaissances motrices en toutes situations ?

Finalement, la question intrinsèque à toutes nos études était de comprendre comment nous sommes capables de traiter une modalité particulière alors que nous n'avons pas ou peu d'expérience/connaissances sur cette modalité. A travers des résultats suggérant une simulation interne des gestes moteurs lors de la vue des mouvements de la langue, une compensation motrice pour pallier un déficit du traitement auditif lors du vieillissement ou encore des mécanismes d'intégration précoce des informations audio-tactiles, nous pourrions envisager une utilisation de nos représentations motrices dans des situations de perceptions complexes et/ou difficiles. Cette hypothèse est tout à fait cohérente avec le modèle de perception de la parole développé par Scott et Rauschecker (2009) par exemple. D'autre part, le fait que nous ayons trouvé une facilitation temporelle des processus de traitement auditif étroitement liée à nos propres connaissances motrices et corrélée négativement à la saillance perceptive suggèrerait une utilisation accrue du système moteur en fonction pourraient également suggérer une utilisation accrue de nos connaissances motrices en fonction de la difficulté de la tâche. Une boucle sensori-motrice pourrait être activée afin de faciliter le traitement auditif à travers des mécanismes prédictifs que nous allons développer dans la partie suivante.

DISCUSSION GÉNÉRALE – D

LA NATURE *PREDICTIVE* DE LA PERCEPTION DE LA PAROLE

Nous avons consacré trois de nos expériences à l'étude des mécanismes prédictifs de la perception de la parole (Etudes 1a, 1b et 3) notamment à travers l'observation des mécanismes précoces d'intégration multisensorielle. Nous avons ainsi confirmé une réduction de l'amplitude ainsi qu'une diminution de la latence des PEAs N1/P2 lors de la perception de syllabes audio-visuelles. Nous avons également montré que l'intégration des informations auditives et tactiles suivait des mécanismes similaires, désignant ainsi les canaux visuels et tactiles comme sources d'indices prédictifs pour faciliter le traitement auditif, c'est ce que nous allons développer dans une première section. Et enfin, nous avons montré une facilitation temporelle des traitements audio-visuels précoces lorsque les stimuli présentés étaient relatifs aux propres productions visuelles du participant, suggérant une utilisation des représentations motrices et des prédictions sensorielles pour favoriser la perception de la parole. Nous développerons cet aspect dans une seconde section.

1. Prédictions sensorielles (visuelles et tactiles)

Dans chacune de nos études, les informations visuelles ou tactiles relatives au stimulus présenté précèdent le signal auditif. En effet, nos stimuli sont des syllabes de type CV (/pa/, /ta/ ou /ka/) qui débutent chacune par une position mi-ouverte du conduit vocal (mâchoire, langue, lèvres), correspondant à une position neutre. Le mouvement débute par une fermeture du conduit vocal plus ou moins saillante visuellement ou tactilement lors de la production de la consonne cible (pour le « p » : fermeture au niveau des lèvres, pour le « t » : fermeture par un contact entre l'apex de la langue et les alvéoles et pour le « k » : fermeture par un contact entre le dos de la langue et le palais mou (ou velum), le tout accompagné dans les trois cas d'un mouvement de montée de la mâchoire). Ce premier geste articulatoire peut être considéré comme un geste préparatoire (Schwartz et Savariaux, 2014) durant lequel aucun son n'est émis, mais qui véhicule déjà beaucoup d'information sur ce qui va suivre, notamment le lieu d'articulation. Survient ensuite la plosion correspondant à l'ouverture rapide de la cavité buccale accompagnée quelques dizaines de millisecondes plus tard par le démarrage du voisement relatif à la voyelle /a/ produite.

Ainsi, des informations visuelles et tactiles sur la syllabe sont déjà disponibles avant même que la syllabe ne soit audible. Ce phénomène d'anticipation se traduit au plan neurophysiologique par une facilitation des traitements auditifs que nous pouvons observer à travers une modulation de la latence et/ou de l'amplitude des potentiels évoqués auditifs N1/P2. Dans nos expériences, nous avons pu retrouver une réduction de l'amplitude des PEAs N1 (Etudes 1a et 1b) et P2 (Etude 3) et une diminution de la latence des PEAs N1 (Etude 1a) et P2 (Etude 1b) lors de la présentation de stimuli audio-visuels par rapport à une présentation auditive seule. D'autre part, dans notre expérience 1b, la modulation de l'amplitude des PEAs N1/P2 observée en condition audio-visuelle n'est pas apparue dépendante de la saillance perceptive des syllabes présentées, ce qui est cohérent avec les

résultats obtenus par van Wassenhove et collègues (2005). Dans le cadre de la théorie du codage prédictif (Friston, 2005), cette diminution de l'amplitude pourrait être le fruit des prédictions visuelles envoyées dans le cortex auditif pour « préparer » le cortex auditif à l'arrivée d'un signal sonore (Arnal et al., 2009). Le traitement serait alors d'ordre temporel, plutôt lié à l'unification perceptuelle des deux signaux qu'au traitement phonétique du stimulus. Une facilitation temporelle des PEAs N1/P2 (réduction de la latence) serait quant à elle plutôt liée au traitement du lieu d'articulation d'après van Wassenhove et collègues (2005) du fait d'une corrélation entre cette modulation et la saillance perceptive des syllabes. Cependant, bien que nous ayons retrouvé une latence plus réduite en perception audio-visuelle par rapport à une perception auditive seule, nos résultats n'ont montré aucune corrélation avec les scores perceptifs visuels. Les résultats EEG étant extrêmement dépendants des conditions expérimentales, cette absence de corrélation pourrait s'expliquer par une variabilité beaucoup plus forte dans nos études (1a et 1b) du fait d'une prononciation en direct de chacun des stimuli présentés aux participants. Sans pour autant contredire les observations de van Wassenhove et collègues, nos résultats montrent qu'il est important d'interpréter avec précaution les résultats expérimentaux utilisant peu d'exemplaires différents d'un même stimulus.

Nous avons également mis en avant, pour la première fois, une réduction de l'amplitude du PEA N1 (Etude 1b) et une diminution de la latence des PEAs N1 (Etude 1a) et P2 (Etude 1b) lors de la perception de stimuli de parole audio-tactile. Ces nouveaux résultats démontrent une utilisation des informations tactiles pour faciliter/améliorer les traitements auditifs au même titre que le signal visuel de parole.

L'existence de connexions entre les régions auditives et visuelles et auditives et somatosensorielles est compatible avec l'idée d'une possible transmission directe des informations visuelles et tactiles au cortex auditif. Nous avons donc ici une première boucle cortico-corticale qui pourrait être activée : les informations visuelles ou tactiles précédant le signal auditif pourraient être envoyées au niveau des régions auditives afin d'alerter le système auditif de l'arrivée d'un son à traiter (Senkowski et al., 2008). Selon Arnal et collègues (2009, voir Figure 16), il pourrait s'agir là d'une première voie « directe » dont le rôle serait de pré-sélectionner des candidats phonétiques. En parallèle, des prédictions sensorielles (auditives, visuelles ou tactiles dans notre cas) seraient envoyées dans le sillon temporal supérieur (STS) pour être intégrées. De là, une seconde voie « feedback » pourrait être activée afin de renvoyer des erreurs de prédiction au cortex auditif afin de finaliser l'interprétation du signal sonore. Dans l'expérience d'Arnal et al. (2009) une modulation du champ magnétique évoqué M1 (équivalent du PEA N1) pourrait être le reflet de cette seconde voie. Cependant nos données ne sont pas suffisantes pour déterminer avec précision si les mécanismes précoces que nous avons observés sont le résultat de l'une ou l'autre de ces trajectoires neuroanatomiques, directe ou indirecte.

