Influence de la lecture labiale sur la ségrégation auditive de flux de parole

Aymeric Devergie, Nicolas Grimault, Etienne Gaudrain, Frédéric Berthommier

10ème **C**ongrès **F**rançais d'**A**coustique

Lyon, 12-16 Avril 2010

Influence de la lecture labiale sur la ségrégation auditive de flux de parole

Aymeric Devergie[1], Nicolas Grimault[1], Étienne Gaudrain[2], Frédéric Berthommier[3]

[1] LNSCC, CNRS-UMR 5020, F-69366 Lyon Cedex 07, {adevergi,ngrimault}@olfac.univ-lyon1.fr

[2] MRC Cognition and Brain Sciences Unit, CB2 7EF Cambridge, etienne.gaudrain@mrc-cbu.cam.ac.uk

[3] GIPSA-LAB, CNRS-UMR 5216, F-38402 Grenoble Cedex, Frederic.Berthommier@gipsa-lab.grenoble-inp.fr

En contexte concurrentiel d'écoute, notre perception des locuteurs nécessite des mécanismes de ségrégation auditive. La lecture labiale améliore l'intelligibilité de la parole dans le bruit et des travaux récents suggèrent que certaines interactions audiovisuelles sont pré-phonétiques. Etant donné que certains mécanismes de ségrégation auditive sont réputés être de bas niveau, il semble pertinent de déterminer si une interaction de type pré-phonétique entre ces mécanismes de ségrégation auditive et l'indice visuel langagier existe et contribue à ce bénéfice. Deux études comportementales permettant d'évaluer objectivement l'état de ségrégation auditive et l'influence de mouvements labiaux sur celui-ci ont été menées. Les stimuli utilisés sont des séquences de voyelles dont la fréquence fondamentale alterne entre deux valeurs. Cet écart en fréquence entre les voyelles gouverne la perception de chaque séquence en un ou deux flux auditifs. À ces séquences auditives sont associées des séquences visuelles de mouvements labiaux articulant une voyelle sur deux. Le degré de congruence audiovisuelle est un paramètre du test. Dans la première étude, les participants doivent rappeler l'ordre de présentation de la séquence de voyelles. Dans la seconde étude, les participants doivent détecter une déviation dans le rythme de présentation de la séquence audiovisuelle. Ces tâches sont réputées difficiles lorsque les séquences sont ségrégées en deux flux. Si les indices visuels et les mécanismes de ségrégation interagissent, l'indice visuel devrait ainsi moduler à la baisse les performances. Les résultats de ces études montrent une diminution significative des performances lorsque l'indice visuel est phonétiquement congruent. Ceci indique que 1) le mouvement des lèvres influence les mécanismes de ségrégation auditive, 2) la congruence audiovisuelle phonétique est nécessaire pour observer cette interaction et 3) l'indice visuel et les indices acoustiques influencent de manière indépendante l'état de ségrégation auditive irrépressible.

1 Introduction

Une littérature conséquente s'est intéressée aux mécanismes de ségrégation auditive (Bregman, 1990) impliqués dans les situations d'écoute concurentielle telle que la perception de la parole en présence de plusieurs locuteurs (i.e. situations *Cocktail Party*, Cherry (1953)). Moore et Gockel (2002) suggérent que toute différence acoustique saillante entre plusieurs événements permet aux auditeurs de regrouper et percevoir distinctement les différentes sources en présence. Des mécanismes de ségrégation reposant sur les attributs perceptifs sont réputés étre précoces et irrépressibles. En effet, deux études neurophysiologiques récentes ont montré des réponses reflétant l'état de ségrégation dans des structures cérébrales tel que le cortex auditif primaire (Micheyl *et al.*, 2005) ou le tronc cérébral (Pressnitzer *et al.*, 2008). En situation *Cocktail Party*, les auditeurs tirent également profil de la lecture labiale pour améliorer la compréhension de la parole (Sumby et Pollack, 1954). Le bénéfice audiovisuel était initialement attribué à des interactions audiovisuelles tardives de nature phonétique (Massaro et Cohen, 1983). Cependant, de nombreuses études en neuroimagerie montrent des interactions dans les structures cérébrales primitives (i.e. Calvert *et al.* (2001); Besle *et al.* (2004)). Comme, à la fois les mécanismes de ségrégation auditive et les interactions audiovisuelles ont été observés dans les structures cérébrales primitives, il devient intéressant d'étudier l'influence de la lecture labiale sur les mécanismes de ségrégation irrépressible. Nous proposons deux études comportementales dans lesquelles nous évaluons par des méthodes objectives l'état de ségrégation auditive et l'influence de la lecture labiale sur celle-ci. Nous proposons des séquences de voyelles qui peuvent étre perçue comme un seul flux intégré (méme fréquence fondamentale) ou comme deux flux distincts (fréquences fondamentale différentes). Les téches proposées exigent des participants qu'ils perçoivent les flux comme étant intégrés, ce qui nous permet d'accéder aux mécanismes de ségrégation irrépressible (i.e. mesure du seuil de cohérence, Van Noorden (1975)). Les gestes labiaux visuels prononcent une voyelle sur deux uniquement. Si l'interaction audiovisuelle se produit alors l'indice visuel doit favoriser la ségrégation ainsi avoir un effet délétére sur les tâches.

