Naissance de la représentation d'une consonne entre les voyelles : Les conditions d'une intégration audiovisuelle

Marie-Agnès Cathiard, Séverine Gedzelman, Christian Abry & Hélène Loevenbruck

Institut de la Communication Parlée, INPG/Université Stendhal,
Domaine Universitaire, BP 25, 38040 Grenoble Cedex 9, France
Tél.: ++33 (0)4 76 82 41 28 - Fax: ++33 (0)4 76 82 43 35
Mél: cathiard@icp.inpg.fr

ABSTRACT

We tested auditory and visual perception of the vowel-to-vowel [yi] gesture via the production of a [ɥ] epenthetic glide in-between. This epenthetic glide can gain in the course of linguistic change the status of a true represented segment, like *v* in French *pouvoir*, hence English *power* (from Old French *t* deletion of Latin *potere*): what we dubbed the *power-effect* (Cathiard et al. [2]). In experiment 1, we showed that the retraction movement in the off-gliding phase of the [y] vowel is misleading for the anticipation of the following [i] vowel, since [i] identification always comes after the minimal [ɥ] constriction event. Moreover the identification boundary in the audio condition is systematically ahead (by at least 20 ms) of the visual condition. In experiment 2, we tested if the epenthetic glide could give birth to a consonant, using an audiovisual McGurk paradigm. We evidenced that this glide needs to be sufficiently lengthened (i.e. maintained in a static phase) in order to be integrated with the sound and represented as a true consonant.

1. INTRODUCTION

Nous nous intéresserons dans cette étude aux glides, non pas comme véritables segments phonologiques entrant dans la syllabation, mais comme existant à l'état latent à l'intérieur de transitions de voyelle à voyelle. Le modèle articulatoire 2-COMP de Abry et al. [1] propose que ces glides soient issus d'une asynchronie entre les différentes composantes de contrôle des voyelles en contact. Ces glides non contrôlés, sous-produits d'une transition intervocalique, peuvent parfois gagner le statut d'un véritable segment représenté, comme on le voit en linguistique diachronique. Ainsi en est-il du *v* dans *pouvoir* : le mot latin potere est devenu en vieux français, par disparition du [t] entre voyelles, *poeir*, lequel, présentant un glide entre les voyelles *o* et *e*, a ainsi donné *v* en français moderne ; *poeir,* emprunté au français, a donné en moyen anglais *poër,* puis *power*. Nous avons appelé l'émergence de ce segment épenthétique, *l'effet power* (Cathiard et al. [2]). Sur quelles bases perceptives repose ce processus évolutif ?

2. EXPERIENCE 1: L'EFFET DU GLIDE EPENTHETIQUE SUR L'ANTICIPATION VOCALIQUE

Dans cette étude portant sur des transitions [yi], nous testerons spécifiquement l'identification de la voyelle [i] en nous demandant si nos sujets pourront l'anticiper rapidement, soit dès l'abandon de la position climax du [y], ou s'ils suivront, dans leur perception auditive et visuelle, les détails de la production au cours du temps : autrement dit, nous testerons si la présence du glide épenthétique peut être trompeuse (*misleading*) en empêchant l'identification du prochain segment [i].

Nous avons enregistré audiovisuellement, à l'aide du poste Visage-parole de l'ICP (Lallouache [3]), un locuteur français masculin entraîné qui prononçait, dans un ordre aléatoire, 10 répétitions de chacune des 3 séquences suivantes : "Tu dis RUHI ise ?" [tydiR**yi**iz], "Tu dis UHI ise ?" [tydi**yi**iz] et "Tu dis ZUHI ise ?" [tydiz**yi**iz] ("RUHI", "UHI" et "ZUHI" sont des pseudo noms propres et "ise" un pseudo verbe à la troisième personne). Ces séquences nous permettent d'observer des transitions de la voyelle [y] vers la voyelle [i] sans consonne intervocalique. L'analyse articulatoire met en évidence, en suivant l'évolution temporelle de l'aire aux lèvres, la présence typique d'un glide [ɥ] entre les voyelles [y] et [i], qui se manifeste par une constriction d'aire minimale. On peut ainsi observer, sur la figure 1, pour une séquence "Tu dis RUHI ise ?", que le plateau de constriction du [y] dans "RUHI" commence alors que l'on est dans la production coarticulée du [R] arrondi initial, avec une aire aux lèvres autour de 90 mm^2, i.e. suffisamment petite pour contribuer aux caractéristiques acoustiques du [y]. A la suite de ce plateau de constriction, l'aire aux lèvres ne réaugmente pas aussitôt après pour la voyelle [i] suivante, mais continue à décroître pour atteindre une valeur minimale de constriction aussi petite que 0,5 mm^2 (sans changer le régime acoustique vocalique) : cet événement de constriction minimale correspond au glide [ɥ]. Ce n'est qu'après cet événement que l'aire réaugmente rapidement pour [i]. Ces transitions vocaliques, avec glide épenthétique, peuvent être transcrites plus précisément [yɥi].