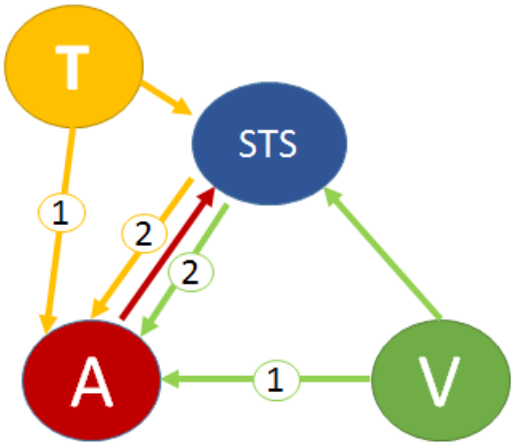

Figure 16 : Boucle multisensorielle : A : Aires auditives, V : aires visuelles, STS : aires intégratives, T : régions somatosensorielles. En plus des prédictions visuelles envoyées simultanément au cortex auditif (voie directe, 1) et au STS selon le modèle d'Arnal et al., 2009 et de la voie « feedback » (2)véhiculant les erreurs de prédiction, nous faisons la supposition d'une boucle parallèle relative aux informations tactiles. Elle serait également basée sur ces deux voies, l'une directe (1) permettant un envoi rapide des prédictions tactiles au cortex auditif et l'autre feedback (2) permettant le renvoi des erreurs de prédictions après intégration dans le STS.

2. Prédictions motrices

Pour l'instant, seules les prédictions sensorielles ont été mises en avant pour faciliter/améliorer les traitements auditifs, or les résultats de notre troisième expérience en EEG sur la perception de soi suggèrent qu'une boucle motrice pourrait également intervenir dans les mécanismes prédictifs de la perception de la parole.

Le système moteur, nous l'avons vu, est fortement impliqué dans la perception d'un mouvement humain, qui plus est lorsque cette action a été produite par le sujet lui-même. La majorité des études montrent que nous sommes meilleurs pour reconnaître nos propres actions (Beardworth et Buckner, 1981 ; Knoblich et al., 2001 ; Loula et al., 2005). Cependant, bien que certaines théories et certains modèles neurobiologiques attribuent un rôle au système moteur dans les mécanismes de perception de la parole, son utilisation reste mineure si les conditions expérimentales ne sont pas complexes ou difficiles et les résultats sur la perception de nos propres productions restent mitigés.

Or notre troisième étude sur les mécanismes d'intégration audio-visuelle de nos propres productions nous a permis de montrer une facilitation temporelle des traitements auditifs précoces lors de la perception visuelle de nos propres gestes articulatoires. Ce nouveau résultat suggère une association entre les représentations sensorielles et motrices, avec un meilleur appariement entre nos propres signaux de parole et les connaissances procédurales motrices que nous en avons. Selon le modèle d'analyse par synthèse de Skipper et collègues (2007), des hypothèses phonémiques relatives à l'entrée sensorielle reçue seraient envoyées au niveau du gyrus frontal inférieur pour être appariées aux buts articulatoires. De là le système moteur va simuler les commandes motrices sous-jacentes afin de renvoyer des prédictions des conséquences sensorielles qui vont être comparées avec l'entrée. Cette facilitation temporelle que nous avons observée semble étroitement liée à nos connaissances motrices et pourrait s'interpréter, selon le modèle d'analyse par synthèse, comme le reflet des prédictions des conséquences sensorielles renvoyées par le système moteur au cortex auditif.

D'autre part, dans le cas d'une entrée audio-visuelle, Skipper et collègues proposent une première étape d'intégration multisensorielle (voir Figure 17A), les hypothèses phonémiques envoyées seraient alors multisensorielles et relatives à la fusion des informations auditives et visuelles. Ainsi, les prédictions des conséquences sensorielles renvoyées sous formes de copies d'efférence seraient elles aussi relatives à ce premier stade d'intégration. Cependant, nos résultats montrent que l'avantage du soi observé serait principalement dû aux informations visuelles contenues dans le signal audio-visuel entrant. Nous pourrions donc envisager l'existence d'une boucle parallèle dans laquelle les informations visuelles précédant le signal auditif seraient envoyées directement au système moteur (voir Figure 17B). Cette boucle serait plus rapide car ne passerait pas par l'étape d'intégration multisensorielle, elle permettrait donc d'expliquer à la fois la facilitation temporelle et le fait que les informations visuelles en soient responsables. Toutefois, nos données ne nous permettent que de poser des hypothèses, et les mécanismes mis en œuvre sont probablement beaucoup plus complexes.

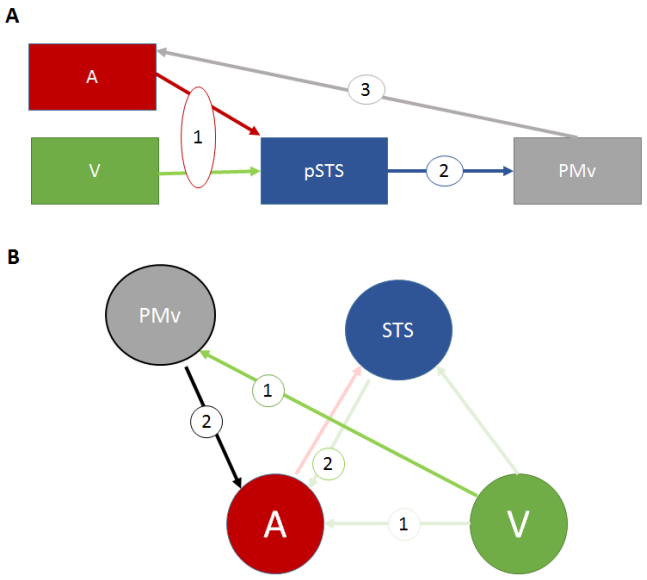

Figure 17 : Prédictions motrices. A – Schéma représentant le modèle de Skipper et al., 2007 avec une première étape (1) d'intégration Audio-Visuelle au niveau du pSTS (partie postérieur du sillon temporal supérieur), puis l'envoie des hypothèses phonémiques multisensorielles (2) au PMv (cortex prémoteur ventral) et finalement un renvoi des conséquences sensorielles sous forme de copies d'efférences au cortex auditif (2). B – Possibilité d'une boucle motrice parallèle, plus rapide grâce à l'envoi des informations visuelles directement au PMv en parallèle de la boucle classique vue en Figure 16.

Notre raisonnement semble cependant cohérent avec le modèle d'Arnal et collègues (2009) qui, sans mentionner le système moteur, propose déjà plusieurs voies possibles et plusieurs destinations pour les informations visuelles. En effet, des projections directes sont envisagées en parallèle entre les régions visuelles et auditives d'une part et visuelles et multisensorielles (intégratives) d'autre part. Finalement, les scenarii proposés par Senkowski et collègues (2008) viennent également supporter cette hypothèse avec de possibles interactions ascendantes et descendantes entre les régions uni- et multi-sensorielles voire frontales.

