HAL
open science

Dynamique temporelle du liage dans la fusion de la parole audiovisuelle

Olha Nahorna, Frédéric Berthommier, Jean-Luc Schwartz

To cite this version:

Olha Nahorna, Frédéric Berthommier, Jean-Luc Schwartz. Dynamique temporelle du liage dans la fusion de la parole audiovisuelle. JEP-TALN-RECITAL 2012 - conférence conjointe 29e Journées d'Études sur la Parole, 19e Traitement Automatique des Langues Naturelles, 14e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues, Jun 2012, Grenoble, France. pp.65. hal-00773192

HAL Id: hal-00773192

https://hal.science/hal-00773192v1

Submitted on 11 Jan 2013

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Dynamique temporelle du liage dans la fusion de la parole audiovisuelle

Olha Nahorna, Frédéric Berthommier, Jean-Luc Schwartz
GIPSA-Lab - DPC
UMR 5216 –CNRS Uiversité de Grenoble
Olha.Nahorna, Jean-Luc.Schwartz, Frederic.Berthommier@gipsa-lab.grenoble-inp.fr
http://www.gipsa-lab.inpg.fr

RESUME

L'effet McGurk met en évidence le phénomène de fusion audiovisuelle : le montage d'un son « ba » avec une vidéo « ga » est souvent perçu comme « da ». Dans un travail précédent nous avons montré que la fusion audiovisuelle peut-être modulée par un processus de liage préalable (Nahorna et al., 2011, 2012). Dans ces expériences, un stimulus McGurk était précédé par un contexte audiovisuel cohérent ou incohérent (son correspondant ou non à la vidéo) et nous avons observé que dans le cas de contexte incohérent l'effet McGurk diminue. Cet effet se produit pour des contextes variant entre 3 et 10 secondes, sans effet significatif de la durée de contexte dans cette plage. Dans le travail actuel, nous étudions des durées de contexte plus courtes. Les résultats montrent qu'une seule syllabe est suffisante pour délier les flux auditif et visuel et produire une forte diminution de l'effet McGurk.

ABSTRACT

Temporal dynamics of binding in audiovisual speech fusion

The McGurk effect demonstrates the phenomenon of audiovisual fusion: a sound "ba" mounted on a video "ga" is often perceived as "da". In a previous work we showed that audiovisual fusion might be modulated by a precedent binding process (Nahorna et al., 2011, 2012). In these experiments a McGurk stimulus was preceded by an audiovisual coherent or incoherent context (sound corresponding or not to the video) and we observed a decrease of the McGurk effect in the incoherent context case. This effect occurs for contexts varying from 3 to 10 seconds, with no significant effect of the context duration in this range. In the present work we study shorter context durations. The results show that one syllable is sufficient to unbind the auditory and visual streams and to produce a strong decrease in the McGurk effect.

MOTS-CLES : Effet McGurk, liage, fusion multisensorielle, perception de la parole audiovisuelle, analyse de scène audiovisuelle.
KEYWORDS : McGurk effect, binding, multisensory fusion, audiovisual speech perception, audiovisual scene analysis.

1 Introduction

Le signal visuel joue un rôle important dans la perception de la parole. L'effet "cocktail party" (Cherry, 1953), les gains d'intelligibilité dans le bruit grâce à la lecture labiale (Sumby et Pollack, 1954), l'effet McGurk (McGurk et McDonald, 1976) montrent bien l'influence de l'information visuelle sur la parole perçue. Jusqu'à présent il n'y a pas de consensus dans la communauté scientifique sur la convergence audiovisuelle et la vision classique considère que l'information des modalités différentes est extraite et traitée indépendamment avant convergence. Plusieurs architectures de fusion audiovisuelle sont proposées dans la littérature. Schwartz et al. (1998) les résument en quatre catégories, selon l'existence et la nature d'une éventuelle représentation

