Traitement audiovisuel lors d'une tâche de discrimination syllabique : une étude EEG/IRMf simultanée

Cyril Dubois[1,2] *Rudolph Sock*[2]

(1) Université de Zürich Romanisches Seminar 8 Zürichbergstr. 8032 Zürich

(2) Université de Strasbourg Institut de Phonétique de Strasbourg (IPS) / Équipe Parole et Cognition (PC), U.R. 1339 – LiLPa, 22, rue René Descartes 67084 Strasbourg cedex

cyril.dubois@uzh.ch, sock@unistra.fr

RÉSUMÉ

Nous avons mené une étude anatomo-fonctionnelle simultanée en Imagerie par Résonance Magnétique fonctionnelle / Électro-encéphalographie (IRMf/EEG), en utilisant une tâche de discrimination à choix forcé, portant sur des syllabes CV, selon deux modalités perceptives : audiovisuelle dynamique et audiovisuelle statique, afin de pouvoir observer les bases neurophysiologiques de la perception audiovisuelle syllabique. La tâche de discrimination portait sur des paires syllabiques, s'opposant sur les trois traits suivants : la labialité vocalique, le lieu d'articulation et le voisement consonantiques. Les résultats IRMf montrent un recrutement de structures corticales dans le gyrus temporal supérieur et dans le cortex occipital des deux hémisphères (correspondant à la perception visuelle), ainsi que des activations du cortex prémoteur gauche. L'analyse des potentiels évoqués (EEG) révèle que l'influence des mouvements est précoce et se manifeste dès 150 millisecondes, mais aussi de façon plus tardive autour de 250 ms, après le début du stimulus d'intérêt.

ABSTRACT

Audiovisual processing in syllabic discrimination task: a simultaneous fMRI-EEG study

We conducted a study based on simultaneous fMRI/EEG recordings, in a discrimination task, comprising CV syllables, in two perception modalities: audiovisual dynamic and audiovisual static, in order to investigate the neural substrates of audiovisual syllabic perception. The discrimination task was based on syllable pairs, contrasting three features: vowel lip rounding, consonant place of articulation and voicing. fMRI results show significant activations in the superior temporal gyrus and in the occipital cortex bilaterally (associated with visual perception), and also a recruitment of the left Premotor cortex. Significant evoked potential responses to syllabic discrimination were recorded around 150 ms and 250 ms following the onset of the second stimulus of the pairs, whose amplitude was greater in the dynamic modality compared to the static audiovisual modality. Our results provide arguments for the involvement of the speech motor cortex in speech perception, and suggest a multimodal representation of speech units.

MOTS-CLÉS : Neurophysiologie, EEG/IRMf, perception audiovisuelle, syllabes.

KEYWORDS: Neurophysiology, EEG/fMRI, audiovisual perception, syllables.

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 1: JEP, pages 851–858,
Grenoble, 4 au 8 juin 2012.

