

Liage et fusion audiovisuelle en perception de la parole : on peut « débrancher » l'effet McGurk par un contexte audiovisuel incohérent

Olha Nahorna, Frédéric Berthommier, Jean-Luc Schwartz

GIPSA-Lab 6 DPC, ICP

UMR 5216 6CNRS Université de Grenoble

Olha.Nahorna, Jean-Luc.Schwartz, Frederic.Berthommier@gipsa-lab.grenoble-inp.fr

http://www.gipsa-lab.inpg.fr

ABSTRACT

The McGurk effect demonstrates the existence of a fusion process in audiovisual speech perception: the combination of the sound "ba" with the face of a speaker who pronounces "ga" is frequently perceived as "da". We assume that in the upstream of this phonetic fusion process, there is another early fusion process, which controls the combination of image and sound, and can block it in the case of audiovisual inconsistencies (conditional binding process), as in the case of a dubbed film. To test this early fusion hypothesis, we designed an experiment in which a consistent or inconsistent audiovisual context is placed before McGurk stimuli, and we show that the inconsistent contextual stimulus can remove the effect McGurk.

Keywords: McGurk effect, binding, multisensory fusion, audiovisual speech perception, audiovisual scene analysis.

1. INTRODUCTION

La perception visuelle fait partie intégrante de la perception de la parole chez les humains. Le célèbre effet McGurk [1] montre bien l'influence de l'information visuelle sur la parole perçue. Le montage du son « ba » avec un film de « ga » est perçu comme « da » chez de nombreux sujets.

Plusieurs architectures de fusion audio-visuelle ont été proposées dans la littérature [2]. Elles ont en commun de considérer des prises d'information auditive et visuelle indépendantes. Or, il y a déjà une quinzaine d'années est apparue l'hypothèse de l'existence de mécanismes précoces pour extraire l'information auditive et visuelle (voir [3]). Pour rendre compte de ce type de phénomène, Berthommier [4] a proposé un modèle dans lequel la fusion audio-visuelle est précédée d'un niveau primitif et pré-phonétique (Figure 1). Le rôle des interactions bas-niveau serait de renforcer la modulation d'amplitude des segments de la parole, sans distorsion des signaux phonétiques, spectrale ou temporelle. Ce niveau précoce permettrait de conditionner les mécanismes de fusion.

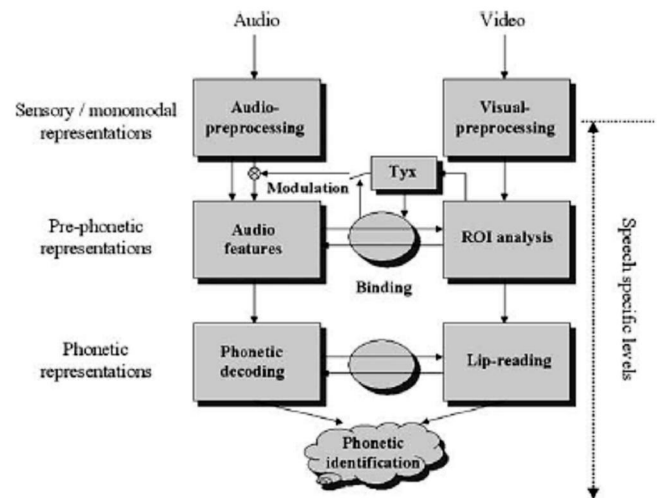


Figure 1 : Un modèle de fusion intégrant une interaction bas-niveau [4]

Ainsi, ce modèle postule deux niveaux d'interaction audiovisuelle, un niveau précoce (détection) et un niveau tardif (fusion). La question expérimentale de ce travail est de savoir si le mécanisme de détection précoce fait partie d'un système plus large assurant un rôle de liage conditionnel. Ce système permettrait, au cas par cas, de lier les entrées auditives et visuelles en un même flux, ou au contraire de les séparer en deux flux différents. Si c'est le cas, on doit pouvoir construire des situations expérimentales où on peut « débrancher » le second niveau de fusion, comme c'est probablement le cas dans les films doublés, où il ne faut pas intégrer les entrées auditive et visuelle dans la reconnaissance, puisqu'elles sont incongruentes et ne portent pas d'information phonétique cohérente.