commune du son et de l'image. Campbell et al. (2008) assignent 2 rôles fonctionnels distincts au signal visuel dans la parole : un rôle de complémentarité où le signal visuel permet de préciser ou rajouter l'information manquante dans le flux de parole auditif, et un rôle de redondance / corrélation, où la vision duplique partiellement l'information de la dynamique articulatoire. Campbell et al. considèrent ces deux rôles comme indépendants et parallèles, mais nous pensons quant à nous que le traitement de la parole AV pourrait impliquer deux étapes, la corrélation des deux entrées étant évaluée au préalable et conditionnant un processus de liage avant la fusion (exploitant elle la complémentarité de Campbell et al.). Ainsi nous pensons que les résultats d'évaluation de corrélation peuvent moduler le niveau de fusion, en indiquant quelle partie du signal visuel peut être prise en compte. Par rapport à la vision classique, nous considérons donc qu'il n'y a pas d'indépendance totale avant convergence et fusion, mais au contraire une interaction à bas niveau permettant d'alimenter un processus de liage modulant la fusion.

L'hypothèse de l'existence de plusieurs niveaux de traitement n'est pas nouvelle (voir Schwartz et al., 2004). Pour rendre compte de ce type de phénomène, Berthommier (2004) a proposé un modèle dans lequel la fusion audio-visuelle est précédée d'un niveau primitif et pré-phonétique. Ainsi, ce modèle postule deux niveaux d'interaction audiovisuelle, un niveau précoce (détection) et un niveau tardif (fusion). Dans notre travail précédent (Nahorna et al., 2011, 2012) nous avons montré que le mécanisme de détection précoce fait partie d'un système plus large assurant un rôle de liage conditionnel. Ce système permet, au cas par cas, de lier les entrées auditives et visuelles, ou au contraire de les séparer. Cet effet apparaît par exemple dans le cas de films doublés, où les entrées auditive et visuelle ne sont pas intégrées dans la reconnaissance qui reste purement auditive.

Pour démontrer cela nous avons construit des situations expérimentales où on peut « débrancher » le niveau de fusion. Nous avons pris l'effet McGurk comme un indicateur de la fusion et cherché à modifier ou supprimer l'effet McGurk en faisant varier le contexte préalable, qui permet de lier/délier les flux auditif et visuel. Nos résultats montrent que par une manipulation du contexte contrastant contexte « cohérent » et « incohérent » (selon que le flux audio est cohérent ou non avec le flux vidéo dans le contexte), on peut produire un « décrochage » du lien audiovisuel, conduisant à une diminution de la fusion (Nahorna et al., 2011, 2012). Dans ces travaux nous avons testé notre hypothèse avec des durées de contexte variables entre 3 et 10 secondes. Nous avons observé une diminution d'effet McGurk en contexte incohérent quelle que soit sa durée, mais pas de différence d'effet McGurk selon la durée du contexte. Dans la présente étude, nous nous demandons si un décrochage de fusion peut se produire avec des durées de contexte incohérent plus courtes et nous évaluons la durée de contexte incohérent minimale nécessaire pour que le décrochage se produise.

2 Méthodologie

Notre paradigme expérimental consiste à présenter à des sujets un flux audiovisuel et de leur demander de détecter en ligne les syllabes cible « ba » ou « da ». Le sujet ne connaît pas a priori la position des cibles dans le flux audiovisuel. Nos stimuli consistent en une cible précédée par un contexte cohérent ou non. Nous avons deux types de cibles : une cible congruente « ba » (audio « ba » + vidéo « ba »), dont on attend qu'elle soit correctement identifiée « ba », et une cible incongruente « McGurk » (audio « ba » + vidéo « ga »), dont on attend qu'elle soit souvent perçue « da ».