Pour résumer, un modèle multisensori-moteur prédictif pourrait être établi à partir de toutes ces données. Différentes entrées sensorielles ont été proposées et validées comme pouvant interagir avec le signal auditif de parole. Ainsi, nous pouvons imaginer un système de perception de la parole composé d'une entrée auditive, d'une entrée visuelle comprenant les mouvements des lèvres et de la langue et d'une entrée tactile. Toutes les régions cérébrales correspondantes seraient interconnectées, favorisant ainsi la propagation de prédictions nécessaires à l'amélioration et à la facilitation du traitement auditif au travers d'une première boucle sensorielle. A cela pourrait s'ajouter l'existence d'une seconde boucle

sensori-motrice qui permettrait de faire appel à nos connaissances procédurales motrices pour venir soutenir et renforcer les traitements auditifs. Les régions sensorielles communiqueraient alors directement avec le système moteur en envoyant des hypothèses phonémiques qui seraient appariées aux commandes motrices afin de générer des prédictions des conséquences sensorielles relatives aux buts articulatoires hypothétiques. Cette seconde boucle pourrait être activée non seulement lorsque les informations sensorielles sont ambigües ou complexes ou bien pour compenser la dégradation des traitements auditifs avec l'âge, mais également et de manière plus générale lorsque des stimuli de parole, quels qu'ils soient, doivent être traités. Les travaux de Scarbel et collègues (2014) appuient cette hypothèse en montrant que même en condition d'écoute confortable, une boucle sensori-motrice est rapidement activée. Ça serait donc à l'aide de prédictions sensorielles et motrices que nous serions capables de percevoir et de traiter n'importe quelle modalité de parole, de la plus courante à la plus étonnante.

Ce modèle est évidemment extrêmement simpliste face à la complexité des processus que nous souhaitions observer, et plusieurs aspects mériteraient d'être approfondis. Tout d'abord il pourrait être intéressant de confirmer l'activation du système moteur lors de la perception de notre propre parole auditive, visuelle et audio-visuelle. Ainsi une étude en IRMf permettrait de mettre en évidence un réseau d'activation neuronale spécifique. D'autre part, il pourrait être utile de vérifier si cet effet du soi relatif aux productions visuelles est bien dû à notre expérience motrice et non à notre expérience visuelle. Pour cela une nouvelle étude en EEG pourrait être envisagée pour comparer les mécanismes mis en œuvre lors de la perception de nos propres productions et celles d'un ami.

Un second point à éclaircir serait le décours temporel des mécanismes d'intégration audio-visuo-lingual. En effet, nous n'avons pas pu montrer de facilitation ou d'amélioration à percevoir les mouvements de la langue tandis qu'une forte activation des systèmes moteur et somatosensoriel a été observée. Etudier les interactions précoces entre les informations auditives et visuo-linguales pourrait nous permettre de mieux comprendre cet effet.

Une troisième perspective serait d'apporter des précisions quant à la latéralisation de la voie dorsale notamment à travers l'étude plus poussée des mécanismes de préservation de l'intégration audio-visuelle de la parole chez les personnes âgées. Une nouvelle étude par TMS répétitive sur une population sénior et une population de jeunes adultes lors de la perception de stimuli de parole accompagné d'un bruit plus important pourrait fournir de nouveaux éléments pour l'interprétation de cette réduction de l'asymétrie hémisphérique que nous avons observée.

CONCLUSION

Cette thèse avait pour objectif d'approfondir les connaissances déjà existantes sur la nature multisensorielle de la parole, en s'intéressant à des modalités peu communes, comme la perception tactile, la visualisation des mouvements de la langue et la perception de nos propres productions. Nous souhaitions appréhender le système moteur au travers de ces différentes conditions, afin d'affiner la compréhension de ce lien perceptivo-moteur qui est aujourd'hui au cœur des théories de la perception de la parole. Pour combiner tous ces aspects, et afin de mieux comprendre la nature prédictive de la parole, nous avons choisi de nous intéresser aux mécanismes d'intégration précoce de ces différentes modalités avec le signal auditif de parole, et les réseaux neuronaux mis en jeu pour y parvenir.

Nos travaux nous ont permis d'élargir la notion de « multisensorialité de la parole », en mettant en évidence une bonne reconnaissance des syllabes présentées dans les différentes conditions unimodales (tactile, visuelle-lèvre, visuelle-langue, auditive) et multimodales (audio-visuelle et audio-tactile) ainsi qu'une modulation des potentiels évoqués auditifs N1/P2 dans les conditions multisensorielles. D'autre part, à différents niveaux, nos études nous ont permis d'établir un rôle fonctionnel et causal du système moteur en montrant une activation plus importante des régions motrices lors de la perception des mouvements de la langue, un recrutement plus bilatéral du PMv au cours du vieillissement ainsi qu'une facilitation des traitements auditifs précoces lors de la perception de nos propres productions visuelles.

Pris ensemble, nos résultats renforcent l'idée d'un couplage fonctionnel, d'une co-structuration des systèmes de perception et de production de la parole. Les études présentées dans cette thèse appuient ainsi l'existence de connexions entre régions sensorielles, intégratives et motrices permettant la mise en œuvre de processus et traitements multisensoriels, sensorimoteurs et prédictifs lors de la perception et compréhension des actions de parole.

RÉFÉRENCES

Alcorn, S. (1932). The Tadoma method. Volta Review, 34: 195-198.

Alsius, A., Navarra, J., Campbell, R. & Soto-Faraco S. (2005). Audiovisual integration of speech falters under high attention demands. Curr. Biol., 15: 839–843.

Alsius, A., Navarra, J. & Soto-Faraco S. (2007). Attention to touch weakens audiovisual speech integration. Exp. Brain Res., 183: 399–404.

Anderson, A., Parbery-Clark, A., Yi, H.G. & Kraus, N. (2011). A neural basis of speech-in-noise perception in older adults. Ear and Hearing, 32(6): 750-757.

Arnal, L.H., Morillon, B., Kell, C.A., & Giraud, A.L. (2009). Dual neural routing of visual facilitation in speech processing. Journal of Neuroscience, 29(43): 13445–13453.

Arnold, P. & Hill, F. (2001). Bisensory augmentation: A speechreading advantage when speech is clearly audible and intact. British Journal of Psychology, 92: 339-355.

Aruffo, C. & Shore, D.I. (2012). Can you McGurk yourself? Self-face and self-voice in audiovisual speech. Psychological Bulletin and Review, 19: 66–72.

Attina, V. (2005). La langue française parlée complétée (LPC) : production et perception. Thèse de doctorat, Grenoble.

Aziz-Zadeh, L., Iacoboni, M., Zaidel, E., Wilson, S. & Mazziotta, J. (2004). Left hemisphere motor facilitation in response to manual action sounds. European Journal of Neuroscience, 19: 2609–2612.

Baart, M., Vroomen, J., Shaw, K. & Bortfeld, H. (2013). Phonetic information in audiovisual speech is more important for adults than for infants; preliminary findings. In S. Ouni, F. Berthommier, & A. Jesse (Eds.). Proceedings of the International Conference on Auditory-Visual Speech Processing (pp. 61 - 64). Annecy, France: Inria.

Baart, M., Stekelenburg, J. J. & Vroomen, J. (2014). Electrophysiological evidence for speech-specific audiovisual integration. Neuropsychologia, 53: 115–121.

Baart, M. & Samuel, A.G. (2015). Turning a blind eye to the lexicon: ERPs show no cross-talk between lip-read and lexical context during speech sound processing. Journal of Memory & Language, 85: 42-59.

Baart, M., (2016). Quantifying lip-read induced suppression and facilitation of the auditory N1 and P2 reveals peak enhancements and delays. Psychophysiology, 53: 1295–1306.

Badin, P., Tarabalka, Y., Elisei, F. & Bailly, G. (2010). Can you 'read' tongue movements? Evaluation of the contribution of tongue display to speech understanding. Speech Communication, 52: 493-503.

Beauchamp, M. S., Argall, B. D., Bodurka, J., Duyn, J. H. & Martin, A. (2004a). Unraveling multisensory integration: Patchy organization within human STS multisensory cortex. Nature Neuroscience, 7: 1190–1192.

Beauchamp, M. S., Lee, K. E., Argall, B. D. & Martin, A. (2004b). Integration of auditory and visual informations about objects in superior temporal sulcus. Neuron, 41: 809–823.

