Le son de tes lèvres : corrélats électrophysiologiques de la perception audio-haptique de la parole.

Camille Cordeboeuf[1], Avril Treille[1], Coriandre Vilain[1], Marc Sato[1]
(1) Département Parole & Cognition, GIPSA-Lab, CNRS & Grenoble Université, France.
Correspondance : marc.sato@gipsa-lab.grenoble-inp.fr

RESUME

Face à la nature multimodale de la perception de la parole, une question fondamentale est celle d'une possible intégration précoce des informations issues des différentes modalités sensorielles et l'existence de mécanismes anticipatoires prédictifs. L'objectif de cette étude pilote était de tester par électroencéphalographie (EEG) une possible modulation du potentiel évoqué auditif précoce N1 lors de la perception audio-haptique de la parole par rapport à une perception auditive seule. Dans ce but, nous avons comparé les réponses électroencéphalographiques de cinq participants obtenues lors de la perception de syllabes selon différentes modalités : auditive, audio-visuelle et audio-haptique. En accord avec de précédentes études, la comparaison des modalités auditive et audio-visuelle montrent une baisse d'amplitude de l'onde N1 lors de la perception audio-visuelle, un résultat suggérant une intégration précoce de ces deux modalités. Des résultats similaires sont observés lors de la comparaison des modalités auditive et audio-haptique pour les électrodes pariétales. De plus, une plus faible latence de l'onde N1 est observée pour la modalité audio-haptique. Pris ensemble, ces résultats suggèrent une intégration précoce des modalités auditive, visuelle et haptique lors de la perception de la parole et soulignent le possible rôle prédictif des informations haptiques dans le décodage et traitement des informations auditives.

ABSTRACT

The sound of your lips: electrophysiological correlates of audio-haptic speech perception

Given the multisensory nature of speech perception, one fundamental question is whether sensory signals are integrated early in the speech processing hierarchy and may reflect predictive, anticipatory, mechanisms. The present pilot EEG study aimed at investigating a possible modulation of auditory-evoked N1 component during audio-haptic compared to purely auditory speech perception. To this aim, we compared auditory-evoked N1 responses from five participants during auditory, audio-visual and audio-haptic perception of syllables. In line with previous studies, auditory-evoked N1 amplitude was attenuated during audio-visual compared to auditory speech perception. Crucially, similar results were observed for audio-haptic compared to auditory speech perception for parietal electrodes, with shortened latency. Altogether, these results suggest some early integrative mechanisms between auditory, visual and haptic modalities in speech perception as well as a predictive role of haptic information in auditory speech processing.

MOTS-CLES : perception de la parole, multimodalité, interactions audio-haptique, EEG.
KEYWORDS : speech perception, multimodality, audio-haptic interactions, EEG.

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 1: JEP, pages 803–810,
Grenoble, 4 au 8 juin 2012.

1 Introduction

Bien que l'audition soit considérée comme la modalité sensorielle principale de la communication parlée, la perception de la parole est par essence fondamentalement multisensorielle. Ainsi les informations visuelles issues du visage de notre interlocuteur modifient profondément le traitement de la parole, notamment en améliorant l'intelligibilité d'un signal de parole présenté dans le bruit (Sumby et Pollack, 1954 ; Benoît, Mohamadi and Kandel, 1994). L'effet McGurk (McGurk and MacDonald, 1976) est une autre démonstration de l'importance et de l'influence des informations visuelles sur le décodage de la parole. En plus des modalités auditives et visuelles, on sait également par la méthode Tadoma (Alcorn, 1932), utilisée par des personnes sourdes et aveugles, que des informations tactiles (perception haptique), obtenues en plaçant une main sur le visage du locuteur, permettent d'accéder à un niveau de communication quasi-normal ce, par la récupération d'informations sur le voisement, le mouvement des lèvres et l'ouverture mandibulaire des gestes de parole produits. Différentes études ont montré que des interactions audio-haptiques pour la parole existent également chez des sujets normaux non entraînés. Ainsi Fowler et Dekle (1991) ont mis en évidence lors d'une tâche de perception catégorielle l'existence d'interactions entre modalités auditive et haptique : l'information tactile influence le décodage de la syllabe auditive et, réciproquement, la syllabe auditive influence le décodage de la syllabe perçue tactilement. De plus, la présentation audio-haptique de syllabes non cohérentes peut produire chez certains sujets un percept illusoire de type McGurk (Fowler et Dekle, 1991) ou, à tout le moins, entrainer une diminution de performance par rapport à une perception auditive seule (Sato et al., 2010). Il a enfin été montré que l'information tactile, ajoutée à une information visuelle ou auditive dans un milieu bruité, améliore la perception de la parole chez des sujets non entrainés (Gick et al., 2008 ; Sato et al., 2010).