Nous construisons trois types de contexte : cohérent (C), incohérent (I) et incohérent phonétique (P). Le contexte cohérent consiste en une séquence de syllabes audiovisuelles : le sujet voit donc le visage du locuteur qui prononce des syllabes synchronisées avec les syllabes audio. Dans le contexte incohérent, nous cherchons à produire une incohérence maximale, en associant le même matériel audio avec la vision du même locuteur, qui prononce de la parole quelconque et

non pas des syllabes. Le contexte incohérent phonétique (ou « phonétique » par la suite) est destiné à produire un niveau d'incohérence intermédiaire, où les syllabes apparaissent au même moment, mais diffèrent phonétiquement. Pour ce faire, nous associons au contenu vidéo du contexte cohérent (séquences de syllabes), un contenu audio dans lequel les syllabes sont remplacées aléatoirement les unes par les autres (permutées) tout en gardant un timing adéquat (synchronisation du son et de l'image, mais incohérence de contenu phonétique) (on trouvera des exemples de stimuli dans http://www.gipsa-lab.inpg.fr/~jean-luc.schwartz/fichiers_public_JLS/AV_Binding_demo/AV_Binding_Demo.html). Pour disposer d'une condition de base pour nos analyses et réflexions nous avons aussi ajouté une condition « sans contexte », où nous ne présentons que la cible pure. La durée des contextes est variable entre 0 et 5 syllabes (soit entre 0 et 3 secondes) (Figure 1).

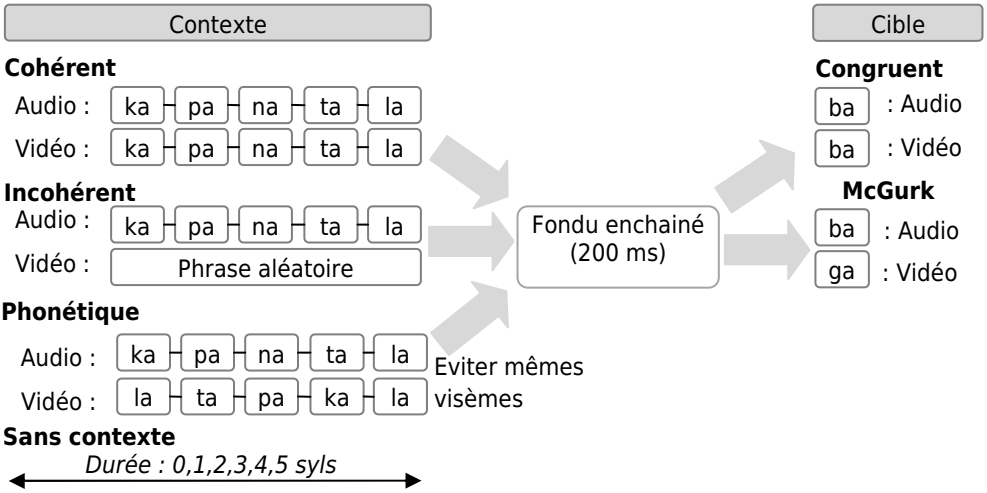

FIGURE 1 – Paradigme expérimental

2.1 Stimuli

Pour préparer l'expérience, nous avons enregistré des séquences avec des syllabes et de la parole quelconque de durée variée, se terminant toujours par la cible « ba » ou « ga ». Le contexte syllabique est constitué de séquences aléatoires de syllabes françaises (syllabes CV, C étant une plosive ou une fricative, à l'exclusion des syllabes «ba », « da » et » ga », soit 13 syllabes possibles : « pa », « ta », « va », « fa », « za », « sa », « ka », « ra », « la », « ja », « cha », « ma », « na » prononcées par un locuteur français, JLS, avec les lèvres maquillées en bleu), enregistré sur un rythme d'environ 1.5 Hz. Dans le contexte de la parole quelconque, le locuteur devait parler librement sur le sujet de son choix. Tous les fichiers acoustiques étaient globalement normalisés en intensité pour assurer qu'ils soient présentés au même niveau sonore global.