Beauchamp, M. S. (2005). Statistical criteria in fMRI studies of multisensory integration. Neuroinformatics, 3: 93–114.

Beauchamp, M.S., Nath, A.R. & Pasalar, S. (2010). fMRI-guided transcranial magnetic stimulation reveals that the superior temporal sulcus is a cortical locus of the McGurk effect. Journal of Neuroscience, 30: 2414–2417.

Berlingeri M., Danelli L., Bottini G., Sberna M. & Paulesu E. (2013). Reassessing the HAROLD model: is the hemispheric asymmetry reduction in older adults a special case of compensatory-related utilisation of neural circuits? Exp. Brain Res. 224: 393–410.

Bernier, V. (2012). Percevoir l'autre et moi-même. Mémoire de master, Grenoble.

Bernstein, L.E., Demorest, P.E. & Tucker, M.E. (2000). Speech perception without hearing. Perception & Psychophysics, 62(2): 233-252.

Besle, J., Fort, A., Delpuech, C. & Giard, M.H. (2004). Bimodal speech: early suppressive visual effects in human auditory cortex. European journal of Neuroscience, 20: 2225-2234.

Bilodeau-Mercure, M., Lortie, C.L., Sato, M., Guitton, M.J. & Tremblay, P. (2015). The neurobiology of speech perception decline in aging. Brain Struct. Funct., 220: 979–997.

Binder, J.R., Liebenthal, E., Possing, E.T., Medler, D.A. & Ward, B.D. (2004). Neural correlates of sensory and decision processes in auditory object identification. Nature Neuroscience, 7: 295-301.

Blakemore, S-J. (2003). Deluding the motor system. Conscious Cogn., 12: 647-655.

Boatmann, D.F. (2004). Cortical bases of speech perception: evidence from functional lesion studies. Cognition, 92: 47-65.

Buccino, G., Binkofski, F., Fink, G.R., Fadiga, L., Fogassi, L., Gallese, V., Seitz, R.J., Zilles, K., Rizzolatti, G. & Freund, H.J. (2001). Action observation activates premotor and parietal areas in a somatotopic manner: an fMRI study. European Journal of Neuroscience, 13: 400-404.

Buccino, G., Lui, F., Canessa, N., Patteri, I., Lagravinese, G., Benuzzi, F., Porro, C.A. & Rizzolatti, G. (2004). Neural circuit involved in the recognition of actions performed by non conspecifics: an fMRI study. Journal of Cognitive Neurosciences, 16: 114-126.

Buckner, R.L., Andrews-Hanna, J.R. & Schacter, D.L. (2008). The brain's default network: anatomy, function, and relevance to disease. Annals of the New York Academy of Sciences, 1124: 1–38.

Beardsworth, T. & Buckner, T. (1981). The ability to recognize oneself from a video recording of one's movements without seing one's body. Bulletin of the Psychonomic Society, 18: 19-22.

Blumstein, S.E. & Stevens, K.N. (1979). Acoustic invariance in speech production: evidence from measurements of the spectral characteristics of stop consonants. Journal of the Acoustical Society ofAmerica, 66: 1001-1017.

Bresciani, J-P., Ernst, M.O., Drewing, K., Bouyer, G; Maury, V. & Kheddar, A. (2005). Feeling what you hear: auditory signals can modulate tactile tap perception. Experimental Brain Research, 162(2): 172-180.

Cabeza, R. (2002). Hemispheric Asymetry Reduction in Older Adults: The HAROLD Model. Psychology and aging, 17: 85-100.

Callan, D.E., Jones, J.A., Munhall, K.G., Callan, A.M., Kroos, C. & Vatikiotis-Bateson, E. (2003). Neural processes underlying perceptual enhancement by visual speech gestures. NeuroReport, 14: 2213-2217.

Callan, D.E., Jones, J.A., Munhall, K.G., Callan, A.M., Kroos, C. & Vatikiotis-Bateson, E. (2004). Multisensory integration sites identified by perception of spatial wavelet filtered visual speech gesture information. Journal of Cognitive Neuroscience, 16: 805-816.

Callan, D., Callan, A., Gamez, M., Sato, M.A. & Kawato, M. (2010). Premotor cortex mediates perceptual performance. NeuroImage, 51(2): 844-58.

Calvert, G.A., Bullmore, E., Brammer, M.J., Campbell, R., Williams, S.C., Mc Guire, P.K., Woodruff, P.W., Iversen, S. D. & David, A.S. (1997). Activation of auditory cortex during silent lipreading. Science, 276: 593–596.

Calvert, G.A., Campbell, R. & Brammer, M.J. (2000). Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. Current Biology, 10(11): 649-657.

Calvert, G.A. & Campbell, R. (2003). Reading speech from still and moving faces: the neural substrates of visible speech. Journal of Cognitive Neuroscience, 15: 57–70.

Calvo-Merino, B., Glaser, D.E., Grèzes, J., Passingham, R.E. & Haggard, P. (2005). Action observation and acquired motor skills: an FMRI study with expert dancers. Cerebral Cortex, 15: 1243-1249.

Calvo-Merino, B., Grèzes, J., Glaser, D.E., Passingham, R.E. & Haggard, P. (2006). Seeing or doing? Influence of visual and motor familiarity in action observation. Current Biology, 16: 1905-1910.

Cappe, C., Rouiller, E.M., Barone, P. (2009). Multisensory anatomical pathways. Hear Res., 258: 28–36.

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A. & Ghazanfar, A.A. (2009). The Natural Statistics of Audiovisual Speech. PLoS Comput Biol., 5(7): e1000436.

Cienkowski, K.M. & Carney, A.E. (2002). Auditory-visual speech perception and aging. Ear and Hearing, 23(5): 439-449.

Colin, C., Radeau, M., Soquet, A., Colin, F. & Deltenre, P. (2002). Mismatch negativity evoked by the McGurk-MacDonald effect: evidence for a phonological representation within auditory sensory short term memory. Clinical Neurophysiology, 113(4) : 495-506.

Colin, C. & Radeau, M. (2003). Les illusions McGurk dans la parole : 25 ans de recherches. L'Année Psychologique, 103 : 497-542.

Cornett, R. O. (1994). Adapting Cued Speech to additional languages. Cued Speech Journal V: 19-29.

Cotton, J. C. (1935). Normal "visual hearing". Science. 82: 592-593.

Cross, E.S., Hamilton, A.F. de C. & Grafton, S.T. (2006). Building a motor simulation de novo: Observation of dance by dancers. NeuroImage. 31: 1257-1267.

Cruickshanks, K.J., Wiley, T.L., Tweed, T.S., Klein, B.E., Mares-Perlman, J.A. & Nondahl, D.M (1998). Prevalence of hearing loss in older adults in Beaver Dam, Wisconsin. Am J Epidemiol, 148: 879–886.

Cvejic, E., Kim, J. & Davis, C. (2010). Prosody off the top of the head: Prosodic contrasts can be discriminated by head motion. Speech Communication, 52: 555-564.

Dancer, J., Krain, M., Thompson, C., Davis, P. & Glenn, J. (1994). A cross-sectional investigation of speechreading in adults: effects of age, gender, practice, and education. Volta Rev., 96: 31–40.

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.

d'Ausilio, A., Jarmolowska, J., Busan, P., Bufalari, I. & Craighero, L. (2011). Tongue corticospinal modulation during attended verbal stimuli: priming and coarticulation effects. Neuropsychologia. 49(13): 3670-6.

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): 882-7.

d'Ausilio, A., Bartoli, E., Maffongelli, L., Berry, J. J. & Fadiga, L. (2014). Vision of tongue movements bias auditory speech perception. Neuropsychologia, 63: 85–91.