1 Introduction

L'objectif général de cette étude est une contribution à la précision du recrutement de zones cérébrales (IRMf) et du *timing* (EEG) impliqués dans les processus de perception de la parole audiovisuelle et, par là même, dans la compréhension du langage articulé. L'intérêt d'une étude couplant simultanément ces deux techniques consiste à recueillir des données sur la localisation et le décours temporel au sein d'une seule et unique session expérimentale. L'avantage de ce recueil simultané repose sur le fait que les phénomènes observés, enregistrés dans des conditions expérimentales similaires, ce qui laisse penser que les processus attentionnels, sensoriels et motivationnels sont identiques (Debener, Ullsperger, Siegel, Fiehler, von Cramon & Engel, 2005). Ainsi, on évite le biais possible de la variation intra-individuelle lors d'enregistrements séparés. Sumby & Pollack (1954) ont démontré que la perception visuelle du visage du locuteur améliorait l'intelligibilité des mots en milieu bruité. Ross, Saint-Amour, Leavitt, Javitt & Foxe (2006) n'observent pas une progression linéaire, le gain étant maximal pour un rapport signal sur bruit de -12 dB. La perception de phonèmes et de syllabes, appartenant à une même classe de visèmes, semble aussi améliorée par la présence d'indices visuels phonologiques (Schwartz, Berthommier & Savariaux, 2004),tout comme la perception des accents lexicaux (Scarborough, Keating, Baroni, Cho, Mattys, Alwan, Auer & Bernstein, 2006). Ces résultats montrent que la perception visuelle a un impact favorable, non seulement, sur l'intelligibilité, mais aussi sur la perception prélexicale. Le cadre théorique sous-jacent que nous avons retenu afin d'ordonner nos résultats est celui de Hickok & Poeppel, (2007) qui reprend l'idée d'un traitement double, et qui postule l'existence d'une voie dite ventrale, qui serait orientée vers la lexicalité (c'est-à-dire principalement vers la compréhension de la parole), et d'une voie dorsale qui serait une interface sensori-motrice (par conséquent impliquée dans la production de la parole). Préalablement à la subdivision en deux voies, deux premières « phases » entrent en jeu : ce sont les traitements acoustique et phonético-phonologique ; les auteurs emploient les notions d'« analyse spectrotemporelle » et de « réseau phonologique ». Notre étude met en jeu deux modalités perceptives se différenciant par la présence ou l'absence de mouvements visuels linguistiquement pertinents. Nous évoquerons par conséquent les modalités audiovisuelles dynamique et statique. À l'aide du paradigme de la soustraction cognitive, nous souhaitons observer les zones cérébrales impliquées dans la perception visuelle de la parole. Dans la perspective d'affiner nos données, nous avons choisi de comparer des syllabes CV se différenciant en fonction d'un seul trait distinctif. Les tâches de discrimination présentent trois contrastes au sein des paires syllabiques. Une opposition portait sur la labialisation vocalique (étirée *vs*. arrondie [i y]). Les deux autres oppositions étaient consonantiques et portaient soit sur le voisement, ou plus précisément sur l'un des indices de l'opposition de sonorité en français, le Délai d'Établissement du Voisement ou "Voice Onset Time" (VOT : sourdes *vs*. sonores : [p b] et [t d]), soit sur les lieux d'articulation (extra *vs.* intra-buccales : [p t] et [b d]). Nous avons retenu ces trois traits en fonction de leur apport visuel à la perception de la parole. La question principale de cette étude est de savoir si l'intégration de la dimension visuelle dans les processus de discrimination phonologique est sous tendue par le recrutement de régions cérébrales dédiées, ou par une modulation de l'activité des réseaux impliqués dans le traitement auditif pur ? Il y a-t-il une implication du cortex auditif primaire lors de la lecture labiale (Calvert, Bullmore, Brammer, Campbell, Williams, McGuire, Woodruff, Iversen & David, 1997) ? On peut aussi s'interroger sur le *timing* des dits processus, en particulier sur une accélération éventuelle de ceux-ci dans le cadre de la perception audiovisuelle dynamique (van Wassenhove, Grant & Poeppel, 2005).

2 Méthode

2.1. Participants

Pour cette étude, nous avons recruté vingt-six participants (quatorze femmes et douze hommes ; âge moyen : 22.6 ± 3.7). Parmi ces vingt-six sujets, seuls onze d'entre eux ont pu être considérés lors de l'analyse des potentiels évoqués (8 femmes et 3 hommes ; âge moyen : 22.55 ± 3.1). Ce sont principalement des artefacts (mouvements oculaires et artefacts cardiaques) qui ont suscité l'exclusion des autres participants.

2.2. Protocole expérimental

Nous avons utilisé un paradigme de discrimination à choix forcé « AX ». Dans ce paradigme, deux stimuli sont présentés l'un après l'autre séparés par une pause. Au sein d'un essai composé de deux syllabes ou de deux stimuli non phonologiques, les participants devaient juger si le second stimulus était identique (AA) ou différent du premier (AB). Ce paradigme a été appliqué aux deux modalités audiovisuelle statique (AVs) et audiovisuelle dynamique (AVd). Chaque tâche de discrimination comprenait huit catégories, constituées de 40 essais, soit un total de 320 paires. Afin d'obtenir une ligne dite de base, 40 essais exempts de tous stimuli étaient présentés durant chaque session (AVs et AVd). L'utilisation de 40 items par catégories est rendue nécessaire par l'IRMf, afin de pouvoir compiler la réponse hémodynamique, ainsi que par l'EEG afin d'obtenir des grandes moyennes statistiquement valides. En raison de l'amplitude très faible du potentiel lié à un événement par rapport à l'activité spontanée du cerveau, il est nécessaire d'enregistrer de nombreuses réponses évoquées par le même événement. Les catégories sont scindées en deux groupes, l'un comprenant les paires appelant une réponse « identique » (AA), et l'autre comprenant les paires appelant une réponse « différente » (AB). Dans ces deux groupes, nous avons introduit des paires ne faisant pas partie du système phonologique du français contemporain. Ces paires non phonologiques sont utilisées afin de faire apparaître les zones cérébrales impliquées dans les processus de traitement de la parole, grâce à la méthode de la « soustraction cognitive ». Cette méthode consiste à mettre en place deux conditions expérimentales en tous points identiques à l'exception du processus d'intérêt, ici la dimension phonologique des stimuli.