Nous avons pris l'effet McGurk comme indicateur de la fusion. Nous allons donc essayer de construire un paradigme expérimental visant à supprimer ou modifier l'effet McGurk. Nous cherchons à déterminer si l'effet McGurk résiste à des variations du contexte préalable, qui permettrait de lier/délier les flux auditif et visuel. Nous supposons que par manipulation du contexte, on peut produire un « décrochage » du lien audiovisuel, conduisant à une diminution de la fusion audio-visuelle.

2. MÉTHODOLOGIE

Le paradigme expérimental, consiste à présenter à des sujets un flux de parole audiovisuelle et de leur demander de détecter en ligne la présentation de stimuli « ba » ou « da ». Nous présentons aux sujets deux types de stimuli cibles : un stimulus cohérent « ba » (audio « ba » + vidéo « ba »), dont on attend qu'il soit correctement identifié « ba », et un stimulus « McGurk » (audio « ba » + vidéo « ga »), dont on attend qu'il soit souvent perçu « da ».

Notre hypothèse est que l'effet McGurk disparaît en fonction du contexte préalable. Pour cela nous construisons deux types de contexte : « cohérent » et « incohérent ». Dans le cas cohérent le contexte consiste en une séquence de syllabes, présentées en modalité audiovisuelle : le sujet voit donc le visage du locuteur qui prononce des syllabes synchronisées avec les syllabes audio que le sujet entend. Le contexte incohérent est constitué du même matériel audio, superposé avec la vision du même locuteur, qui prononce de la parole quelconque et non pas des syllabes. Les cibles auditives sont les mêmes dans les deux cas. Comme nous ne savons pas a priori combien de temps il faut présenter le contexte incohérent pour perturber l'effet McGurk, nous utilisons des contextes de durées variables (Figure 2).

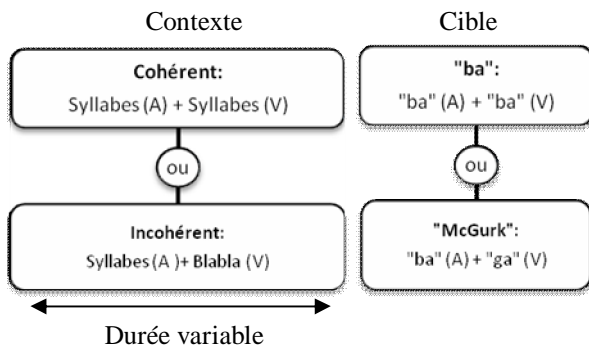


Figure 2 : Principe expérimental

2.1. Mise en oeuvre

Pour préparer l'expérience nous avons enregistré des séquences avec des syllabes et de la parole quelconque. Le contexte acoustique 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 possible) et les syllabes « ba » ou « ga » servent de cible. Nous avons utilisé des contextes préalables incohérents de 5, 10, 15, 20 syllabes.

Les stimuli cibles « ba » ne présentent pas d'intérêt direct dans cette expérience, puisque nous prédisons qu'ils devraient être identifiés correctement « ba » quelque soit le contexte. Seuls les stimuli McGurk nous intéressent, la prédiction étant qu'ils produisent moins de réponses de fusion « da » (et plus de réponses « auditives » « ba ») dans le cas de contexte incohérent. Mais pour être sûr que les sujets sont attentifs pendant toute l'expérience et

répondent correctement, on ne peut présenter uniquement des stimuli McGurk. 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 présenter les stimuli dans les proportions : $\frac{1}{4}$ des stimuli « ba » et $\frac{3}{4}$ des stimuli « McGurk ».

Pour résumer, nous avons 3 facteurs principaux à contrôler dans la préparation de l'expérience :

- Stimuli : $\frac{3}{4}$ de « McGurk » versus $\frac{1}{4}$ de « ba »
- Cohérence : contexte cohérent versus incohérent
- Durée : 5, 10, 15, 20 syllabes.

2.2. Préparation des matériaux expérimentaux

Enregistrement

Nous avons enregistré 80 séquences audiovisuelles de contextes de durée variée, se terminant toujours par la cible « ba » ou « ga » (prononcées par un locuteur français, JLS, avec les lèvres maquillées en bleu). Les 40 séquences destinées à produire le contexte audio pour toute l'expérience, et le contexte vidéo pour le cas de contexte cohérent, sont produites par des arrangements aléatoires de 13 syllabes françaises : « pa », « ta », « va », « fa », « za », « sa », « ka », « ra », « la », « ja », « cha », « ma », « na ». 20 séquences se terminent par une syllabe « ba » et 20 par une syllabe « ga ». La longueur des séquences est 5, 10, 15, 20 syllabes, correspondant à des durées de l'ordre de 3, 7, 10 et 13 s. Les séquences étaient présentées au locuteur sur un écran de contrôle. Le locuteur devait répéter les séquences proposées, en laissant à chaque fois un silence court entre deux syllabes consécutives, de façon à fournir des points de montage acoustique simples.