Diehl, R.L., Lotto, A.J. & Holt, L.L. (2004). Speech Perception. Annual Review of Psychology, 55: 149-179.

Di Pellegrino, G., Fadiga, L., Fogassi, L., Gallese, V. & Rizzolatti, G. (1992). Understanding motor events: a neurophysiological study. Experimental Brain Research, 91: 176-180.

Fadiga, L., Craighero, L., Buccino, G. & Rizzolatti, G. (2002). Speech listening specifically modulates the excitability of tongue muscles: A TMS study. European Journal of Neuroscience, 15: 399-402.

Fischer, C. (1968). Confusions among visually perceived consonnants. Journal of Speech and Hearing Research, 11: 796-804.

Fogassi, L., Ferrari, P.F., Gesierich, B., Rozzi, S., Chersi, F. & Rizzolatti, G. (2005). Parietal lobe: from action organization to intention understanding. Science, 308: 662-667.

Fowler, C.A. (1986). An event approach to the study of speech perception from a direct-realist perspective. Journal of Phonetics, 14: 3-28.

Fowler, C. & Dekle, D. (1991). Listening with eyes and hands: cross modal contributions to speech perception. Journal of experimental psychology: human performance. 17: 816-828.

Fowler, C.A. (1996). Listeners do hear sounds, not tongues. Journal of the Acoustical Society of America, 99: 1730–41.

Friston, K.(2005) A theory of cortical responses. Philos Trans R Soc Lond B Biol Sci., 360: 815–836.

Frith, C.D. & Frith, U. (1999). Interacting minds—A biological basis. Science, 286(5445): 1692–1695.

Füllgrabe, C. (2013). Age-dependent changes in temporal-fine-structure processing in the absence of peripheral hearing loss. Am. J. Audiol. 22: 313–315.

Gagne, J. P., Rochette, A. J., & Charest, M. (2002). ‘‘Auditory, visual and audiovisual clear speech,’’ Speech Commun. 37: 213–230.

Gallese, V., Fadiga, L., Fogassi, L. & Rizzolatti, G. (1996). Action recognition in the premotor cortex. Brain, 119: 593-609.

Gelfand, S.A., Piper, N. & Silman, S. (1985). Consonant recognition in quiet as a function of aging among normal hearing subjects. J. Acoust. Soc. Am., 78: 1198–1206.

Gelfand, S. A., Piper, N. & Silman, S. (1986). Consonant recognition in quiet and in noise with aging among normal hearing listeners. J. Acoust. Soc. Am., 80: 1589–1598.

Getzmann, S., Hanenberg, C., Lewald, J., Falkenstein, M. & Wascher, E. (2015). Effects of age on electrophysiological correlates of speech processing in a dynamic “cocktail-party” situation. Front Neurosciences, 9: 341.

Gentil, M. (1981). Etude de la perception de la parole : Lecture labiale et sosies labiaux. Rapport technique, IBM, France.

Gick, B., Johannsdottir, K., Gibraiel, D., & Muhlbauer, J. (2008). Tactile enhancement of auditory and visual speech in untrained pervceivers. Jasa express letters, 123: 1145-1154.

Gick, B. & Derrick, D. (2009). Aero-tactile integration in speech perception. Nature, 462: 502-504.

Gick, B., Ikegami, Y. & Derrick, D. (2010). The temporal window of audio-tactile integration in speech perception. Jasa express letters, 128: 342-346.

Gordon-Salant, S. (1986). Effects of aging on response criteria in speech-recognition tasks, J Speech Hear Res, 29: 155-162

Gordon-Salant, S. & Fitzgibbons, P.J. (1993). Temporal factors and speech recognition performance in young and elderly listeners. J Speech Hear Res., 36: 1276.

Grabski, K. (2012). Les cartes sensorimotrices de la parole. Thèse de Doctorat, Grenoble.

Grady, C.L., Maisog, J.M., Horwitz, B., Ungerleider, L.G., Mentis, M.J., Salerno, J.A., Pietrini, P., Wagner, E. & Haxby, J.V. (1994) Age-related changes in cortical blood flow activation during visual processing of faces and location. Journal of Neuroscience, 14: 1450–1462

Graux, J., Gomot, M., Roux, S., Bonnet-Brilhault, F., Camus, V. & Bruneau, N. (2013). My voice or yours?An electrophysiological study.Brain Topography, 26: 72–82.

Graux, J., Gomot, M., Roux, S., Bonnet-Brilhault, F., Camus, V. & Bruneau, N. (2014). Is my voice just a familiar voice? An electrophysiological study. Social Cognitive and Affective Neuroscience, 2-5.

Grèzes J., Costes N. & Decéty J. (1998). Top down effect of strategy on the perception of human biological motion: A PET investigation. Cognitive Neuropsychology, 15: 553-582.

Grossman M., Cooke A., DeVita C., Chen W., Moore P., Detre J., Alsop, D. & Gee, J. (2002). Sentence processing strategies in healthy seniors with poor comprehension: an fMRI study. Brain Lang., 80: 296–313.

Guenther, F.H. (1994). A neural network model of speech acquisition and motor equivalent speech production. Biological Cybernetics, 72(1): 43–53.

Guenther, F.H. & Vladusich, T. (2012). A neural theory of speech acquisition and production. Journal of Neurolinguistics, 25(5): 408–422

Guiard-Marigny, T., Tsingos, N., Adjoudani, A., Benoît, C. & Gascuel, M.-p. (1996). 3D Models of the Lips for Realistic Speech Animation. Computer Animation '96 (pp. 80-89). Genève, Suisse.

Haueisen, J. & Knösche, T. R. (2001). Involuntary motor activity in pianists evoked by music perception. Journal of Cognitive Neuroscience, 13: 786–792.

Hickok, G. & Poeppel, D. (2000). Towards a functional neuroanatomy of speech perception. Trends in Cognitive Science, 4(4): 131-138.

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

Hickok, G. & Poeppel, D. (2007).The cortical organization of speech processing. Nature Review Neurosciences, 8: 393-402.

Hommel B., Müsseler J., Aschersleben G. & Prinz W. (2001). The Theory of Event Coding (TEC): A Framework for Perception and Action Planning, Behavioral and Brain Sciences, 24: 849-937.

Howard, R. J., Brammer, M., Wright, I., Woodruff., P. W., Bullmore, E. T. & Zeki, S. (1996). A direct demonstration of functional specialization within motion-related visual and auditory cortex of the human brain. Current Biology, 6: 1015–1019.

Hueber, T., Chollet, G., Denby, B. & Stone, M. (2008). Acquisition of ultrasound, video and acoustic speech data for a silent-speech interface application. Proceedings of International Seminar on Speech Production (Strasbourg, France), 365-369.

Hueber, T. (2009). Reconstitution de la parole par imagerie ultrasonore et vidéo de l'appareil vocal : vers une communication parlée silencieuse. Thèse de doctorat, Paris.

Iacoboni, M., Molnar-Szakacs, I., Gallese, V., Buccino, G., Mazziotta, J.C., Rizzolatti, G., (2005). Grasping the intentions of others with one's own mirror neuron system. PLoS Biology, 3: 529–535.

Ito, T., Tiede, M. & Ostry, D.J. (2009). Somatosensory function in speech perception. Proceedings of the National academy of sciences of the United States of America, 106(4): 1245-1248.

Johansson, R. (1973). Visual perception of biological motion and a model for its analysis. Perception and psychophysics, 14: 201-211.

Jones, J. & Callan, D. E. (2003). Brain activity during audio-visual speech perception: An fMRI study of the McGurk effect. NeuroReport, 14: 1129–1133.