2.2.1. Stimuli

Nous avons filmé une locutrice francophone, sans accent identifiable, âgée de 23 ans prononçant les syllabes isolément. Enregistrée à l'aide d'une caméra (Sony DXC D30-Pal), chaque séquence vidéo était constituée de neuf images, soit une durée de 360 millisecondes par syllabe. Pour la modalité AV statique, nous avons utilisé le signal acoustique acquis durant l'enregistrement des séquences filmées. Une image fixe était projetée afin de pouvoir observer l'influence des indices visuels dynamiques lors des analyses comparant les deux modalités perceptives. Nous avons effectué un cadrage de sorte que n'apparaisse que le bas du visage de notre locutrice, afin de limiter les distracteurs et de concentrer l'attention des sujets.

2.2.2. Corpus

Nos stimuli sont constitués d'une part de huit syllabes naturelles du type Consonne – Voyelle (CV) : [pi bi ti di pu bu tu du] et, d'autre part, de huit syllabes naturelles modifiées à l'aide du logiciel Audacity® (paires non phonologiques). Ces paires non phonologiques ont été créées à partir de syllabes naturelles prononcées par la même locutrice. Nous les avons modifiées afin

de les rendre méconnaissables, tout d'abord en inversant le décours temporel des images et des sons, puis en appliquant une distorsion aux signaux acoustiques Les syllabes nous ont permis de réaliser des paires minimales. Les paires syllabiques s'opposant par un seul trait articulatoire pertinent ont été construites afin d'étudier trois contrastes opératoires en français contemporain. Ces paires minimales diffèrent en fonction de la labialisation (étirée *vs.* arrondie [i y]) pour l'opposition vocalique, en fonction du voisement (sourdes *vs.* sonores : [p b] et [t d]), et des lieux d'articulation (extra *vs.* intra-buccales : [p t] et [b d]).

2.2.3. Paradigme IRMf/EEG

Nous avons utilisé un paradigme d'IRMf événementiel à intervalle fixe. Les réponses hémodynamiques liées aux stimuli d'intérêt, dans une condition donnée, ont été moyennées. Ensuite, la réponse moyenne, liée aux stimuli dits de « contrôle », a été soustraite à la réponse aux stimuli audiovisuels dynamiques et statiques. Durant l'acquisition d'un volume cérébral (4 sec.), 1800 ms étaient réservées pour la présentation des paires de stimuli. La période pré-stimuli durait 100 ms. Nos syllabes durant 360 ms et l'intervalle inter-stimuli étant de 400 ms, la durée totale de présentation est de 1120 ms (deux fois 360 ms plus 400 ms). La fenêtre d'analyse des potentiels évoqués débute 100 ms avant le début du second stimulus (760 ms). La période post-stimuli durait 580 ms. Par conséquent, la fenêtre d'analyse des potentiels évoqués se terminait 340 ms avant le début des gradients subséquents. Les gradients sont les périodes d'acquisition d'un volume cérébral durant lesquelles un bruit important est généré par le scanner. Ce bruit est dû à l'alternance des champs magnétiques intenses. Des électrodes amagnétiques en chlorure d'argent ont été utilisées, reliées à des amplificateurs différentiels à faible bruit. De plus, les signaux EEG sont parasités par l'activité cardiaque du sujet. Par conséquent, il est nécessaire de réaliser une acquisition simultanée de l'électrocardiogramme (ECG), à l'aide d'un amplificateur (Physiogard, Bruker SARL, Wissembourg France), afin de pouvoir procéder au filtrage de cet artefact cardiaque (Otzenberger, Gounot, Marrer, Namer & Metz-Lutz, 2005). Les signaux EEG et ECG ont été enregistrés à la fréquence de 1000 Hz. Dix-neuf électrodes d'intérêt ont été utilisées durant chaque session, elles étaient placées selon une disposition normalisée appelée « système 10-20 ».