2 Expérience 1

Dans cette expérience, nous mesurons l'état de ségrégation en demandant aux participants de rappeler l'ordre de présentation de séquences de six voyelles al-

ternant en fréquence fondamentale (Gaudrain *et al.*, 2007). Des performances faibles sont supposées refléter la ségrégation irrépressible des séquences en deux flux distincts. Dès lors que les séquences sont ségrégées, il est alors impossible pour les participants de rappeler l'ordre de la séquence de six voyelles. Simultanément, des mouvements labiaux visuels articulent une voyelle audio sur deux. Différentes conditions de cohérence audiovisuelle sont proposées. Si l'interaction audiovisuelle se produit, nous devons observer une diminution du taux de rappel correct de l'ordre de présentation des séquences puisque les séquences ont été séparée en deux flux.

2.1 Participants

Dix participants, âgés entre 18 et 24 ans, ont pris part à l'expérience. Tous sont de langue maternelle française et ont des pertes auditives inférieures à 15 dB HL pour des sons purs dont les fréquences sont comprises entre 250 et 4000Hz (American National Standards Institute, 2004). Les participants ont reçu une indemnisation forfaitaire et ont signé un consentement de participation. Cette étude a été approuvée par un comité d'éthique local (CPP Sud-Est II No. 06035).

2.2 Stimuli

Des séquences de six voyelles /a e i o y u/ alternées en fréquence fondamentale (F0) ont été générées. La fréquence du flux grave (F0(1)) est égale à 100Hz. La fréquence du flux aigu peut prendre une valeur entre 100 et 238Hz. Chaque voyelle dure 166ms incluant une rampe *raised-cosine* de 10ms en onset et offset. Ces voyelles sont générées avec l'algorithme de Klatt (Klatt, 1980). Les formants des voyelles sont identiques à ceux utilisés dans Gaudrain *et al.* (2008). Simultanément, des séquences de mouvements labiaux prononçant une voyelle audio sur deux sont présentées. Les mouvements labiaux sont des mouvements générés à partir de captures video (i.e. Berthommier (2003)). Les séquences audio et vidéo sont générées séparement et mixées dans un second temps. Trois types de mouvements labiaux ont été proposé. La condition 'V' correspond au mouvement labial articulant la voyelle audio correspondante (contenu phonétique). La condition 'O' correspond à un mouvement labial neutre d'ouverture-fermeture. Cette condition ne fournit pas d'indice phonétique mais uniquement un indice de rythme. La condition 'C' correspond à la condition contrôle. Les lèvres restent immobiles durant toute la séquence. Toutes les séquences ont été générées avec une routine matlab avant le début de l'expérience. Les séquences sont ajustées en valeurs RMS à 85 dB SPL (Larson Davis AEC101 et 824, American National Standards Institute (1995)).