A partir de ces séquences nous avons construit 4 exemplaires de contextes audiovisuels pour chaque durée de contexte (1, 2, 3, 4, 5 syllabes), avec trois types de contexte et deux types de cible, soit 120 exemplaires de contextes. Pour préparer les cibles McGurk nous avons extrait la dernière syllabe « ga » enregistrée dans une groupe de séquences syllabiques et nous avons fait un montage audio en remplaçant la syllabe « ga » par une syllabe « ba », prise dans l'autre groupe de séquences se terminant par « ba ». Nous avons repéré et sélectionné l'instant de l'explosion de la consonne plosive comme le repère de montage. La cible montée a été ensuite normalisée en amplitude.

Un stimulus complet consiste en un exemplaire de contexte suivi d'une cible. Comme nous avons des contextes visuels différents avec de légères modifications de position de la tête, nous avons systématiquement introduit un fondu enchaîné progressif noir sur 5 images pour minimiser la perturbation perceptive entre contexte et cible. Chaque stimulus complet est séparé du suivant par une pause de 840 ms qui consiste à voir une image fixe du même locuteur avec du silence. Nous avons besoin de cette pause pour que le sujet puisse prendre sa décision avant que le prochain stimulus arrive.

Les cibles « ba » sont des contrôles et ne présentent pas d'intérêt direct dans cette expérience, puisque nous prédisons qu'elles devraient être identifiées correctement « ba » quel que soit le contexte. Seuls les stimuli McGurk nous intéressent, la prédiction étant qu'ils produisent moins de réponses de fusion « da » (donc plus de réponses « auditives » « ba ») dans le cas de contexte incohérent et phonétique. Les données empiriques montrent que l'effet McGurk apparaît en moyenne dans 35-50% des cas, tandis que les stimuli « ba » produisent des réponses « ba » dans presque 100% des cas. Pour équilibrer dans notre expérience la fréquence attendue des réponses « ba » et « da », et pour optimiser le nombre de cibles «McGurk » qui concentrent notre intérêt, nous avons décidé de présenter les stimuli dans les proportions : ¼ des stimuli « ba » et ¾ des stimuli « McGurk ». Au total nous avons donc présenté 256 stimuli répartis aléatoirement (64 cibles congruentes « ba » et 192 cibles incongruentes McGurk) dans un bloc de 14 minutes (les différentes conditions de stimuli et de contexte sont donc mélangées au sein du bloc).

2.2 Procédure expérimentale

Les instructions données aux sujets étaient de détecter en ligne les syllabes « ba » ou « da » (tâche de « monitoring » syllabique avec un choix forcé de réponse) et d'y répondre le plus rapidement possible en appuyant sur le bouton correspondant, sans savoir quand ils apparaissent dans la séquence. Ainsi, les réponses peuvent apparaître à tout moment. L'ordre des boutons était également distribué parmi tous les sujets.

L'expérience a été conduite dans une chambre sourde en utilisant le logiciel Presentation® (Version 0.70, www.neurobs.com). Le signal sonore était présenté sous casque avec un niveau de volume confortable et fixe pour tous les sujets (environ 60 dB SPL). Le signal visuel était présenté sur un moniteur avec un taux de 25 images/s. Le sujet était positionné à environ 50 cm de l'écran pour être dans une position confortable.

2.3 Analyse des réponses

Pendant l'expérience les stimuli sont fournis en ligne, et le sujet peut répondre à chaque instant, qu'il y ait une cible ou non. Il peut donc se produire deux types d'erreurs : fausses alarmes (la présence d'une réponse « ba » ou « da » en l'absence de cible) ou absence de réponse à une cible. Pour traiter correctement les réponses, nous avons mis en place la méthodologie suivante. Pour chaque stimulus, nous comptons les réponses qui sont apparues après sa présentation (repérée par l'instant d'explosion acoustique de la plosive dans la cible) et avant la cible suivante, puisque nous avons limité la validité temporelle de réponse dans une fenêtre de 1200 ms. Dans les expériences précédentes, nous avons vérifié que la plupart des réponses données par les sujets rentrent dans cette période. S'il n'y a pas de réponse dans cet intervalle, on compte une « absence de réponse » pour ce stimulus. S'il y a plusieurs réponses, on fait une vérification de l'identité des réponses, si elles sont identiques, nous prenons la première d'entre elles, sinon nous les éliminons toutes et considérons une « absence de réponse » pour ce stimulus. Le taux de non-réponses pour toute l'expérience est 5,8%. Ce score assez élevé n'est pas surprenant, vu que les sujets étaient limités dans le temps et que les cibles McGurk peuvent être perçues différentes de « ba » ou « da » en français (Cathiard et al., 2001).