Jordan, T.R. & Thomas, S.M. (2011).When half a face is as good as a whole: effects of simple substantial occlusion on visual and audiovisual speech perception. Atten. Percept. Psychophys., 73: 2270–2285

Jousmäki, V. & Hari, R. (1998). Parchment-skin illusion: sound-biased touch. Current Biology. 8: 190.

Kaganovich, N. & Schumaker, J. (2014). Audiovisual integration for speech during mid-childhood: Electrophysiological evidence. Brain and Language, 139: 36–48.

Kaplan, J.T., Aziz-Zadeh, L., Uddin, L. & Iacoboni, M. (2008). The self across the senses: The neural response to one's own face and voice. Social, Cognitive, and Affective Neuroscience, 3: 218-223.

Katz, W., Campbell, T., Wang, J., Farrar, E., Eubanks, J.C., Balsubramanian, A., Prabhakaran, B. & Rennaker, R. (2014). Opti-Speech: A real-time, 3D visual feedback system for speech training. Interspeech Conference.

Katz, W. & Mehta, S. (2015). Visual feedback of tongue movement for novel speech sound learning. Frontiers in Human Neuroscience, 9(612).

Keenan, J.P., Freund, S., Hamilton, R.H., Ganis, G. & Pascual-Leone, A. (2000). Hand response differences in a self-face identification task. Neuropsychologia, 38: 1047-1053.

Keyes, H., Brady, N., Reilly, R. B. & Foxe, J. J. (2010). My face or yours? Event-related potential correlates of self-face processing. Brain and Cognition, 72: 244-254.

Keysers, C., Kohler, E., Umilta, M.A., Fogassi, L., Gallese, V. & Rizzolatti, G. (2003). Audiovisual mirror neurons and action recognition. Experimental Brain Research, 153: 628-636.

Klucharev, V., Möttönen, R. & Sams, M. (2003). Electrophysiological indicators of phonetic and non-phonetic multisensory interactions during audiovisual speech perception. Brain Res. Cogn. Brain Res., 18: 65-75.

Knoblich, G. & Prinz, W. (2001). Recognition of self-generated actions from kinematic displays of drawing. J Exp Psychol Hum Percept Perform., 27: 456-465.

Kohler, E., Keysers, C., Umilta, M.A., Fogassi, L., Gallese, V. & Rizzolatti, G. (2002). Hearing sounds, understanding actions: action representation in mirror neurons. Science, 297: 846-848.

Kuhl, P.K. & Miller, J.D. (1975). Speech perception by the chinchilla: Voiced-voiceless distinction in alveolar plosive consonants. Science, 190: 69-72.

Kuhl, P.K. & Miller, J.D. (1978). Speech perception by the chinchilla: identification functions for synthetic VOT stimuli. Journal of the Acoustical Society of America, 63: 905–917.

Lahav, A., Saltzman, E. & Schlaug, G. (2007). Action representation of sound: Audiomotor recognition network while listening to newly acquired actions. Journal of Neuroscience, 27: 3008–3014.

Laurienti, P.J., Burdette, J.H., Maldjian, J.A. & Wallace, M.T. (2006). Enhanced multisensory integration in older adults. Neurobiology of aging, 27: 1155-1163.

Liberman, A.M., Cooper, F.S., Shankweiler, D.P. & Studdert-Kennedy, M. (1967). Perception of the speech code. Psychological Review, 74: 431-461.

Liberman, A.M. & Mattingly, I.G. (1985). The motor theory of speech perception revised. Cognition, 21: 1-36.

Lombardo, M.V., Chakrabarti, B., Bullmore, E.T., Sadek, S.A., Pasco, G., Wheelwright, S.J. et al. (2009). Atypical neural self-representation in autism. Brain, 133(Pt 2): 611-624.

Loula, F., Prasad, S., Harber, K. & Shiffrar, M. (2005). Recognizing people from their movements. Journal of Experimental Psychology: Human Perception & Performance, 31: 210-220.

Madden, D.J., Turkington, T.G., Provenzale, J.M., Denny, L.L., Hawk, T.C., Gottlob, L.R. & Coleman, R.E. (1999). Adult age differences in functional neuroanatomy of verbal recognition memory. Hum. Brain Mapp., 7: 115–135.

Marian, V. (2009). Audio-visual integration during bilingual language processing, in The Bilingual Mental Lexicon: Interdisciplinary Approaches, ed Pavlenko A., editor. (Bristol, UK: Multilingual Matters), 52–78.

McGrath, M. (1985). An examination of cues for visual and audiovisual speech perception using natural and computer generated faces. Thèse de doctorat, University of Nottingham, Angleterre.

McGurk, H. & MacDonald, J. (1976). Hearing lips and seeing voices. Nature, 264: 746-748.

Meister, I.G., Wilson, S.M., Deblieck, C., Wu, A.D. & Iacoboni, M. (2007). The Essential Role of Premotor Cortex in Speech Perception. Current Biology, 17(19): 1692-1696.

Mesch, J. (2000). Tactile Sign Language, Turn Taking in Signed Conversations of Deaf-Blind People, International studies on sign language and communication of the deaf, volume 38

Morin, M. (2011). Perception tactile de la parole. Expérimentation en Tadoma auprès de deux sujets entendants-voyants. Mémoire de master 1, Grenoble.

Möttönen, R., Järveläinen, J., Sams, M. & Hari, R. (2004). Viewing speech modulates activity in the left SI mouth cortex. NeuroImage, 24: 731-737.

Möttönen, R. & Watkins, K.E. (2009). Motor representations of articulators contribute to categorical perception of speech sounds. The Journal of Neuroscience, 29(31): 9819-9825.

Munhall, K. G., & Vatikiotis-bateson, E. (1998). The moving face during speech communication. In R. Campbell, B. Dodd & D. Burnham (Eds.), Hearing by Eye II: Advances in

the psychology of speechreading and audiovisual speech (pp. 123-139). Hove: Psychology Press.

Nahorma, O., Berthommier, F. & Schwartz, J.-L. (2010). Binding and unbinding in audiovisual speech fusion: Removing the McGurk effect by an incoherent preceding audiovisual context. International Conference on Auditory-Visual Speech Processing, AVSP2010. Hakone, Kanagawa, Japon.

Navarra, J. & Soto-Faraco, S. (2007). Hearing lips in a second language: visual articulatory information enables the perception of second language sounds. Psychological Research, 71: 4-12.

Nixon, P., Lazarova, J., Hodinott-Hill, I., Gough, P. & Passingham, R. (2004). The inferior frontal gyrus and phonological processing: an investigation using rTMS. Journal of Cognitive Neuroscience, 16(2): 289-300.

Northoff, G., Heinzel, A., de Greck, M., Bermpohl, F., Dobrowolny, H. & Panksepp, J. (2006). Self-referential processing in our brain — A meta-analysis of imaging studies on the self. NeuroImage, 31: 440–457.

Ojanen, V., Möttönen, R., Pekkola, J., Jääskeläinen, I.P., Joensuu, R., Autti, T. & Sams, M. (2005). Processing of audiovisual speech in Broca's area. NeuroImage, 25: 333-338.

Paris, T., Kim, J., & Davis, C. (2016). Using EEG and stimulus context to probe the modelling of auditory-visual speech. Cortex, 75: 220–230.

Pekkola, J., Ojanen, V., Autti, T., Jääskeläinen, I.P., Möttönen, R., Tarkiainen, A. & Sams, M. (2005). Primary auditory cortex activation by visual speech: an fMRI study at 3 T. Neuroreport, 6(2): 125-8.