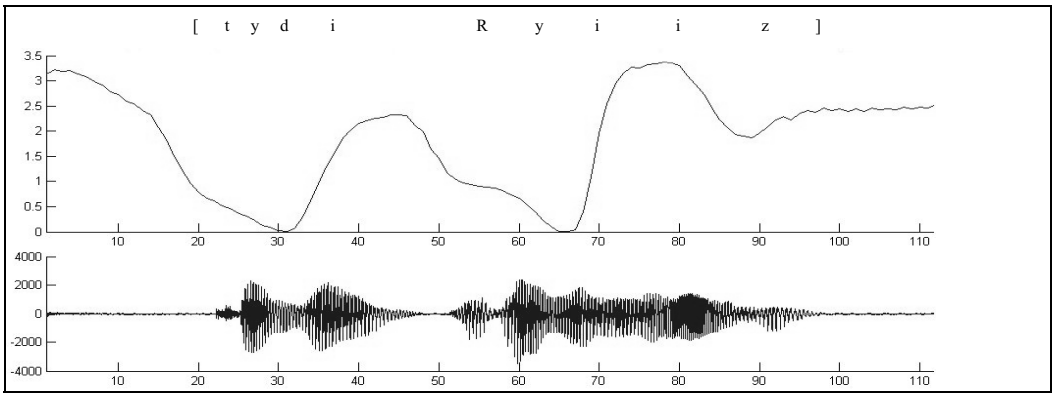

Figure 1 : Signal acoustique (en bas ; en abscisse, numéros de trames : 200 ms dans chaque dizaine) et évolution de l'aire aux lèvres (en haut ; en cm2) pour la séquence «Tu dis RUHI ise?». N.B. : l'aire aux lèvres décroît jusqu'à 1 mm2 à la transition [ydi] et 0,5 mm2 à la transition [Ry(ɥ)i] (et non 0 mm2 comme peut le laisser croire l'échelle en cm2).

Pour les 30 transitions, nous avons observé systématiquement un minimum d'aire de constriction avec une valeur moyenne de 8 mm^2. Ce minimum est clairement inférieur à la valeur de la partie stable du [y], soit 50 mm^2 en moyenne (de 26,6 mm^2 en moyenne pour les 10 réalisations de UHI, à 44,5 mm^2 pour ZUHI, et 81,4 mm^2 pour RUHI). Cette valeur minimale de constriction du glide ne peut être prédite par la valeur du plateau de constriction.

L'analyse acoustique de ces transitions [yɥi] montrent une baisse d'intensité de 2 dB autour de l'aire minimale et une baisse des formants de 122,5 Hz pour F2 et de 176,5 Hz pour F3, par rapport aux valeurs cibles du [y] (cette baisse formantique n'empêche pas la focalisation des formants F2 et F3 typique de la voyelle [y]). Ces observations sont similaires à celles de Chafcouloff [4].

Nous avons exploré par gating auditif et visuel (Grosjean [5]) le rôle perceptif de l'événement de constriction minimale. Six séquences ont été choisies parmi les 30 étudiées, deux de chaque contexte. Les séquences tronquées commençaient en début de phrase et s'arrêtaient autour de la constriction minimale — à plus ou moins 60 ms avant et après cet événement — par pas de 20 ms (aboutissant à 7 pas de gating). Dans la condition audio, 20 sujets français entendaient les séquences tronquées jusqu'aux points de gating. Dans la condition visuelle, ils entendaient et voyaient le début de la phrase porteuse ("Tu dis ...") et voyaient seulement la suite. Dans les deux conditions, la tâche du sujet était de décider si la séquence "Tu dis ..." se terminait par [y] ou [i] (par exemple par "HUE" [y] ou "UHI" [yi]).

Les courbes d'identification auditive et visuelle obtenues pour chacun des 6 stimuli montrent des bascules perceptives rapides (la figure 2 donne l'exemple d'une séquence Ruhi). Trois séquences montrent une bascule perceptive de [y] vers [i] en 20 ms (deux en 40 ms et une en 60 ms). Pour chaque stimulus, la frontière d'identification (à 50%) dans la condition auditive est systématiquement en avance par rapport à celle de la condition visuelle.