2.3 Méthode

Les participants commencent par une tâche d'identification des voyelles Klatt présentées seules. Les voyelles /a e i o y u/ sont présentées dans un ordre aléatoire à différentes fréquences (100, 147 ou 238Hz). Tous les participants ont obtenu un score d'identification de 100% (sauf un avec seulement 87.5%). Ensuite, ils commencent la tâche de rappel d'ordre. Chaque séquence audiovisuelle est répétée en boucle pendant dix secondes. Deux secondes après le début de la séquence, les participants peuvent donner l'ordre des voyelles. La séquence suivante est jouée si ils valident leur réponse ou bien si la séquence se termine. Les dix premiers essais constituent la phase d'entraînement. Toutes les combinaisons de conditions visuelles (CV) et de F0 (100 à 238Hz) sont présentées. Les essais suivant constituent la phase de test. Quatre blocs de deux essais sont proposés. Dans chaque essai, chaque combinaison de CV et F0 est répétée cinq fois et présentée aléatoirement. L'expérience dure six heures et est divisée en sessions de deux heures.

2.4 Résultats

La figure 1 représente le pourcentage de rappel correct de l'ordre de présentation des six voyelles en fonction de F0 et de la CV moyenné sur vingt participants. Une ANOVA mesures-répétées avec comme facteurs F0 et CV a été réalisée. Une correction de Bonferroni planifiée a été appliquée aux données. La fréquence fondamentale a un effet significatif sur le taux de rappel correct [$F(9,81)= 22.14$, $p<0.01$]. La condition visuelle a un effet sensible sur les performances [$F(2,18)=3.64$, $p=0.05$] démontrant que les taux de rappel correct diminue lorsque les lèvres prononcent les voyelles audio correspondantes. Aucune interaction entre l'effet de la condition visuelle et la fréquence fondamentale n'a été observée [$F(8,182)=0.89$, $p=0.58$].

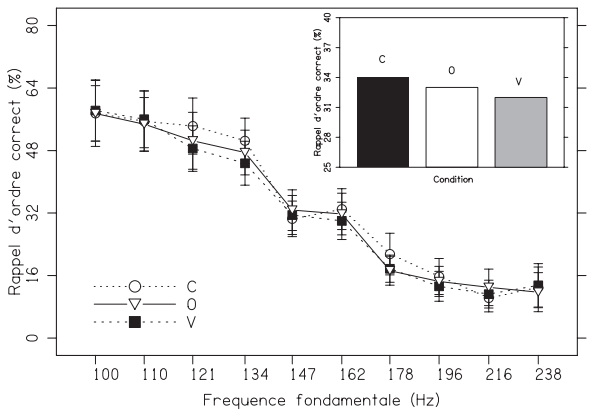

FIG. 1: Pourcentage de rappel correct de l'ordre de présentation des 6 voyelles en fonction de la condition visuelle et de la fréquence fondamentale

2.5 Discussion

Les données suggèrent qu'un mouvement de lèvres congruent phonétiquement peut influencer la ségrégation auditive irrépressible. En revanche, quand le mouvement de lèvres est un mouvement d'ouverture-fermeture, la ségrégation irrépressible n'est pas influencée. Dans cette condition visuelle, l'interaction audiovisuelle est probablement moins forte que dans le cas de lèvres congruente phonétiquement. En effet, Devergie

et al. (2009) montrent que l'appariement entre les indices visuels et audio est plus fort lorsque la cohérence est phonétique. Ici, la cohérence audiovisuelle semble déterminante pour qu'un indice visuel de parole vienne influencer la ségrégation auditive. Les données semblent indiquer également que l'indice visuel augmente la ségrégation irrépressible méme en l'absence de différence de fréquence fondamentale. L'absence d'interaction entre l'effet de l'indice visuel et l'effet de la fréquence fondamentale repose sur une faible significativité statistique. A l'issue de cette étude, nous nous pouvons rien conclure concernant les interactions entre ces deux facteurs.

3 Expérience 2

Dans cette étude, nous mesurons l'habilité des participants à détecter une variation du rythme de présentation des séquences de voyelles alternant en F0 (similaires à celle de l'expérience 1). La tâche repose sur la détection d'une variation du silence séparant deux voyelles successives. Comme dans la première expérience, une voyelle sur deux est prononcée visuellement. Tout d'abord, nous suggérons qu'il est difficile de juger de la durée du silence séparant deux événements qui sont perçus dans des flux auditifs différents (Cusack et Roberts (2000); Stainsby *et al.* (2004)). Des faibles performances de détection indiqueraient que la ségrégation irrépressible s'est mise en place. Ensuite, si l'indice visuel affecte les performances de détection, nous pouvons penser qu'une interaction de bas niveau entre les modalités visuelle et auditive s'est produite.