Pris ensemble, ces résultats soulèvent d'importantes questions sur les interactions entre la modalité auditive et les autres modalités sensorielles et sur un possible couplage fonctionnel entre systèmes de perception et de production de la parole (Schwartz et al., 2010 ; Grabski et al., 2010). Notamment, une question fondamentale est celle d'une possible intégration précoce des informations issues des différentes modalités sensorielles via l'existence de mécanismes anticipatoires prédictifs. Certaines études suggèrent en effet que dans le cas d'un signal visuel de parole précédant l'information auditive, cette avance temporelle serait exploitée par notre système perceptif afin d'extraire des indices permettant d'anticiper leur conséquence acoustique (Cathiard, 1994). Ces mécanismes anticipatoires prédictifs sont également à la base de modèles récents neurobiologiques de la perception de la parole (Skipper et al., 2007 ; Rauschecker and Scott, 2009). Ainsi, d'après Skipper et collègues (2007), les informations auditives et visuelles convergeraient au niveau des aires temporales associatives postérieures supérieures. De là, un mécanisme de simulation motrice permettrait alors d'associer les mouvements articulatoires associés aux phonèmes perçus et, en retour, de prédire les états auditifs et somatosensoriels associés à ces mouvements simulés afin de contraindre l'interprétation phonétique finale de l'auditeur.

En accord avec ces hypothèses, des études EEG ont démontré une diminution de l'amplitude du potentiel évoqué auditif précoce N1 lors de la perception audio-visuelle de syllabes par rapport à une perception auditive seule (Klucharev, Möttönen and Sams, 2003 ; Besle et al., 2004 ; Van Wassenhove, Grant and Poeppel, 2005; Stekelenburg and Vroomen, 2007 ; Pilling, 2009 ; Vroomen and Stekelenburg, 2009). L'onde N1 auditive apparaissant environ 100ms

après l'onset d'un stimulus acoustique et étant traditionnellement reliée à une analyse précoce des indices acoustiques de ce stimulus dans le cortex auditif, la diminution d'amplitude observée dans ces études pourrait refléter une facilitation de traitement des syllabes auditives due à la présence d'informations phonétiques visuelles, à une latence où les différents traits acoustiques n'ont pas encore abouti à une représentation intégrée.

Dans cette étude, nous avons utilisé la méthode Tadoma pour évaluer l'interaction entre information tactile et information auditive lors de la perception de la parole en comparant les amplitudes et latences du potentiel évoqué auditif précoce N1 lors d'une tâche d'identification syllabique selon les modalités auditive, audio-visuelle et audio-haptique (grâce à une méthode similaire à la méthode TADOMA pour cette dernière modalité). En accord avec de précédentes études, la comparaison des modalités auditive et audio-visuelle devraient montrer une baisse d'amplitude de l'onde N1 lors de la perception audio-visuelle. De plus, des résultats similaires lors de la comparaison des modalités auditive et audio-haptique suggéreraient l'existence d'un mécanisme d'intégration précoce des modalités auditive et haptique.

2 Méthodes

2.1 Participants

Six sujets adultes, âgés de 26 à 42 ans, ont participé à l'expérience. Tous les participants étaient droitiers, locuteurs natifs du français et ne présentaient pas de troubles de compréhension ou de production de la parole. Tous les sujets ont donné préalablement à l'étude leur consentement éclairé.

2.2 Procédure

L'expérience consistait en la perception des syllabes /pa/ et /ta/ produites individuellement par une expérimentatrice de langue maternelle française. Cinq modalités perceptives ont été testées : Auditive (A : le sujet garde les yeux fermés et seule la voix de l'expérimentatrice est perçue), Visuelle (V : le sujet a les yeux ouverts et regarde l'expérimentatrice prononcer les syllabes silencieusement), Audio-Visuelle (AV : le sujet a les yeux ouverts et regarde l'expérimentatrice prononcer les syllabes à haute voix), Haptique (H : le sujet garde les yeux fermés, la main droite disposée sur les lèvres et la mandibule de l'expérimentatrice qui prononce les syllabes silencieusement) et Audio-Haptique (AH : le sujet a les yeux fermés et la main droite disposée sur les lèvres et la mandibule de l'expérimentatrice qui prononce les syllabes à haute voix).