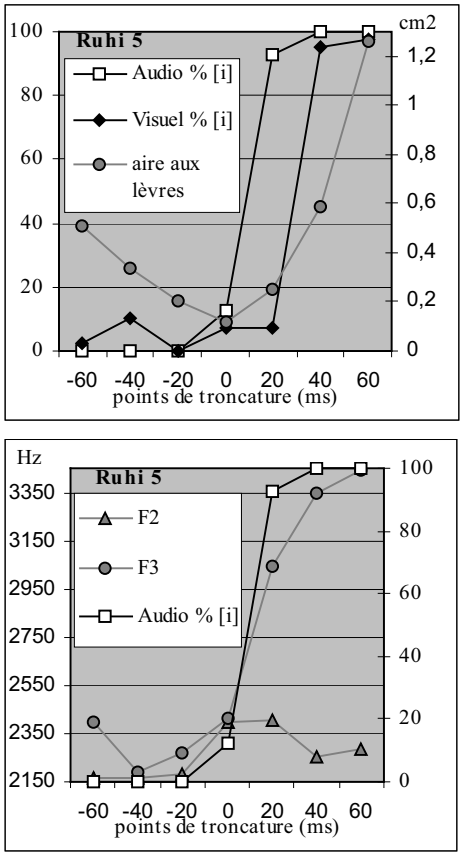

Figure 2 : Exemple d'une séquence "RUHI". L'axe horizontal indique la date de troncature (en ms) par rapport à la constriction minimale (date 0). En haut: évolution de l'aire aux lèvres (cm2) avec courbes d'identification auditive et visuelle en %[i] (la frontière à 50% en audio est de 19 ms plus précoce que la frontière visuelle). En bas : évolution des formants F2 et F3 (Hz) et courbe d'identification auditive en % [i].

Les analyses Probit effectuées sur les courbes audio et visuelle de chaque stimulus indiquent une avance significative ($p<0.05$) en audio de 19 à 29 ms selon le stimulus. Notons que les études menées jusque-là sur l'anticipation visuelle (Cathiard et al. [6]) montraient une avance de la vision sur l'audition.

L'explication que nous proposons de ce décalage systématique est l'explication causale articulatori-acoustique suivante, en prenant l'exemple du stimulus "Ruhi 5" (figure 2). La courbe d'identification auditive suit le formant F3 et le moment de la bascule auditive correspond à l'augmentation d'aire, petite (moins de 20 mm2) mais détectable auditivement, puisqu'elle fait grimper F3 d'une manière significative. Pourquoi ? Partant de [y], le modèle d'affiliation de F3 à la cavité avant ([Maeda & Carré [7]) peut être approximé par un tube étroit de 9 cm de longueur (correspondant à la constriction linguale de 5 cm plus la constriction labiale de 4 cm) avec une résonance demi-onde ($\lambda/2$), soit en-dessous de 2kHz. Au moment de l'aire aux lèvres minimale, soit après la rétraction de la protrusion, on approxime une résonance quart d'onde ($\lambda/4$) avec une longueur de 4,5 cm (ce qui correspond à la constriction linguale pour [i] moins 0,5 cm pour la constriction des lèvres), restant ainsi toujours en-dessous de 2 kHz. Ce changement dans le mode des résonances de $\lambda/2$ [y] à $\lambda/4$ [ɥ] correspond aux petits changements dans la focalisation F2-F3, mesurés plus haut. On peut penser qu'aussitôt qu'est dépassée la constriction minimale, la moindre réouverture des lèvres peut laisser entendre la constriction linguale caractéristique du [i], passant ensuite rapidement de 2 kHz à plus de 3 kHz (avec un tube de 5 cm de longueur et une résonance demi-onde). A l'inverse, cette légère augmentation d'aire n'est pas encore suffisante en perception visuelle pour permettre d'identifier la forme labiale du [i] alors qu'elle permet en acoustique un changement de régime. On a par conséquent une perception visuelle en retard, puisqu'il faut une augmentation d'aire aux lèvres plus importante (d'environ 40 mm^2) pour que la configuration du [i] soit reconnue.