3.1 Participants

Dix participants, âgés entre 18 et 24 ans, ont pris part à l'expérience. Les critères de sélection vis-à-vis de leurs pertes auditives étaient les mêmes que pour l'expérience 1.

3.2 Stimuli

Six voyelles /a e i o y u/ prononcées par un locuteur masculin ont été enregistrées au format wave (44.1Khz, 16 bits). Les fréquences fondamentales (F0) des voyelles sont corrigées avec STRAIGHT. La fréquence F0(1) est égale à 100Hz. La fréquence F0(2) peut être soit égale à 100Hz ou bien à 224Hz. Chaque voyelle dure 166ms incluant une rampe *raised-cosine* de 10ms en onset et offset. Les séquences sont ajustées en valeurs RMS à 70 dB SPL (Larson Davis AEC101 et 824, American National Standards Institute (1995)). Les voyelles sont présentées dans des séquences de vignts paires alternant en F0. Les voyelles sont choisies aléatoirement pour chaque séquence. Comme dans l'expérience 1, des mouvement de lèvres sont proposés. Les mouvements de lévres sont enregistrés en même temps que les voyelles audio. Deux types de mouvement de lèvres sont proposés. La première condition 'V' prononce la voyelle audio correspondante. Pour la seconde condition 'C', les lèvres restent fermées durant toute la séquence. Chaque séquence est générée, pendant l'expérience, en fonction de la réponse des participants.

3.3 Méthode

La méthode utilisée est similaire à celle utilisée par Roberts *et al.* (2002). Une procédure adaptative à choix forcé *3 down-1 up* (Levitt, 1971) a été utilisée pour mesurer la plus petite variation de rythme détectable dans une séquence de voyelles. Les séquences de voyelles étaient présentées par paire. L'une des séquences (séquence *Contrôle*) avait un rythme de présentation qui restait constant. Chaque voyelle était séparée de la suivante par un silence de 40ms. L'autre séquence (séquence *Test*) avait un rythme de présentation qui devenait irrégulier au fur et à mesure (voir détails Figure 2). À chaque essai, le sujet devait identifier la séquence présentant une irrégularité rythmique. Le délai temporel variait géométriquement au cours de la procédure adaptative par pas de $2^{1/2}$ (pas réduit à $2^{1/4}$ après deux changements de direction). La valeur initiale du décalage temporel était fixée à 20 ms. La valeur maximale du décalage était de 40ms afin d'éviter toute superposition entre voyelles successives. Deux secondes de silence séparaient les deux intervalles à comparer. La mesure du seuil de détection prenait fin aprés six changements de direction. Le seuil de détection était défini comme étant la moyenne géométrique des délais temporels lors des quatre derniers changements de direction. Un essai était abandonné si le participant donnait dix mauvaises réponses successives. Les participants étaient rejetés si le taux de ce type d'essais était supérieur à 50%. Deux sujets ont été rejeté sur ce critére (75% et 62%). Quatre mesures ont été faites pour chaque combinaison de F0 et de CV. La moyenne géométrique de ces quatre mesures était prise comme valeur finale de la variation.

3.4 Résultats

La figure 3 représente le seuil de détection moyen en fonction de la différence de fréquence et de la condition visuelle. Une ANOVA mesures-répétées avec comme facteurs la fréquence fondamentale et la condition visuelle a été calculée. La fréquence fondamentale a un effet significatif sur les performances de détection [$F(1,9)=25.32$, $p<0.01$]. La condition visuelle a, elle aussi, un effet significatif sur les performances [$F(1,9)=8.19$, $p=0.024$]. Nous n'observons pas d'interaction entre les facteurs [$F(1,9)=0.21$, $p=0.66$].

3.5 Discussion

Un mouvement de lèvres congruent augmente la ségrégation auditive irrépressible. Ceci confirme l'effet sensible rapporté dans l'expérience 1. L'utilisation de voyelles naturelle a probablement renforcé la cohérence audiovisuelle et contribué à l'interaction entre le mouvement labial et la ségrégation auditive. De plus, la tâche de détection de déviation était problabement plus facile à réaliser que la tâche de rappel d'ordre. La charge cognitive impliquée dans l'expérience 2 est potentiellement moins importante que dans l'expérience 1. Enfin, l'absence d'interaction entre les facteurs laisse penser que les mécanismes de ségrégation basés sur la fréquence fondamentale et les mécanismes d'intégration audiovisuelle participent indépendamment à la ségrégation auditive.