Les 40 séquences destinées à produire le contexte incohérent consistent en un flux de parole quelconque de durée 4, 7, 10, 13 secondes, se terminant dans la moitié des cas par une séquence « ba » et dans l'autre moitié par une séquence « da ». Le locuteur devait parler librement sur le sujet de son choix, et au bout d'une durée correspondant à la condition correspondante (4, 7, 10, 13 secondes), l'indication de la syllabe terminale apparaissait, indiquant au locuteur qu'il devait conclure en prononçant cette syllabe.

Sélection et montage

Pour préparer les stimuli McGurk nous avons fait un montage audio en remplaçant le son « ga » par le son « ba », pris dans l'autre groupe des séquences avec « ba » à la fin (Fig. 3). Les données ont été ensuite normalisées en amplitude, et sélectionnées sur des critères d'amplitude de mouvement visuel, de manière à ce que les stimuli « cohérents » et « incohérents » ne diffèrent pas en terme de contenu audiovisuel des stimuli cible McGurk [5].

Nous avons ainsi préparé 16 stimuli originaux de chaque type (4 par durée de contexte, avec 4 durées de contexte), et ce pour les 4 types définis par la cible (« ba » vs. McGurk) et le contexte (cohérent vs. incohérent). Nous les avons combinés dans l'expérience avec les proportions : $\frac{3}{4}$ de « McGurk » versus $\frac{1}{4}$ de « ba ». Pour ce faire les stimuli de type McGurk ont été répétés 3 fois. Au total nous avons donc présenté 128 stimuli (16 « ba » dans les 2 contextes, 48 McGurk dans les deux contextes) répartis en 4 blocs de 32 stimuli, répartis aléatoirement.

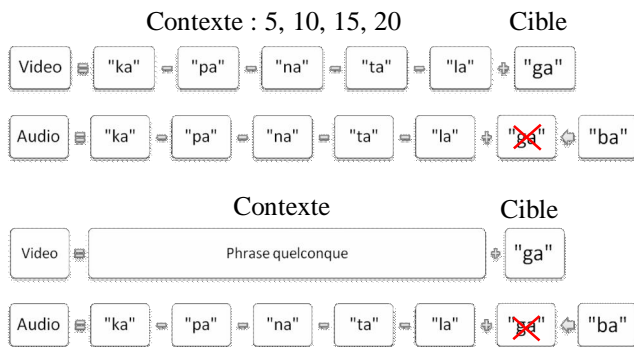


Figure 3 : Montage des stimuli "McGurk" à contexte cohérent (en haut) ou incohérent (en bas)

2.3. Passation du test et analyse des réponses

Protocole et sujets

Le protocole consistait, selon l'instruction donnée aux sujets, à observer les films et, chaque fois que le sujet entendait le son "ba" ou "da", à appuyer immédiatement sur le bouton correspondant, indiqué par le présentateur au début de l'expérience. Les boutons de réponse ont été spécifiés par des marques « ba » et « da ». Pour la moitié des sujets le bouton « ba » était à gauche et le bouton « da » était à droite, pour l'autre moitié les boutons ont été inversés. Les mêmes marques étaient également affichées sur l'écran, de chaque côté de l'écran au niveau des yeux. La durée de l'expérience était environ 25 minutes. Entre les blocs le sujet pouvait faire une pause de durée arbitraire. Les réponses des sujets, avec leur date précise, étaient enregistrées automatiquement au cours de l'expérience, par le logiciel de test.

19 sujets français ont participé à l'expérience avec vision et audition normale ou corrigée, soit 8 femmes et 11 hommes, entre 22 et 51 ans (17 droitiers et 2 gauchers).