3 Résultats

3.1. Résultats : Comportementaux

Une interaction significative est observée entre nos deux modalités et les différentes paires pour les scores de discrimination ($F(3,45) = 11{,}30$ $p < 0.0001$) et pour les temps de réponse ($F(3,45) = 3.05$ $p < 0.04$). Les paires syllabiques discriminées le moins efficacement sont celles mettant en jeu la labialité en AV statique (82,8 %) ; c'est aussi l'indice le plus lentement identifié (827 ms). Les stimuli non phonologiques en AV statique (92,5 %) et en AV dynamique (95 %) sont significativement moins bien discriminées que les autres, hormis les paires s'opposant sur le lieu d'articulation en AV statique (95,8 %).

3.2. Résultats : Potentiels évoqués

La figure 1 montre les potentiels évoqués recueillis sur l'électrode Fz (vertex frontal). On peut observer deux pics significatifs en modalité AV dynamique (ligne rouge) une onde positive autour de 150 ms et une négative autour de 250 ms. On constate que l'amplitude des potentiels évoqués en modalité AV dynamique est supérieure à celle enregistrée en modalité AV statique, et ce, de façon significative autour de 250 ms. Pour les trois contrastes (labialité:

T = 4.67 p < 10^{-4}; lieux d'articulation : T = 3.37 p < 0.008 ; voisement : T = 2.95 p < 0.009).

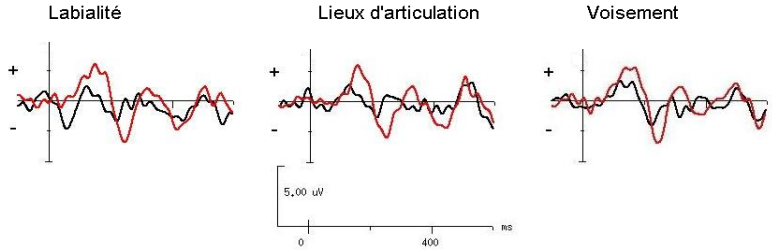

FIGURE 1 : Potentiels évoqués par les réponses correctes à la discrimination des trois contrastes phonologiques (en rouge : AV dynamique ; en noir : AV statique). La fenêtre temporelle s'étend de -100 à + 400 ms ; la ligne verticale représente le début du second stimulus.

3.3. Résultats : IRMf

La figure 2 montre les résultats IRMf pour les réponses aux paires différentes dans les deux modalités. Le cortex temporal supérieur est recruté dans les deux hémisphères pour les deux modalités et par tous les types de stimuli. La région MT/V5 du cortex occipital apparait activée uniquement en AV dynamique. De plus, une zone du cortex prémoteur est activée par les contrastes de voisement et de lieux d'articulation, lors de la présentation AV dynamique.

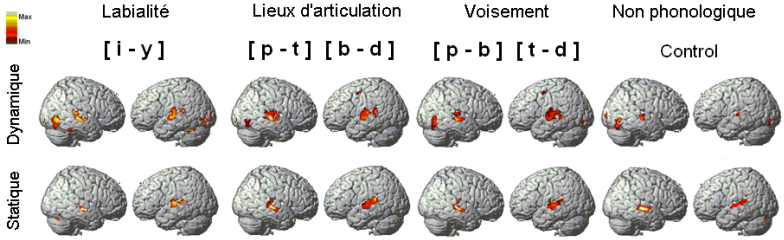

FIGURE 2 : Activations significatives à un niveau p < .005, ayant une étendue supérieure à 25 voxels.

4. Discussion

L'influence de la dimension dynamique des mouvements visuels se caractérisent dans nos résultats à deux niveaux. Tout d'abord, les potentiels évoqués par la modalité AV statique ont une amplitude plus faible que ceux évoqués en modalité AV dynamique. L'étude de Ponton, Auer & Bernstein (2002) rapporte un effet similaire, observant un renforcement lors des présentations audiovisuelles de l'onde. Nous n'observons pas de différence significative entre les deux modalités à l'instar de l'étude de Van Wassenhove *et al.* 2005. L'influence des mouvements dynamiques apparait de façon précoce dès 150 ms. et se poursuit aux alentours de 250 ms. La soustraction des activations IRMf suscitées par la modalité AV dynamique, par