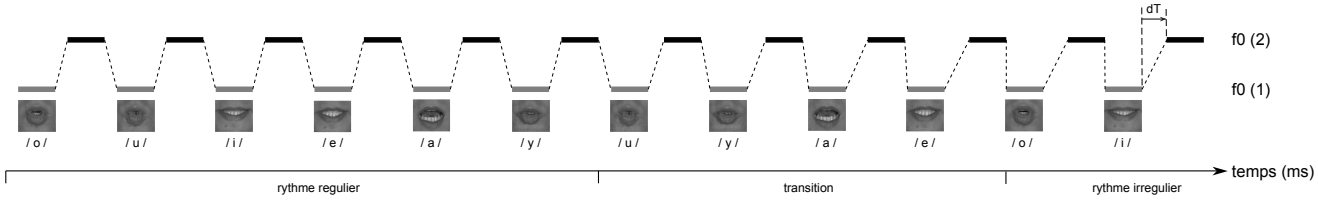

FIG. 2: Représentation d'une séquence *Test*. Dans une premiére phase, le rythme de présentation est régulier. Dans une seconde phase le rythme devient irrégulier progressivement en décalant un des deux flux de voyelles. Dans la derniére phase, le rythme est maintenu irrégulier

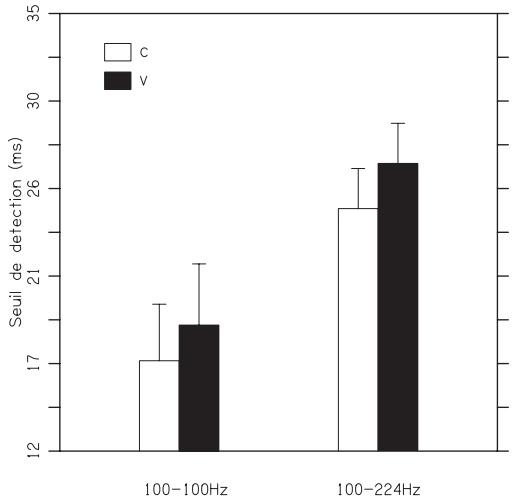

FIG. 3: Seuil de détection moyen de la déviation de rythme pour chaque différence de fréquence et chaque condition visuelle pour les dix participants

4 Discussion générale

Effet de la lecture labiale sur la ségrégation irrépressible Le bénéfice apporté par la lecture labiale en situation concurrentielle d'écoute repose en partie sur une interaction audiovisuelle précoce. La ségrégation auditive irrépressible peut être influencée par des mouvements de lèvres phonétiquement congruent avec le signal audio de parole. Cependant, cette interaction de bas niveau n'exclue pas la contribution d'interactions audiovisuelles de plus haut niveau. Arnal *et al.* (2009) ont par ailleurs montré que les deux niveaux d'interactions étaient présent au cours de la perception de la parole audiovisuelle.

Cohérence audiovisuelle La cohérence audiovisuelle semble être un facteur déterminant. En effet, dans Rahne *et al.* (2007) des séquences de sons purs sont associés à des formes géométriques. Cette cohérence artificielle explique probablement pourquoi les auteurs ne parviennent pas à mettre en évidence une interaction audiovisuelle lorsque la ségrégation repose sur un indice acoustique saillant comme la fréquence fondamentale. La cohérence est également artificielle dans notre tâche de rappel d'ordre. Les signaux audio et vidéo ont été générés séparément. De plus, les mouvements labiaux ont été généré à partir d'un nombre limité d'images capturées de lévres. Certaines voyelles étaient donc hypo-articulées. Cela a pu engendré de la confusion entre des voyelles proches comme /o/, /u/ et /y/ par exemple. En revanche, les signaux utilisés dans notre tâche de détection étaient plus naturels. Les signaux audio et vidéo étaient enregistrés simultanément à partir d'un même locuteur. L'articulation des voyelles était mieux contrôlée, les rendant plus discriminable les une des autres.

Remerciements

Ces travaux ont été financés par une bourse du Cluster HVN 2007 de la Région Rhônes-Alpes Auvergne et par l'Agence Nationale de Recherche (ANR-08-BLAN-0167-01).