Analyse des résultats

Pendant l'expérience les stimuli sont fournis en ligne, et le sujet peut répondre à chaque instant, qu'il y ait ou non la présence d'une cible perceptible « ba » ou « da ». Il peut donc se produire deux types d'erreurs : la présence d'une réponse « ba » ou « da » en l'absence de cible (stimulus « ba » ou « McGurk ») ou l'absence de réponse à une cible. Pour traiter correctement les réponses nous avons mis en place la méthodologie suivante. (1) Pour un

stimulus on compte les réponses qui sont apparues après sa présentation, mais avant le stimulus suivant. (2) Des analyses nous avons été conduites à limiter la validité temporelle de réponse par un seuil, qui est égal à la durée d'une séquence minimale (3500ms). Sur l'histogramme temporel de toutes les réponses (correctes et incorrectes), données par tous les sujets, nous avons pu observer que les plupart des réponses sont à l'intérieur de ce seuil. (3) Pour déterminer les réponses incorrectes, nous distinguons 2 types d'erreurs : « Fausses alarmes » et « Absence de réponse ». Toutes les réponses au-delà du seuil sont considérées comme « fausses alarmes ». S'il n'y a pas de réponse dans l'intervalle entre le stimulus et le seuil, on compte une « Absence de réponse » pour ce stimulus. S'il y a plusieurs réponses dans cet intervalle, on fait une vérification de l'identité des réponses. Si elles sont identiques, nous ne prenons que l'une d'entre elles et la comptons comme une réponse normale, sinon nous les éliminons toutes, et considérons une « absence de réponse » pour le stimulus.

3. RESULTATS

Les résultats bruts sont présentés dans la Table 1. Il apparaît une tendance à obtenir plus d'absence de réponse et moins de réponses multiples proportionnellement avec les stimuli McGurk qu'avec les « ba », sans que le contexte ne joue fortement sur ces tendances (voir les deux colonnes de droite). Si l'on en vient à ce qui est le focus de notre étude, la proportion de réponses « ba » par rapport au nombre total de réponses (« ba » + « da »), on obtient les données de la Figure 4.

Une analyse de la variance à trois facteurs (stimulus, contexte, sujets) sur ces proportions (après transformation en $\text{asin}(\sqrt{x})$, pour assurer la gaussianité) montre que les 3 effets sont fortement significatifs. L'effet significatif du stimulus traduit l'effet McGurk (moins de réponses « ba » pour les stimuli McGurk : $F(1,18)=61.77$, $p<0.0001$). L'effet sujet traduit les fortes différences interindividuelles classiques dans l'effet McGurk ($F(18,18)=3.76$, $p<0.004$). L'effet contexte, traduisant la chute du nombre de réponses « ba » en contexte incohérent ($F(1,18)=35.67$, $p<0.0001$) est essentiellement produit par les stimuli McGurk, ainsi que le montre l'existence d'une interaction entre stimulus et contexte ($F(1,18)=24.14$, $p<0.0001$).

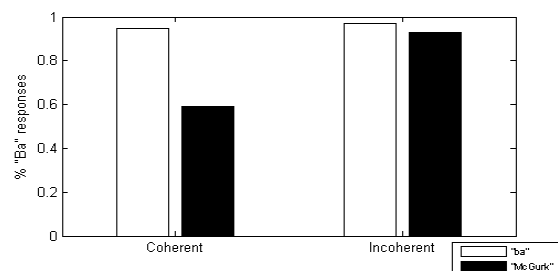


Figure 4 : Pourcentage de réponses « ba » rapportées à l'ensemble des réponses (« ba »/(« ba » + « da »))

Table 1 : Résultats des tests.

Stimuli	Stimuli présentés	Réponse « ba »	Réponse « da »	Absence de réponses	Plusieurs réponses
Cohérent	« ba »	304	12	25	17
	« McGurk »	912	455	301	39
Incohérent	« ba »	304	7	21	16
	« McGurk »	912	724	53	25

Une seconde analyse de la variance, centrée sur les stimuli d'intérêt, les stimuli McGurk, à trois facteurs, sujet, contexte et durée, ne fait pas apparaître d'effet durée global ($F(3,54)=2.07$, $p=0.1156$) mais un effet d'interaction durée-contexte ($F(3,54)=2.85$, $p<0.05$) faiblement significatif. La Figure 5 montre que cet effet est dû au contexte cohérent, pour lequel un allongement du contexte augmente légèrement l'effet McGurk, effet confirmée par une analyse par régression linéaire.