3 Résultats

20 sujets français ont participé à cette étude (16h et 4f), avec parmi eux 19 droitiers et 1 gaucher. Nos hypothèses sont que l'effet McGurk, estimé par la proportion des réponses « da » sur les cibles incongruentes McGurk doit diminuer dans le cas des contextes incohérent et phonétique par rapport au contexte cohérent. L'effet McGurk peut aussi dépendre de la durée d'un contexte. La quantité des réponses « ba » et « da » est calculée pour chaque sujet et chaque condition (contexte, durée d'un contexte, cible congruente vs McGurk).

Des ANOVAs à mesures répétées ont été effectuées sur les proportions de réponses « ba » sur la totalité de réponses « ba » plus « da » en ignorant les cas d'absence de réponse. Ces taux de réponses ont été transformés en asin(sqrt) pour assurer une distribution quasi gaussienne des variables. Nous avons systématiquement vérifié que nos résultats ne différent pas en faisant l'analyse sur les proportions de réponses « ba » rapportées au nombre total de stimuli (« ba » plus « da » plus « réponses absentes ») ou sur la proportion de réponses « da » rapportées à la totalité des réponses. Nous avons systématiquement exclu la condition « sans condition » ou « durée de contexte 0 syllabes » de l'analyse ANOVA, vu que le nombre de stimuli présentés aux sujets est différent par rapport aux autres conditions de contexte. Mais nous présentons systématiquement les scores associés à cette condition pour disposer d'un repère. Nous avons également effectué des ANOVAs à mesures répétées sur les temps de réponse, en appliquant un logarithme pour assurer la normalité.

3.1 Taux de réponses

Il apparaît que les cibles Ba ont été bien identifiées dans tous les contextes (Figure 2). Les cibles McGurk produisent des taux d'identification « ba » moindres. L'ANOVA à deux facteurs « cible », « contexte » confirme l'effet significatif du facteur « cible » ($F(1,19)=55.1$, $p<.001$). Dans le cas de contexte cohérent nous avons obtenu ~55% d'effet McGurk. Ce résultat est classique en français (Cathiard et al., 2001) et typiquement plus réduit qu'en anglais (Colin and Radeau, 2003).

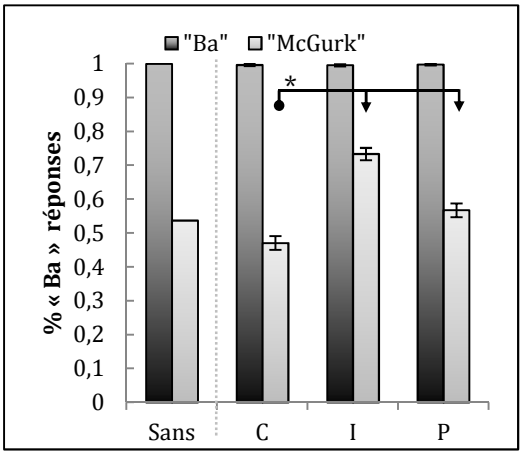

FIGURE 2 –Taux de réponses. Pourcentage de réponses « ba » rapportées a l'ensemble des réponses (« ba »/(« ba » + « da »)) ; On a indiqué les différences significatives (voir données précises dans le texte).