L'expérience s'est déroulée dans une chambre sourde et consistait en 5 sessions expérimentales indépendantes correspondant aux modalités perceptives A, V, AV, H et AH. Chaque session est basée sur l'identification par le participant des syllabes /pa/ ou /ta/ prononcées individuellement par l'expérimentatrice (procédure à choix forcé). L'ordre de passage de ces différentes conditions a été randomisé entre les sujets. Pour chaque session, 80 syllabes ont été présentées de manière aléatoire (40 /pa/ et 40 /ta/). La durée de chaque essai était de 3 secondes. 600ms après la prononciation de la syllabe par l'expérimentatrice, une alerte sonore indiquait aux participants le moment de délivrer leur réponse. Pour ce faire, le sujet disposait de deux touches clavier et répondait avec sa main gauche.

Avant l'expérience, les participants étaient informés qu'il leur serait présenté les syllabes /pa/ ou /ta/ soit auditivement, soit visuellement, soit tactilement par contact entre leur main et le visage de l'expérimentatrice, soit par deux modalités en même temps. Un court entrainement était donné préalablement à l'expérience pour chacune des modalités. La procédure expérimentale dans les conditions H et AH est inspirée de celle de Fowler et Dekle (1991) et Sato et collègues (2010). Les participants étaient assis face à l'expérimentatrice, leur main droite placée sur son visage, le pouce posé verticalement sur les lèvres et les autres doigts placés horizontalement sur la mandibule. Cette position permettait de capter les mouvements des lèvres et de la mandibule lors de la production des syllabes /pa/ et /ta/. Le participant avait le coude posé sur la table et surélevé par un support en mousse. Pour éviter que les sujets ne regardent l'expérimentatrice, ils fermaient les yeux lors de ces deux conditions. L'expérimentatrice était assise face au sujet et à un écran d'ordinateur. À chaque essai, l'écran lui indiquait la syllabe à prononcer et affichait des indices temporels liés à la production de la syllabe. De manière à restreindre la variabilité des ses productions, elle était entraînée avant l'expérimentation à articuler à haute voix et silencieusement chaque syllabe en synchronie avec les indices affichés sur son écran. L'ensemble des productions de l'expérimentatrice lors des sessions A, AV et AH a été enregistré de manière à permettre une synchronisation des signaux EEG avec l'onset des syllabes produites.

2.3 EEG

Lors de chacune des sessions, un enregistrement continu des signaux EEG provenant de 9 électrodes représentatives (F3, Fz, F4, C3, Cz, C4, P3, Pz, P4 selon le système international 10-20) a été effectué via le système BIOSEMI. Ces 9 électrodes frontales, centrales et pariétales ont été sélectionnées du fait d'une réponse maximale du PE auditif N1 précédemment observée pour les électrodes centrales et permettaient de couvrir une partie conséquente du scalp des participants. Une électrode externe de référence a été placée sur l'extrémité du nez et les mouvements oculaires verticaux (VEOG) et horizontaux (HEOG) ont été enregistrés via deux électrodes placées sur le coté externe de chaque œil et une autre électrode placée sous l'œil gauche. Avant chaque expérience, l'impédance de toutes les électrodes était inférieure à 20 KΩ. Lors des enregistrements, la fréquence d'échantillonnage était fixée à 256 hz. En vue de permettre l'analyse des données EEG, un étiquetage semi-automatique des onsets syllabiques produits par l'expérimentatrice lors des sessions A, AV et AH a été réalisé via le logiciel Praat. Les triggers des enregistrements EEG ont ensuite été resynchronisés de manière à correspondre aux onsets syllabiques pour chaque essai et chaque session. Du fait de l'absence de marqueurs temporels précis pour les conditions de production silencieuse, les sessions H et AV n'ont pas été analysées. Pour les sessions A, AV et AH, les données EEG ont été prétraitées et analysées via le logiciel EEGlab sous environnement Matlab. Suite à l'indexation des signaux des 9 électrodes (F3, Fz, F4, C3, Cz, C4, P3, Pz, P4) par rapport à l'électrode de référence, un filtre passe-bande 1-40Hz a été appliqué. Pour l'ensemble des essais, les données ont ensuite été segmentées en évènements de 1000ms centrés sur l'onset de la syllabe, incluant une baseline de 100ms (de -500 à -400ms). Les évènements impliquant un changement d'amplitude supérieur à $\pm$60 µV pour tout électrode (y compris les électrodes HEOG et VEOG) ont été éliminés (en moyenne 6% des essais $\pm$5%).