rapport à la modalité AV statique, met en évidence, dans les deux hémisphères, des activations significativement plus élevées dans la partie inférieure du cortex occipito-temporal, et localisées dans les aires de Brodmann 19 et 37. Cette région est activée par tous types de mouvements (Tong, 2003), incluant les mouvements orofaciaux inhérents à la production de la parole. Contrairement aux données de Calvert *et al.* (1997), les résultats de la soustraction entre les cartes d'activation obtenues, lors de la présentation AV dynamique par rapport à l'AV statique, ne montrent pas d'activation plus marquée du cortex auditif primaire (gyrus de Heschl / Aires de Brodmann 41 et 42), en présence des seuls mouvements orofaciaux langagiers. Nos résultats plaident en faveur d'une modulation de l'activité cérébrale lors de la perception AV dynamique, sans pour autant qu'une région soit dévolue spécifiquement au traitement des indices visuels. L'activation du cortex prémoteur évoque les neurones miroirs (Rizzolatti & Craighero, 2004). L'éventualité que ceux-ci prennent part dans la perception de la parole est encore en débat (Hickok, 2009 ; Skipper, van Wassenhove, Nusbaum & Small, 2007), mais semble appuyée par nos résultats. Les premières questions relatives au rôle et l'implication des aires motrices dans la perception verbale ont été soulevées par la présence d'activation de l'aire de Broca, dans des tâches n'impliquant aucune production articulée (Price, Wise, Warburton, Moore, Howard, Patterson, Fracowiak & Friston, 1996). La découverte du système des neurones miroirs a multiplié les interrogations sur l'implication des régions motrices dans la perception de la parole (Gallese, Fadiga, Fogassi & Rizzolatti, 1996 ; Rizzolatti, Fadiga, Gallese & Fogassi, 1996). Le modèle à deux voies de Hickok & Poeppel (2007) postule que l'aire de Broca et la partie supérieure de l'aire prémotrice (AB 6) font partie de la voie dorsale. Dans notre étude, la discrimination, en modalité AV dynamique, des syllabes s'opposant sur le voisement et sur les lieux d'articulation est associée à l'activation significative d'une région du cortex prémoteur (AB 6). Les coordonnées des pics observés (voisement : x = -50 ; y = -2 ; z = 52 ; lieux d'articulation : x = -50 ; y = -4 ; z = 52) sont à mettre en parallèle avec celles rapportées (x = -50 ; y = -6 ; z = 47) dans l'étude de Wilson, Saygin, Sereno, & Iacoboni (2004). Ces auteurs ont constaté un recouvrement des activations liées à la perception passive de syllabes CV, avec celles liées à la production orale des mêmes syllabes. Contrairement à ces travaux, nos analyses en cerveau entier révèlent des activations du cortex prémoteur, alors que Wilson *et al.* (2004) ont procédé à une analyse en régions d'intérêt. Néanmoins, n'ayant pas mené de phase de localisation motrice, nous pouvons seulement constater la similarité de nos pics. L'implication du cortex prémoteur peut constituer un lien entre la perception et la production de la parole, *via* le réseau articulatoire de la voie dorsale tel que proposé par le modèle de Hickok et Poeppel (2007). À l'instar de ce modèle qui reconnait une influence réciproque des deux voies l'une sur l'autre, le concept de représentation perceptuo-motrice avancé par Schwartz, Basirat, Ménard, Sato (2010) retient notre attention. Les auteurs considèrent que l'action façonne la perception et vice-versa. Certaines composantes du système des neurones miroirs pourraient constituer un substrat important dans un tel mécanisme. Au vu de nos résultats, on peut envisager que dans des conditions perturbées, la composante motrice des représentations intervienne dans le processus de perception de la parole. Parmi les nombreuses questions qui demeurent, on peut s'interroger sur la nécessité de la perception des indices visuels dans un tel mécanisme. En effet, si la perception visuelle est nécessaire pour l'accès à la composante motrice des représentations, pourquoi n'observons-nous pas des activations du cortex prémoteur lors de la discrimination des paires s'opposant sur le degré de labialisation des voyelles ? Cela pourrait suggérer que l'implication du cortex prémoteur est davantage liée à la composante motrice des représentations plutôt qu'à la perception des indices visuels (Dubois, Otzenberger, Gounot, Sock, Metz-Lutz, 2012)

Remerciements

Ce travail a été financé par un programme de la Maison Interuniversitaire des Sciences de l'Homme Alsace (MISHA), 2008-2012 « Perturbations et Réajustements : parole normale vs. parole pathologique », par une ANR "DOCVACIM" attribuée à l'Institut de Phonétique de Strasbourg / U.R. LiLPa, E.R. Parole et Cognition et par le projet du CS de UdS Gutenberg-Strasbourg, 2009-2011.