Perrier, P. & Schwartz, J.L. (2016). De la bouche à l'oreille : éléments d'anatomie et de physiologie fonctionnelles des systèmes auditif et articulatoire. In S. Pinto & M. Sato (Eds.) Traité de Neurolinguistique (pp. 13-30). Louvain-la-Neuve, Belgique: De Boeck-Supérieur.

Pilling, M. (2009). Auditory event-related potentials (ERPs) in audiovisual speech perception. Journal of Speech, Language, and HearingResearch, 52: 1073-1081.

Pisoni, D.B. (1977). Identification and discrimination of the relative onset of two component tones: Implications for voicing perception in stop consonants. Journal of the Acoustical Society of America, 61: 1352-1361.

Pizzamiglio, L., Aprile, T., Spitoni, G., Pitzalis, S., Bates, E., D'Amico, S. & Di Russi, F. (2005). Separate neural systems for processing action- or non-action related sounds. Neuroimage, 24: 852–861.

Potts, G.F. (2004). An ERP index of task relevance evaluation of visual stimuli. Brain Cogn., 56(1): 5-13.

Preminger, J.E., Lin, H.B., Payen, M. & Levitt, H. (1998). Selective visual masking in speechreading. Journal of Speech, Language and Hearing Research, 41: 564-575.

Pulvermüller, F., Huss, M., Kherif, F., Moscoso del Prado Martin, F., Hauk, O. & Shtyrov, Y. (2006). Motor cortex maps articulatory features of speech sounds. Proceedings of the National Academy of Sciences of the USA, 103(20): 7865-70.

Rapin, L. (2011). Hallucinations auditives verbales et trouble du langage intérieur dans la schizophrénie: traces physiologiques et bases cérébrales. Thèse de Doctorat, Grenoble.

Rauschecker, J. & Scott, S. (2009). Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. Nature Neuroscience, 12: 718-725.

Rauschecker, J.P. (2011). An expanded role for the dorsal auditory pathway in sensorimotor control and integration. Hearing Research, 271: 16-25.

Reed, C.M., Rabinowitz, W.M., Durlach, N.I., Braida, L.D., Conway-Fithian, S., & Schultz, M.C. (1985). A Research on the Tadoma Method of Speech Communication. Journal of the Acoustical Society of America, 77: 247-257.

Reed, C.M., Rabinowitz, W.M., Durlach, N.I., Delhorne, L.A., Braida, L.D., Pemberton, J.C., Mulcahey, B.D. & Washington, D.L. (1992). Analytic study of the Tadoma method: Improving performance through the use of supplementary tactual displays. JSHR, 35(2): 450-465.

Reisberg, D., McLean, J. & Goldfield, A. (1987). Easy to hear but hard to understand: a lipreading advantage with intact auditory stimuli. In: Campbell, R., Dodd, B. (Eds.), Hearing by Eye: The Psychology of Lipreading. Lawrence Erlbaum Associates, London (UK), 97–113.

Remez, R.E., Rubin, P.E., Pisoni, D.B. & Carrell, T.D. (1981). Speech perception without traditional speech cues. Science, 212: 947–949.

Repp, B.H. & Knoblich, G. (2004). Perceiving action identity: How pianists recognize their own performances. Psychological Science, 15: 604–609.

Reuter-Lorenz, P., Jonides, J., Smith, E. S., Hartley, A., Miller, A., Marshuetz, C. & Koeppe, R.A. (2000). Age differences in the frontal lateralization of verbal and spatial working memory revealed by PET. J. Cogn. Neurosci., 12: 174–187.

Rizzolatti, G. & Freund, H.J. (2001). Action observation activates premotor and parietal areas in a somatotopic manner: an fMRI study. European Journal of Neuroscience, 13: 400-404.

Rizzolatti, G., Craighero, L. (2004). The mirror-neuron system. Annu Rev Neurosci, 27: 169–192.

Rizzolatti, G. & Fabbri-Destro, M. (2008). The mirror system and its role in social cognition. Current Opinion in Neurobiology, 18(2): 179-84.

Romero, L., Walsh, V. & Papagno, C (2006). The neural correlates of phonological short-term memory: a repetitive transcranial magnetic stimulation study. Journal of Cognitive Neuroscience, 18(7): 1147-1155.

Rosa, C., Lassonde, M., Pinard, C., Keenan, J. P. & Belin, P. (2008). Investigations of hemispheric specialization of self-voice recognition. Brain Cogn. 68: 204–214.

Rosenblum, L.D. (2010). See what I'm saying: the extraordinary power of our five senses.

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): 1147-53.

Roy, A.C., Craighero, L., Fabbri-Destro, M. & Fadiga, L. (2008). Phonological and lexical motor facilitation during speech listening: A transcranial magnetic stimulation study. Journal of physiology, Paris, 102(1-3): 101-105.

Sato, M., Tremblay, P. & Gracco, V. (2009). A mediating role of the premotor cortex in phoneme segmentation. Brain and Language, 111(1): 1-7.

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.

Sato, M., Cavé, C., Ménard, L. & Brasseur, A. (2010b). Auditory-tactile speech perception in congenitally blind and sighted adults. Neuropsychologia. 48(12): 3683-3686.

Saygin, A. (2007). Superior temporal and premotor brain areas necessary for biological motion perception. Brain, 130(9): 2452-2461.

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. Front. Psychol., 5: 568.

Schürmann, M., Caetano, G., Jousmäki, V. & Hari, R. (2003). Hands help hearing: facilitatory audiotactile interaction at low sound-intensity levels. Journal of the Acoustical Society of America. 115(2): 830-832.

Schwartz, J-L., Robert-ribes, J. & Escudier, P. (1998). Ten years after Summerfield ... a taxonomy of models for audiovisual fusion in speech perception. In R. Campbell, et al. (eds.) Hearing by eye, II (pp. 85-108). Hove (UK): Psychology Press.

Schwartz, J.-L, Abry, C., Boë, L.-J. & Cathiard, M.-A. (2002). Phonology in a theory of perception-for-action-control. In Durand, J., Lacks, B (Eds.), Phonology: From Phonetics to Cognition. Oxford University Press, Oxford, 240-280.

Schwartz, J.L., Berthommier, F. & Savariaux C. (2004). Seeing to hear better: evidence for early audio-visual interactions in speech identification. Cognition, 93: 69–78.

Schwartz, J.-L., Sato, M. & Fadiga, L. (2008). The common language of speech perception and action: a neurocognitive perspective. Revue Française de Linguistique Appliquée, 13(2): 9-22.

Schwartz, J.L., Sato, M. & Fadiga, L. (2011). Le langage commun de la perception et de l'action dans la communication parlée : une perspective neurocognitive. Faits de Langue, 37: 117-136.

Schwartz, J.-L, Ménard, L., Basirat, A. & Sato, M. (2012). The Perception for Action Control Theory (PACT): a perceptuo-motor theory of speech perception. Journal of Neurolinguistics, 25(5): 336-354.

Schwartz, J-L & Savariaux, C. (2014) No, There Is No 150 ms Lead of Visual Speech on Auditory Speech, but a Range of Audiovisual Asynchronies Varying from Small Audio Lead to Large Audio Lag. PLoS Comput Biol., 10(7): e1003743.

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

Sekiyama, K. & Tohkura, Y. (1991). McGurk effect in non-English listeners: few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility. J Acoust Soc Am., 90(4): 1797-805.

Sekiyama, K., Kanno, I., Miura, S. & Sugita Y. (2003). Auditory-visual speech perception examined by fMRI and PET. Neurosci. Res., 47: 277–287.

Sekiyama, K., Soshi, T. & Sakamoto, S. (2014). Enhanced audiovisual integration with aging in speech perception: a heightened McGurk effect in older adults. Frontiers in psychology, 5(323).