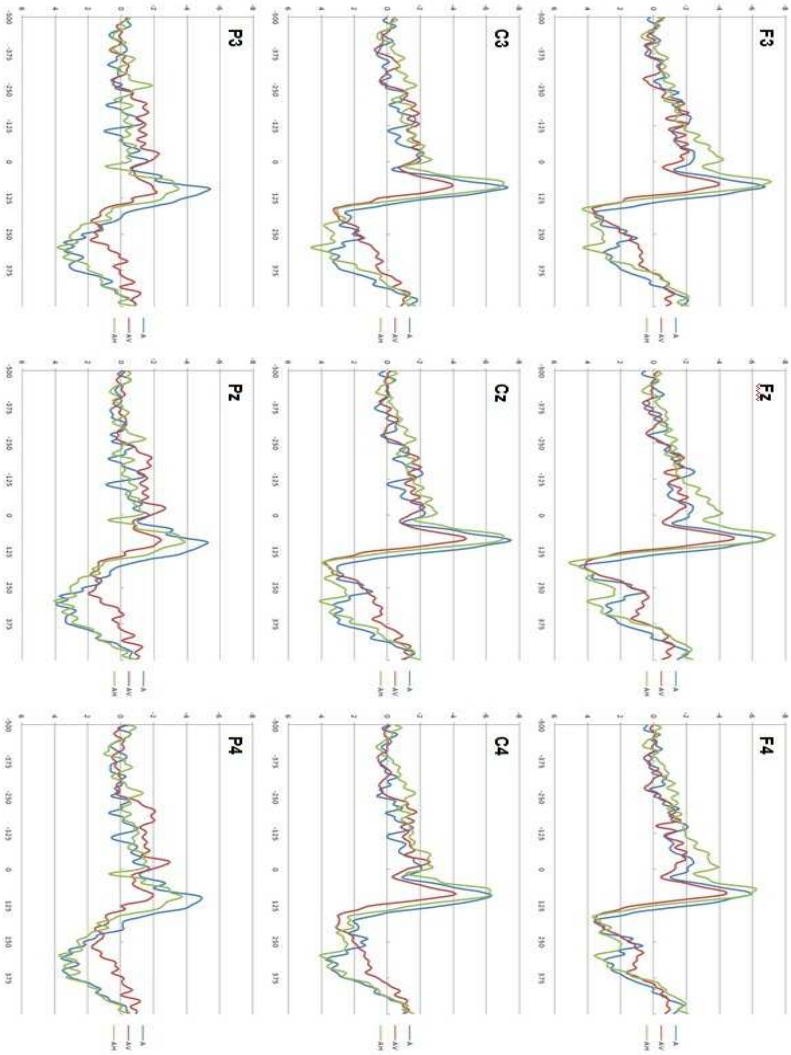

Figure 1 : Réponses EEG pour les conditions A (bleu), AV (rouge) et AH (vert). Chaque courbe représente la réponse moyenne d'une électrode : F3, Fz, F4 (frontales), C3, Cz, C4 (centrales) et P3, Pz, P4 (pariétales) (nombre impaire : position gauche, z : position centrale, paire : position droite).

3 Résultats

Pour toutes les analyses, le niveau de significativité a été fixé à $p < 0.05$, un test de Mauchley a été effectué de manière à vérifier l'hypothèse de sphéricité des données, enfin des tests de Newman-Keuls ont été utilisés pour les analyses post-hoc.

3.1 Réponses comportementales

L'ensemble des réponses recueillies a été analysé pour chaque participant et chaque condition. Les données ont été traitées par une analyse de variance (ANOVA) à mesures répétées avec pour variable intra-sujets la syllabe présentée (/pa/, /ta/), et la modalité de présentation (A, V, AV, H, AH). Les scores observés sont très élevés pour toutes les conditions (en moyenne 99%). Néanmoins, un effet significatif de la modalité de présentation est observé ($F_{(4,20)} = 3.85$, $p < 0.02$) avec un score perceptif plus faible en modalité haptique p/r à toutes les autres modalités (en moyenne, 100%, 99%, 100%, 97%, 99% pour les modalités A, V, AV, H et AH). Il n'y a pas d'effet de la syllabe ni d'interaction 'modalité x syllabe'.