Références

CALVERT, G. A., BULLMORE, E. T., BRAMMER, M. J., CAMPBELL, R., WILLIAMS, S. C. R., MCGUIRE, P. K., WOODRUFF, P. W. R., IVERSEN, S. D. & DAVID, A. S. (1997). Activation of auditory cortex during silent lipreading. *Science, 276*, 593-596.

DEBENER, S., ULLSPERGER, M., SIEGEL, M., FIEHLER, K., VON CRAMON, Y. D. & ENGEL, A. K. (2005). Trial-by-trial coupling of concurrent electroencephalogram and functional magnetic resonance imaging identifies the dynamics of performance monitoring. *The Journal of Neurosicence, 25*(50).

DUBOIS, C., OTZENBERGER, H., GOUNOT, D., SOCK, R. & METZ-LUTZ, M.-N. (2012). Visemic processing in audiovisual discrimination of natural speech: A simultaneous fMRI-EEG study. Neuropsychologia

GALLESE, V., FADIGA, L., FOGASSI, L. & RIZZOLATTI, G. (1996). Action recognition in the premotor cortex. *Brain, 119*, 535-609.

HICKOK, G. (2009). Eight Problems for the Mirror Neuron Theory of Action Understanding in Monkeys and Humans. *Journal of Cognitive Neuroscience, 21*(7), 1229-1243.

OTZENBERGER, H., GOUNOT, D., MARRER, C., NAMER, I. J. & METZ-LUTZ, M.-N. (2005). Reliability of Individual Functional MRI Brain Mapping of Language. *Neuropsychology, 19*(4).

PONTON, C. W., AUER, E. T. & BERNSTEIN, L. E. (2002). Neurocognitive basis for audio-visual speech perception: evidence from event-related potentials. *7th International Conference on Spoken Language Processing, DENVER, USA.* .

PRICE, C. J., WISE, R. J. S., WARBURTON, E. A., MOORE, C. J., HOWARD, D., PATTERSON, K., FRACOWIAK, R. S. J. & FRISTON, K. J. (1996). Hearing and saying The functional neuro-anatomy of auditory word processing. *Brain, 119*.

RIZZOLATTI, G. & CRAIGHERO, L. (2004). The mirror-neuron system. *Annual Review 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*, 131-141.

ROSS, L. A., SAINT-AMOUR, D., LEAVITT, V. M., JAVITT, D. C. & FOXE, J. J. (2006). Do You See What I Am Saying? Exploring Visual Enhancement of Speech Comprehension in Noisy Environments. *Cerebral Cortex 10*.

SCARBOROUGH, R., KEATING, P., BARONI, M., CHO, T., MATTYS, S., ALWAN, A., AUER, E. J. & BERNSTEIN, L. (2006). Optical Cues to the Visual Perception of Lexical and Phrasal Stress in English. *UCLA Working Papers in Phonetics, 105*.

SCHWARTZ, J. L., BERTHOMMIER, F. & SAVARIAUX, C. (2004). Seeing to hear better: evidence for

early audio-visual interactions in speech identification. *Cognition 93*.

SCHWARTZ, J.-L., BASIRAT, A., MÉNARD, L., SATO, M. (2010) The Perception-for-Action-Control Theory (PACT): A perceptuo-motor theory of speech perception. *Journal of Neurolinguistics,* vol. In Press, Corrected Proof.

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*.

SUMBY, W. H. & POLLACK, I. (1954). Visual Contribution to Speech Intelligibility in Noise. *JASA, 26*(2).

TONG, F. (2003). Primary visual cortex and visual awereness. *Cognitive Neuroscience, 4*.

VAN WASSENHOVE, V., GRANT, K. W. & POEPPEL, D. (2005). Visual speech speeds up the neural processing of auditory speech. *PNAS, 102*(4), 1181-1186.

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*(7), 701-702.