Senkowski, D., Schneider, T. R., Foxe, J. J., and Engel, A. K. (2008). "Crossmodal binding through neural coherence: implications for multisensory processing," Trends Neurosci, 31: 401-409.

Skipper, J.I., Nusbaum, H.C. & Small, S.L. (2005). Listening to talking faces: Motor cortical activation during speech perception. NeuroImage, 25: 76-89.

Skipper, J., Van Wassenhove, V, Nussman, H. & Small, S. (2007). Hearing lips and seeing voices: How cortical areas supporting speech production meditate audiovisual speech perception. Cerebral Corte, 17: 2387-2399.

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 Hear., 26: 263–275.

Stebbins, G.T., Carrillo, M.C., Dorman, J., Dirksen, C., Desond, J., Turner, D.A., Bennett, D. A., Wilson, R. S., Glover, G. & Gabrieli, D.E. (2002). Aging effects on memory encoding in the frontal lobes. Psychol. Aging, 17: 44–55.

Stefanics, G., Kremláèek, J. & Czigler, I. (2014). Visual mismatch negativity: a predictive coding view. Front. Hum. Neurosci., 8: 666.

Stekelenburg, J.J. & Vroomen, J. (2007). Neural correlates of multisensory integration of ecologically valid audiovisual events. Journal of Cognitive Neuroscience, 19: 1964–1973.

Stekelenburg, J. J., & Vroomen, J. (2012). Electrophysiological correlates of predictive coding of auditory location in the perception of natural audiovisual events. Frontiers in Integrative Neuroscience, 6: 26.

Stevens, K.N. & Klatt, D.H. (1974). Role of formant transitions in the voiced-voiceless distinction for stops. Journal of the Acoustical Society of America, 55: 653–59.

Stevens, K.N. & Blumstein, S.E. (1978). Invariant cues for place of articulation in stop consonants, Journal of the Acoustical Society ofAmerica, 64: 1358-68.

Stevenson, R.A., Nelms, C., Baum, S.H., Zurkovsky, L., Barense, M.D., Newhouse, P.A. & Wallace, M.T. (2015). Deficits in audiovisual speech perception in normal aging emerge at the level of whole-word recognition. Neurobiol Aging, 36(1): 283-291.

Strouse, A., Ashmead, D., Ohde, R. & Grantham, D. (1998). Temporal processing in the aging auditory system. J Acoust Soc Am., 104: 2385–2399.

Sumby, W.H. & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. The Journal of the Acoustical Society of America, 26: 212-215.

Summerfield, Q. A. (1987). Some preliminaries to a comprehensive account of audio-visual speech perception Hearing by Eye: The Psychology of LipReading (pp. 3-51). Londres: Erlbaum Associates.

Summerfield, Q. A. (1991). Visual perception of phonetic gestures. In I. G. Mattingly & M. Studdert-Kennedy (Eds.), Modularity and the Motor Theory of Speech Perception (Vol. 335, pp. 117-137). Hillsdale: Erlbaum Associates.

Sundara, M., Namasivayam, A.K. & Chen, R. (2001). Observation-execution matching system for speech: A magnetic stimulation study. Neuroreport, 12(7): 1341-1344.

Tai, Y.F., Scherfler, C., Brooks, D.J., Sawamoto, N. & Castiello, U. (2004). The human premotor cortex is 'mirror' only for biological actions. Current Biology, 14: 117-120.

Thomas, S. M., & Jordan, T. R. (2004). Contributions of Oral and Extraoral Facial Movement to Visual and Audiovisual Speech Perception, 30: 873- 888.

Titova, N. & Naatanen, R. (2001). Preattentive voice discrimination by the human brain as indexed by the mismatch negativity. Neurosci Let., 308: 63–65.

Tremblay, P. & Small, S.L. (2011). On the context-dependent nature of the contribution of the ventral premotor cortex to speech perception. NeuroImage, 57(4): 1561-71.

Troille, E., Cathiard, M-A. & Abry, C. (2010) Speech face perception is locked to anticipation in speech production. Speech Comm., 52: 513–524.

Tourville, J.A. & Guenther, F.H. (2011). The DIVA model: A neural theory of speech acquisition and production. Lang Cogn Process, 26(7): 952–981.

Tye-Murray, N., Sommers, M.S. & Spehar, B.P. (2007). Audiovisual integration and lipreading abilities of older adults with normal and impaired hearing. Ear and Hearing, 28(5): 656–668.

Tye-murray, N., Sommers, M., Spehar, B., Myerson, J. & Hale, S. (2010). Aging, audiovisual integration and the principle of inverse effectiveness. Ear and Hearing, 31(5): 636-644.

Tye-Murray, N., Spehar B., Myerson, J., Hale, S. & Sommers, M.S. (2012). Reading your own lips: Common coding theory and visual speech perception. Psychonomic Bulletin & Review, 20: 115-119.

Uddin, L.Q., Kaplan, J.T., Molnar-Szakacs, I., Zaidel, E. & Iacoboni, M. (2005). Self-face recognition activates a frontoparietal ‘mirror’ network in the right hemisphere: an event-related fMRI study. Neuroimage, 25: 926–35.

van Wassenhove, V., Grant, K.W. & Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. Proceedings of the National Academy of Sciences U.S.A., 102: 1181-1186.

van Wassenhove, V., Grant, K.W. & Poeppel, D. (2007). Temporal window of integration in auditory–visual speech perception. Neuropsychologia, 45: 598–607.

Vatikiotis-Bateson, E., Eigsti, I. M., Yano, S. & Munhall, K. G. (1998). Eye movement of perceivers during audiovisual speech perception. Perception & Psychophysics, 60: 926-940.

Vilain, C. (2002). Contribution à la synthèse de parole par modèle physique. Application à l'étude des voix pathologiques. Thèse de doctorat, Grenoble et Eindhoven.

Vroomen, J. & Stekelenburg, J.J. (2010). Visual anticipatory information modulates multisensory interactions of artificial audiovisual stimuli. Journal of Cognitive Neuroscience, 22: 1583-1596.

Wilson, S.M. & Iacoboni, M. (2006). Neural responses to non-native phonemes varying in producibility: evidence for the sensorimotor nature of speech perception. NeuroImage, 33(1): 316-25.

Wilson, S.M., Saygin, A.P., Sereno, M.I. & Iacoboni, M. (2004). Listening to speech activates motor areas involved in speech production. Nature Neuroscience, 7: 701-702.

Watkins, K. E., Strafella, A. P. & Paus, T. (2003). Seeing and hearing speech excites the motor system involved in speech production. Neuropsychologia, 41: 989–994.

Watkins, K.E. & Paus, T. (2004). Modulation of motor excitability during speech perception: the role of Broca's area. Journal of Cognitive Neuroscience, 16(6): 978-987.

Wik, P. & Engwall, O. (2008). Can visualization of internal articulators support speech perception?, Proceedings of Interspeech (Brisbane), 2627–2630.

Winneke, A.H. & Phillips, N.A. (2011). Does audio visual speech offer a fountain of youth for old ears? An event-related brain potential study of age differences in audio visual speech perception. Psychol. Aging, 26: 427–438.

Wolpert, D.M. (1997). Computational approaches to motor control. Trends in Cognitive Sciences, 1(6): 209-216.

Wong, P.C.M., Jin, J.X., Gunasekera, G.M., Abel, R., Lee, E.R. & Dhar, S. (2009). Aging and cortical mechanisms of speech perception in noise. Neuropsychologia, 47: 693–703.

Zekveld, A.A., Heslenfeld, D.J., Festen, J.M. & Schoonhoven, R. (2006). Top–down and bottom–up processes in speech comprehension. Neuroimage, 32: 1826-1836.