3.2 Réponses EEG

Pour chaque participant et condition (A, AV, AH), les signaux EEG des électrodes frontales (C3, Cz, C4), centrales (F3, Fz, F4) et postérieures (P3, Pz, P4) ont été moyennés par électrode pour les 80 essais. L'amplitude et la latence du potentiel évoqué N1 ont ensuite été calculés. Pour l'amplitude et la latence, les données ont été traitées par une ANOVA à mesures répétées avec pour variable intra-sujets la modalité de présentation (A, AV, AH), la position de l'électrode sur l'axe latéral (gauche, centre, droite) et sur l'axe caudo-rostral (antérieur, centre, postérieur). Du fait d'un signal EEG bruité, un sujet n'a pu être analysé. Les réponses EEG moyennées par condition pour les 5 sujets et pour les électrodes frontales (F3, Fz, F4), centrales (C3, Cz, C4) et postérieures (P3, Pz, P4) sont indiqués sur la Figure 1.

Amplitude N1: L'ANOVA réalisée sur l'amplitude des PE auditifs N1 (voir la Figure 2) montre un effet significatif de la modalité ($F_{(2,8)} = 8.89$, $p < 0.01$) avec une amplitude plus faible pour la modalité AV p/r aux deux autres modalités A et AH. Un effet significatif de la position caudo-rostrale des électrodes est observé ($F_{(2,8)} = 12.52$, $p < 0.004$) avec une amplitude inférieure pour les électrodes postérieures par rapport aux électrodes antérieures et centrales. Enfin, une interaction 'modalité x position caudo-rostrale' est observée ($F_{(4,16)} = 4.11$, $p < 0.02$). Cette interaction provient du fait de différences d'amplitudes significatives entre les conditions A et AH pour les électrodes postérieures mais non pour les électrodes antérieures et centrales.

Latence N1: L'ANOVA réalisée sur la latence des potentiels évoqués N1 (voir Figure 3) montre un effet significatif de la modalité ($F_{(2,8)} = 6.89$, $p < 0.02$) avec une latence plus faible pour la modalité AH p/r aux deux autres modalités A et AV. Les interactions 'modalité x position latérale' ($F_{(4,16)} = 3.24$, $p < 0.04$) et 'modalité x position caudo-rostrale' ($F_{(4,16)} = 4.00$, $p < 0.02$) sont également significatives. L'interaction 'modalité x position latérale' démontre une latence plus faible pour les électrodes gauches en modalité AH par rapport aux autres électrodes. L'interaction 'modalité x position caudo-rostrale' démontre une

latence plus importante pour les électrodes postérieures par rapport aux électrodes centrales et antérieures pour la modalité A.

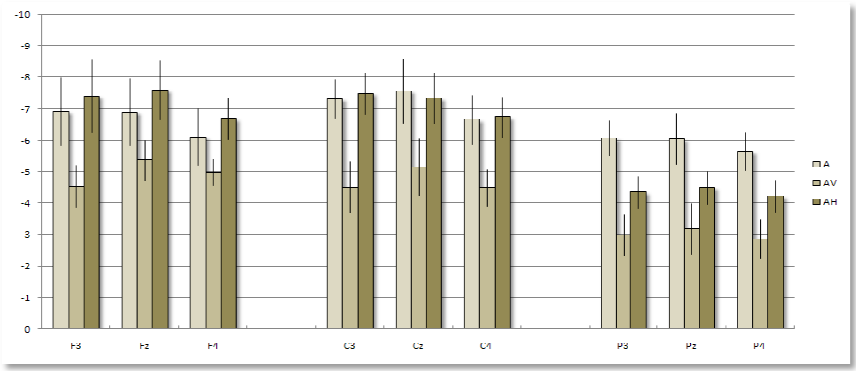

Figure 2 : Amplitude moyenne (en µV) du PE auditif N1 en fonction des conditions A, AV et AH et des électrodes F3, Fz, F4 (frontales), C3, Cz, C4 (centrales) et P3, Pz, P4 (pariétales).