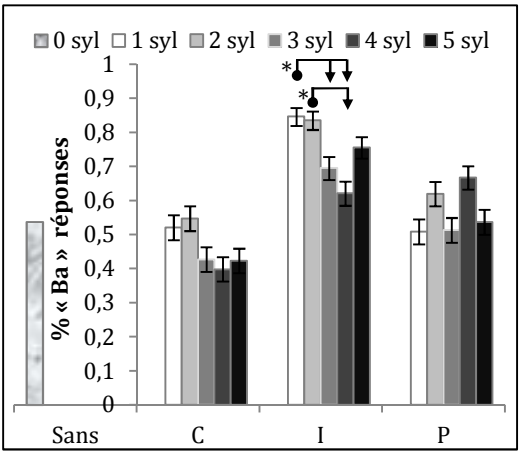

FIGURE 3 – Facteur durée pour les cibles McGurk - Pourcentage de réponses « ba » rapportées à l'ensemble des réponses (« ba »/(« ba » + « da »))

Le facteur « contexte » est aussi significatif (F(2.2,40.8)=10.3, p<.001), ce qui montre que le nombre des réponses « ba » augmente dans les contextes phonétique et incohérent, essentiellement grâce aux cibles McGurk comme indiqué par l'interaction significative entre « cible » et « contexte » (F(3,57)=14.7, p<.001). Une analyse post-hoc confirme l'augmentation des proportions de réponses « ba » pour les cibles McGurk dans les contextes phonétique (55%) et incohérent (75%) par rapport au contexte cohérent (45%) (p<.05). Donc nous observons un déliage dans les deux contextes incohérents, et moins fort pour le contexte phonétique, qui est moins incohérent.

L'autre question principale de cette étude est l'évaluation de l'effet McGurk selon la durée de contexte. Nous avons donc fait une seconde ANOVA à deux facteurs « durée », « contexte » centrée sur les cibles McGurk, avec un effet significatif de ces facteurs (« durée » F(4,76)=7.2, p<.001 ; « contexte » F(2,38)=46.7, p<.001, interaction F(8,152)=4.7, p<.001).

L'analyse post-hoc indique que la signification de l'effet « durée » est due plutôt au contexte incohérent, où l'effet McGurk est globalement plus faible pour les durées 1,2 syllabes que 4 syllabes (p<.005) (Figure 3). L'autre résultat important qui nous pouvons tirer de cette analyse est que la réduction de l'effet McGurk se produit dès les durées d'incohérence les plus courtes. Donc une syllabe est suffisante pour décrocher jusqu'à un certain point les flux auditif et visuel.

3.2 Temps de réponse

Nous avons effectué une ANOVA sur les temps de réponse avec les facteurs « cible » et « contexte », avec un effet significatif du seul facteur « cible » (F(1,19) =37.9, p<.001). Il y a ainsi une différence de durée de réponse entre les cibles Ba (600 ms) et cibles McGurk (675 ms) (Figure 4), probablement due à l'incongruence dans une cible McGurk, qui prendrait alors plus de temps de traitement et d'identification, ce qui est un résultat classique. Par contre, le temps de réponse ne varie pas en fonction du contexte. Ce résultat, très intéressant, avait déjà été obtenu dans nos études précédentes (Nahorna et al., 2011, 2012), mais sans obtenir alors d'effet suffisant de la cible pour être concluant. Nous obtenons donc une confirmation d'un fait très intéressant : le contexte module l'effet McGurk mais ne module pas les temps de réponse associés.