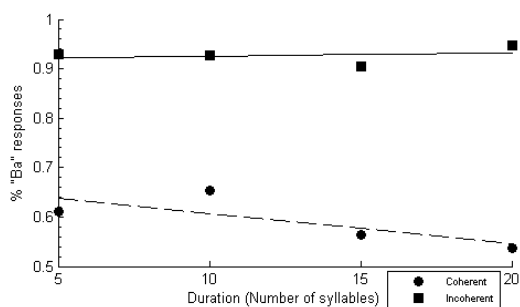


Figure 5 : La régression moyenne des réponses en fonction du contexte et sa durée (en contexte cohérent, $r = -0,31$ significativement différent de zéro, $p < 0.0059$).

4. DISCUSSION ET CONCLUSION

Les résultats obtenus montrent clairement que l'effet McGurk dépend du contexte préalable. La suppression d'effet McGurk signifie que l'on peut bloquer la fusion audio-visuelle. Dans le cas cohérent les flux auditif et visuel se combinent dans un même percept, avec des taux de McGurk classiques (60% de réponses « ba » contre 40% de réponses « da », ce qui correspond aux données d'autres études en français, voir [6]. Dans le cas incohérent l'effet McGurk disparaît complètement et la réponse est essentiellement gérée par l'information auditive, avec des scores supérieurs à 90% de réponses « ba » pour des stimuli McGurk, presque identique à la réponse à un stimulus cohérent « ba ».

Après avoir longtemps considéré l'effet McGurk comme automatique [1], des données récentes ont indiqué qu'il était sous la dépendance de mécanismes attentionnels de divers types [7,8]. Néanmoins, c'est la première fois, à notre connaissance, qu'il est démontré sa sensibilité à des mécanismes vraisemblablement préattentionnels ou de liage conditionnel, qui se réfèrent à des travaux préalables que nous avons menés sur l'analyse de scènes audiovisuelle [9]. Il nous faudra montrer par la suite comment fonctionnent ces mécanismes de liage, à quels mécanismes ils sont reliés à de quels paramètres expérimentaux ils dépendent, ainsi que d'en mettre à jour les corrélats neuronaux (voir par exemple les données

récentes de Bernstein et col. [10] sur le rôle potentiel du Gyrus Supramarginal comme « hub », lieu de fusion de l'information provenant de flux différents, et de contrôle des différences et incohérences entre flux).

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

BIBLIOGRAPHIE

- [1] H. McGurk & J. MacDonald, "Hearing lips and seeing voices," *Nature*, 264, 746-748, 1976.
- [2] J.-L. Schwartz, J. Robert-Ribes, & P. Escudier, "Ten years after Summerfield. a taxonomy of models of audiovisual fusion in speech perception," in *Hearing by Eye*, R. Campbell and et al., Eds. Hove, UK: Psychology Press, 1998, pp. 85-108.
- [3] J.-L. Schwartz, F. Berthommier, & C. Savariaux, "Seeing to hear better: evidence for early audio-visual interaction in speech identification," *Cognition*, 93, B69-B78, 2004.
- [4] F. Berthommier, "A phonetically neutral model of the low-level audio-visual interaction," *Speech Communication*, 44, 31-41, 2004.
- [5] O. Nahorna, "L'émergence des formes audiovisuelles dans le traitement multisensoriel de la parole : expériences et modélisation," Rapport de stage, Master IC2A-AST, Grenoble INP, 2009.
- [6] M.A. Cathiard, J.L. Schwartz, & C. Abry, "Asking a naive question about the McGurk Effect: why does audio [b] give more [d] percepts with visual [g] than with visual [d]?" *Proc. AVSP-2001*, 138-142, 2001.
- [7] K. Tiippana, M. Sams, and K.S. Andersen, "Visual attention modulates audiovisual speech," *Proc. AVSP-2001*, 167-171, 2001.
- [8] A. Alsius, J. Navarra, R. Campbell & S. Soto-Faraco, "Audiovisual Integration of Speech Falters under High Attention Demands," *Current Biology*, 15, 839-843, 2005.
- [9] J. Barker, F. Berthommier, & J.L. Schwartz, "Is primitive coherence an aid to segment the scene?" *Proc. AVSP-08*, Terrigal, Australia, 1036108, 1998.
- [10] L.E. Bernstein, Lu Z.-L., & J. Jiang, "Quantified acoustic-optical speech signal incongruity identifies cortical sites of audiovisual speech processing," *Brain Research*, 1242:172-84, 2008.