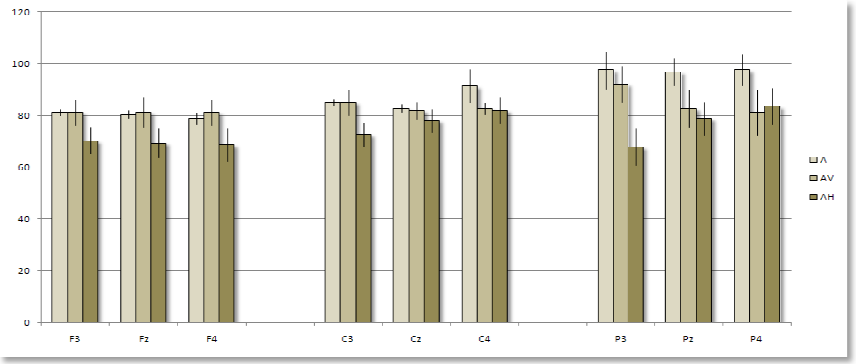

Figure 3 : Latence moyenne (en ms) du PE auditif N1 en fonction des conditions A, AV et AH et des électrodes F3, Fz, F4 (frontales), C3, Cz, C4 (centrales) et P3, Pz, P4 (pariétales).

4 Conclusion

En accord avec de précédentes études, la comparaison des modalités auditive et audio-visuelle démontre une moindre amplitude de l'onde N1 lors de la perception audio-visuelle. Bien qu'aucune modulation d'amplitude ne soit constatée entre les modalités audio-haptique et auditive au niveau des électrodes frontales et centrales, une moindre amplitude pour la modalité audio-haptique est cependant observée au niveau des électrodes pariétales. De plus, une moindre latence de l'onde N1 est observée lors de la condition audio-haptique par rapport aux conditions auditive et audio-visuelle. Bien que ces résultats doivent être confirmés par l'examen d'un plus grand nombre de sujets, ils suggèrent néanmoins une

intégration précoce des modalités auditive, visuelle et haptique lors de la perception de la parole et soulignent le possible rôle prédictif des informations haptiques dans le décodage et traitement des informations auditives.

Références

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

BENOÎT, C., MOHAMADI, T., & KANDEL, S. (1994). Effects of 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. *Eur. J. Neurosci.*, 20: 2225-2234.

CATHIARD, M. A. (1994). La perception visuelle de l'anticipation des gestes vocaliques: cohérence des évènements audibles et visibles dans le flux de la parole. *Thèse de doctorat. Université Stendhal, Grenoble, France*.

FOWLER, C. & DEKLE, D. (1991). Listening with eye and hand: crossmodal contributions to speech perception. *J. Exp. Psychol. Hum. Percept. Perform.*, 17: 816–828.

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., LAMALLE, L., VILAIN, C., SCHWARTZ, J.-L, VALLÉE, N. TROPRÈS, I., BACIU, M. LE BAS, J.-F & SATO, M. (2010). Corrélats neuroanatomiques des systèmes de perception et de production des voyelles du Français. *Proceedings of the XXVIIIèmes Journées d'Étude sur la Parole*.

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.

MCGURK, H. & MACDONALD, J. (1976). Hearing lips and seeing voices. *Nature*, 264: 746-748.

PILLING, M. (2009). Auditory event-related potentials (ERPs) in audiovisual speech perception. *Journal of Speech, Language, and Hearing Research*, 52: 1073–1081.

RAUSCHECKER, J.P., & SCOTT, S.K. (2009). Maps and streams in the auditory cortex: Nonhuman primates illuminate human speech processing. *Nature Neuroscience*, 12(6): 718–724.

SATO, M., CAVE, C., MENARD, L. & BRASSEUR, A. (2010). Auditory-tactile speech perception in congenitally blind and sighted adults. *Neuropsychologia*, 48(12): 3683-3686.

SCHWARTZ, J.-L., MÉNARD, L., BASIRAT, A. & SATO, M. (IN PRESS). The Perception for Action Control Theory (PACT): a perceptuo-motor theory of speech perception. *Journal of Neurolinguistics*.

SKIPPER, J.I., VAN WASSENHOVE, V., NUSBAUM, H.C. & SMALL, S.L. (2007). Hearing lips and seeing voices: how cortical areas supporting speech production mediate audiovisual speech perception. *Cerebral Cortex*, 17(10): 2387-2399.

STEKELENBURG, J.J. & VROOMEN, J. (2007). Neural correlates of multisensory integration of ecologically valid audiovisual events. *Journal of Cognitive Neuroscience*, 19(12): 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. *Proc. Natl. Acad. Sci. USA*, 102: 1181-1186.

VROOMEN, J. & STEKELENBURG, J.J. (2009). Visual anticipatory information modulates multisensory interactions of artificial audiovisual stimuli. *Journal of Cognitive Neuroscience*, 22(7): 1583-1596.