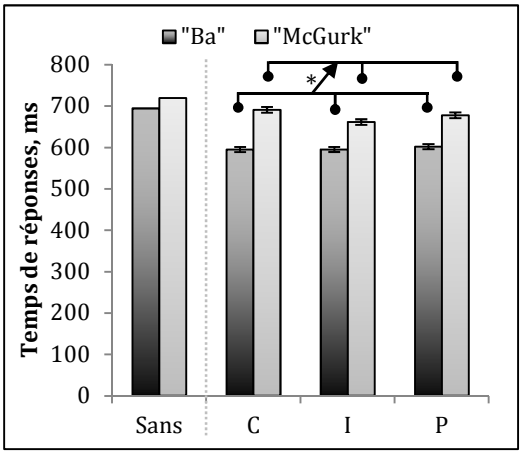

FIGURE 4 – Temps de reponse

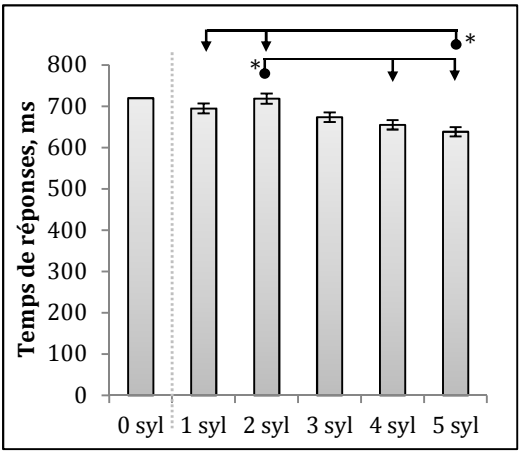

FIGURE 5 – Temps de réponse pour les cibles McGurk

L'ANOVA centrée sur les stimuli McGurk et sur le facteur « durée » (tous contextes confondus) donne un résultat significatif ($F(4,76)= 7.1$, $p<.001$). L'analyse post-hoc ($p<.05$) montre que globalement les durées de contexte courtes (1, 2 syllabes) produisent des temps de réponse significativement plus longs que des durées plus longues (4 et 5 syllabes) (figure 5).

4 Discussion et conclusion

Dans cette expérience, nous avons réussi une nouvelle fois à démontrer un effet de déliage du niveau de fusion par un contexte préalable. Ces résultats confirment les résultats obtenus précédemment (Nahorna et al., 2011, 2012). Les contextes incohérent et phonétique sont suffisants pour produire un déliage, mais l'effet est moins fort dans le contexte phonétique, qui est aussi moins incohérent. Ceci confirme notre hypothèse d'un schéma de fusion audiovisuelle de la parole à deux étapes, où une première étape est un étage de liage/déliage qui évalue la cohérence de deux signaux (Figure 6).

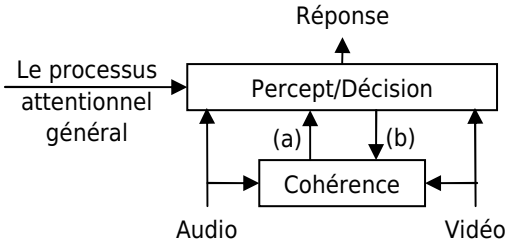

FIGURE 6 – Modèle à deux étages

Le nouveau résultat principal de ce travail est que la dynamique temporelle du processus de liage suggère un effet de déliage très rapide : une incohérence d'une syllabe est suffisante pour produire un décrochage des deux flux.

Sur la Figure 3, nous avons vu que dans le contexte incohérent les durées de contexte très courtes sont plus perturbantes que les durées longues. Nous n'avons pas d'interprétation claire de cet effet inattendu. On pourrait proposer une première piste qui serait l'existence d'un effet d'adaptation sur le déliage, avec amplification de l'effet à durées courtes, puis saturation et décroissance à durées plus longues. Evidemment, cette hypothèse, basée sur un effet faible et n'apparaissant pas dans le cas du contexte phonétique (ou tous les effets sont réduits) reste à tester et à confirmer.

Nous observons par ailleurs une tendance significative et claire de diminution de temps de réponse entre durées de contexte courtes (1,2 syllabes) et longues (4,5 syllabes) (figure 5) et on pourrait modéliser globalement l'effet durée du contexte par une régression linéaire. Une interprétation simple et logique est que cet effet est dû à la « surprise » du sujet qui voit apparaître très vite une cible après le début du stimulus, et répond ainsi plus lentement que si le contexte est plus long.