Mais la frontière d'identification, qu'elle soit auditive ou visuelle, se situe toujours après l'événement de constriction minimale. Ainsi : (i) l'identification acoustique attend l'augmentation de F3 ; (ii) et, de même, l'identification visuelle suit le décours temporel de l'aire aux lèvres. Il est certain qu'au niveau acoustique, les petits changements dans la focalisation, après la partie stable du [y], ne peuvent pas être un indice pour l'identification de la voyelle suivante [i] : on reste sur une focalisation de type [y]. Et on ne peut pas dire non plus que la diminution d'aire aux lèvres après le plateau de constriction du [y] permette d'anticiper la voyelle suivante [i] (en dépit de la fin de la rétraction des lèvres que l'on trouve dans [ɥ]). Les sujets doivent donc attendre l'augmentation formantique puis celle de l'aire aux lèvres.

3. EXPERIENCE II : L'INTEGRATION AUDIOVISUELLE DU GLIDE COMME CONSONNE

La seconde question que nous nous sommes posée est de savoir si ce glide peut ou ne peut pas donner naissance à une véritable consonne. On peut remarquer (figure 1) qu'une même constriction minimale, due au geste de voyelle à voyelle à travers la consonne, est observée pour la transition CVC [ydi] de "Tu dis ..." (le [d] pouvant être considéré comme neutre pour les lèvres). Nous nous sommes demandé sous quelles conditions, notre glide — qui se manifeste par une très petite aire intéro-labiale dans nos données, toute proche de l'occlusion bilabiale de [b] — pourrait être intégré perceptivement comme un [b].

Nous avons eu recours au paradigme dit de McGurk (McGurk & Mac Donald [8]) qui consiste à monter des séquences auditive et visuelle non congruentes. En prenant exemple sur un McGurk de combinaison avec [aba] visuel monté sur un audio [ada], on obtient classiquement le percept [abda] (et non [adba]). Obtiendra-t-on de la même façon une intégration audiovisuelle [ybdi], en montant sur un audio [ydi] un visuel [yɥi] (avec [ɥ] proche d'une aire nulle) ?

Après enregistrement audiovisuel de séquences [ydi], [ybi], [yi], [yɥi] et [ybdi] (insérées dans une phrase porteuse du type : « t'as dit : UDI ise ? »), nous avons constitué un continuum naturel visuel avec 5 stimuli, allant de [ydi] à [ybi]. Après (1) [ydi], nous avons (2) [yi] avec un glide présent sur le signal articulatoire mais peu marqué (aire minimale de 17 mm2), puis (3) [yɥi] avec un glide produit volontairement qui se traduit par une petite constriction de 6 mm^2, (4) [ybi] "raccourci" dont le temps d'occlusion bilabiale du [b] a été manipulé, passant de deux images (80 ms) à une image (40 ms) et enfin (5) [ybi]. Ces 5 stimuli visuels ont tous été montés sur une même séquence audio [ydi]. Pour le calage de ces séquences visuelles sur l'audio, nous avons pris comme référence la séquence d'aire aux lèvres [ydi] dans laquelle une aire minimale existe et permet de repérer sur l'audio correspondant le point de synchronisation.

Nous avons présenté à 15 sujets francophones : (1) les 5 séquences *audiovisuelles* répétées 10 fois (en ordre aléatoire), (2) les 5 séquences *visuelles* répétées 5 fois et (3) les 2 séquences *auditives* [ybdi] et [ydi] répétées 10 fois. La tâche consistait, dans chacun des 3 tests, à identifier "udi" ou "ubdi".

Le test en modalité auditive seule a conduit à 100% de réponses correctes pour [ybdi] et [ydi]. Les courbes d'identification [ybdi] sont données pour les deux conditions audiovisuelle et visuelle seule dans la figure 3. On peut remarquer que notre continuum naturel est bien valide puisque nous obtenons une augmentation régulière des réponses "ubdi" de [ydi] à [ybi].

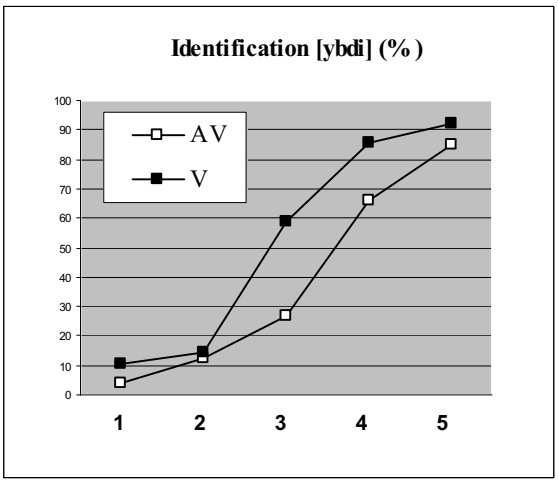

Figure 3 : Pourcentages d'identification [ybdi] pour 5 stimuli visuels (1 = [ydi], 2 = [yi], 3 = [yɥi], 4 = [ybi] raccourci et 5 = [ybi]) présentés en condition visuelle seule (V) et en condition audio-visuelle (AV ; dans ce cas, montés sur un audio [ydi]).