Références

American National Standards Institute (**1995**). "Ansi s3.7-r2003 : Methods for coupler calibration of earphones", .

American National Standards Institute (**2004**). "Ansi s3.21-2004 : Methods for manual pure-tone threshold audiometry", .

Arnal, L. H., Morillon, B., Kell, C. A., et Giraud, A.-L. (**2009**). "Dual neural routing of visual facilitation in speech processing.", J Neurosci **29**, 13445–13453.

Berthommier, F. (**2003**). "A phonetically neutral model of the low-level audiovisual interaction", *in Proceedings of the International Conference Audio Visual Speech Processing.*

Besle, J., Fort, A., Delpuech, C., et Giard, M.-H. (**2004**). "Bimodal speech : early suppressive visual effects in human auditory cortex.", Eur J Neurosci **20**, 2225–2234.

Bregman, A. (**1990**). *Auditory Scene Analysis : The Perceptual Organzation of Sounds* (MIT Press, Massachussets, USA).

Calvert, G. A., Hansen, P. C., Iversen, S. D., et Brammer, M. J. (**2001**). "Detection of audio-visual integration sites in humans by application of electrophysiological criteria to the bold effect.", Neuroimage **14**, 427–438.

Cherry, E. C. (**1953**). “Some experiments on the recognition of speech, with one and with two ears”, J Acoust Soc Am **25**, 975–979.

Cusack et Roberts (**2000**). “Effects of differences in timbre on sequential grouping”, Perception and Psychophysics **62**, 1112–1120.

Devergie, A., Berthommier, F., et Grimault, N. (**2009**). “Pairing audio speech and various visual displays : binding or not binding ?”, *in Proceedings of the International Conference on Audio-Visual Speech Processing*, 140–144.

Gaudrain, E., Grimault, N., Healy, E. W., et Béra, J.-C. (**2007**). “Effect of spectral smearing on the perceptual segregation of vowel sequences.”, Hear Res **231**, 32–41.

Gaudrain, E., Grimault, N., Healy, E. W., et Béra, J.-C. (**2008**). “Pitch-based streaming of vowel sequences, speech-in-speech segregation, and frequency selectivity.”, J Acoust Soc Am **123**, 3301.

Klatt, D. (**1980**). “Software for a cascade/parallel formant synthesizer”, J Acoust Soc Am **67**, 971–995.

Levitt, H. (**1971**). “Transformed up-down methods in psychoacoustics.”, J Acoust Soc Am **49**, Suppl 2 :467+.

Massaro, D. W. et Cohen, M. M. (**1983**). “Evaluation and integration of visual and auditory information in speech perception.”, J Exp Psychol Hum Percept Perform **9**, 753–771.

Micheyl, C., Tian, B., Carlyon, R. P., et Rauschecker, J. P. (**2005**). “Perceptual organization of tone sequences in the auditory cortex of awake macaques.”, Neuron **48**, 139–148.

Moore, B. C. J. et Gockel, H. (**2002**). “Factors influencing sequential stream segregation”, Acta Acustica **88**, 320–333.

Pressnitzer, D., Sayles, M., Micheyl, C., et Winter, I. M. (**2008**). “Perceptual organization of sound begins in the auditory periphery.”, Curr Biol **18**, 1124–1128.

Rahne, T., Böckmann, M., von Specht, H., et Sussman, E. S. (**2007**). “Visual cues can modulate integration and segregation of objects in auditory scene analysis.”, Brain Res **1144**, 127–135.

Roberts, B., Glasberg, B. R., et Moore, B. C. J. (**2002**). “Primitive stream segregation of tone sequences without differences in fundamental frequency or passband.”, J Acoust Soc Am **112**, 2074–2085.

Stainsby, T. H., Moore, B. C., et Glasberg, B. R. (**2004**). “Auditory streaming based on temporal structure in hearing-impaired listeners”, Hearing Research **192**, 119–130.

Sumby, W. H. et Pollack, I. (**1954**). “Visual contribution to speech intelligibility in noise”, Journal of th Acoustical Society of America **26**, 212–215.

Van Noorden, L. (**1975**). “Temporal coherence in the perception of tone sequences”, Unpublished doctoral dissertation, Technische Hogeschool Eindhovern, Eindhoven, The Netherlands.