On peut alors se demander si l'existence d'un pic de déliage pendant les deux premières syllabes dans le contexte incohérent pourrait être une conséquence de cet effet « surprise », susceptible de produire une charge cognitive, dont on sait qu'elle peut diminuer l'effet McGurk (Alsius, 2005). Ceci pourrait fournir une seconde explication à la décroissance de l'effet McGurk dans le cas de contexte incohérent d'une durée de 1-2 syllabes à une durée de 4 syllabes.

Néanmoins, si l'hypothèse de charge cognitive avait un rôle majeur, alors nous devrions l'observer également dans la condition « sans contexte », où la durée de contexte est la plus faible. Or évidemment ceci ne se produit pas puisque l'effet McGurk est quasiment maximal en l'absence de contexte. Donc, même si l'effet de charge cognitive pouvait jouer un rôle mineur dans nos résultats, ceci ne remet pas en cause le résultat principal de nos travaux : l'existence d'un processus de déliage, susceptible de moduler la fusion audiovisuelle en perception de parole.

C'est sur la caractérisation cognitive et neurophysiologique de ce processus et sur ses implications pour le traitement de la parole dans le cerveau humain que nous continuerons donc à porter nos efforts par la suite.

Remerciements

Cette étude est financée par le projet ANR-08-BLAN-0167 MULTISTAP

Références

ALSIUS, A., NAVARRA, J., CAMPBELL, R., & SOTO-FARACO, S.S. (2005). Audiovisual integration of speech falters under high attention demands. *Current Biology* **15**, 839-843.

BERTHOMMIER, F. (2004). A phonetically neutral model of the low-level audio-visual interaction. *Speech Communication* **44**, 31-41.

CAMPBELL, R. (2008). The processing of audio-visual speech : empirical and neural bases. *Philosophical Transactions of the Royal Society of London Biological Science* **363**, 1001-1010

CATHIARD, M.A., SCHWARTZ, J.L. & ABRY, C. (2001). Asking a naive question about the McGurk Effect: why does audio [b] give more [d] percepts with visual [g] than with visual [d]?. *Proceedings of 5th International Conference on Auditory-Visual Speech Processing (AVS 2001)*, 138-142.

CHERRY, E. C. (1953). Some experiments on the recognition of speech, with one and two ears. *Journal of the Acoustical Society of America* 25, pp. 975–979.

COLIN, C., & RADEAU, M. (2003). Les illusions McGurk dans la parole : 25 ans de recherche (The McGurk illusions in speech : 25 years of research). *L'Année Psychologique* **104**, 497-542.

MCGURK H., & MACDONALD J. (1976). Hearing lips and seeing voices. *Nature* **264** (5588): 746–8.

NAHORNA, O., BERTHOMMIER, F., & SCHWARTZ, J.L. (2012). Binding and unbinding the auditory and visual streams in the McGurk effect. *Journal of the Acoustical Society of America,* (en révision).

NAHORNA, O., BERTHOMMIER, F., & SCHWARTZ, J.L. (2011). Binding and unbinding the McGurk effect in audiovisual speech fusion: Follow-up experiments on a new paradigm. *Proceedings of 10th International Conference on Auditory-Visual Speech Processing (AVSP 2011)* IT 2011-08-31

SCHWARTZ, J.L., BERTHOMMIER, F., & SAVARIAUX, C. (2004). Seeing to hear better : Evidence for early audio-visual interactions in speech identification. *Cognition* **93**, B69-B78.

SCHWARTZ, J.-L., ROBERT-RIBES, J. & ESCUDIE, P. (1998). Ten years after Summerfield. a taxonomy of models of audiovisual fusion in speech perception. *Hearing by Eye*, R. Campbell and et al., Eds. Hove, UK: Psychology Press, pp. 85-108.

SUMBY WH, & POLLACK I (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America* **26**, 212-215.