L'analyse de variance à 2 facteurs (condition et stimulus) indique un effet stimulus significatif ($F(4, 56)= 91.5$, $p<.01$), un effet condition seulement significatif à $p=0.014$ ($F(1,14)=7.9$) et un effet d'interaction significatif ($F(4,56)=4.36$, $p<.01$). Les comparaisons *posthoc* indiquent une différence significative entre les scores audiovisuel et visuel uniquement pour le stimulus [yɥi]. Ce stimulus [yɥi] obtient 58% d'identification [ybdi] en perception visuelle seule : le glide a pu ainsi être vu dans plus de 50% des cas comme une consonne. En condition audiovisuelle, nous constatons que ce glide n'est plus intégré au flux consonantique (seulement 25% de réponses [ybdi]). Ce ne sont que le stimulus [ybi] raccourci et le [ybi] naturel qui permettent un percept de combinaison. Nous pouvons en conclure que, pour qu'un mouvement transitionnel de type glide entre deux voyelles donne naissance à une consonne, il faut qu'il soit suffisamment allongé, autrement dit maintenu dans une phase statique, comme l'est la séquence [ybi] même avec une occlusion raccourcie.

4. DISCUSSION

Dans cette étude de la production et de la perception du glide épenthétique d'arrondissement en français, nous avons montré, par notre expérience I, que la présence du glide épenthétique [ɥ] entre les voyelles [yi] semble bien jouer le rôle de verrou perceptif pour passer d'une voyelle à l'autre. Auditivement il faut attendre une remontée formantique importante (au-dessus de 2600-2900 Hz) pour que les sujets commencent à identifier la voyelle [i] : ils suivent en fait l'évolution de la résonance de la cavité avant. Visuellement les sujets attendent aussi d'avoir une forme labiale suffisamment claire de [i]. Notre expérience II montre que ce glide peut être utilisé pour faire naître une consonne. Nous avons déterminé qu'un des facteurs *visuels* pour son émergence semble être la *tenue* de ce glide. Il nous reste encore à explorer les conditions *acoustiques* propres à favoriser la perception d'une consonne.

Remerciements : à Christophe Savariaux et Alain Arnal pour leur support technique; à Caroline Brunière et Suzie Bianciotto pour leur aide à la passation de l'expérience I et enfin à tous nos sujets d'expérience. Cette recherche a été financée par un programme cognitique Act1b du Ministère de la Recherche.

BIBLIOGRAPHIE

[1] C. Abry, R. Laboissière, H. Loevenbruck, M.-A. Cathiard and Schwartz, J.-L. Glide production and control in the two-component vowel model. In *Proc. of the 5th Seminar of Speech Production : Models and Data & CREST Workshop of Models of Speech Production : Motor Planning and Articulatory Modelling*, pages 37-40, Kloster Seeon, Bavaria, May 1-4 2000.

[2] M.-A. Cathiard, C. Abry, S. Gedzelman and H. Loevenbruck. Visual and auditory perception of epenthetic glides. In *Proceedings of the AudioVisual Speech Processing'03*, pages 61-66, St-Jorioz, France, 4-7 sept. 2003.

[3] M.-T. Lallouache. *Un poste Visage-Parole couleur. Acquisition et traitement automatique des contours des lèvres*. PhD Thesis. I.N.P. Grenoble, 1991.

[4] M. Chafcouloff. Les caractéristiques acoustiques de [j, ɥ ,w, l, r] en français. *Travaux de l'Institut de Phonétique d'Aix*, volume 7, pages 7-53, 1980.

[5] F. Grosjean. Spoken word recognition processes and the gating paradigm. *Perception & Psychophysics*, 28: 267-283, 1980.

[6] M.-A. Cathiard, C. Abry and M.-T. Lallouache. Does movement on the lips mean movement in the mind? In D. Stork & M. Hennecke (Eds.), *Speechreading by Humans and Machines*, 211-219, Berlin, Springer-Verlag, 1996.

[7]S. Maeda and R. Carré. Modèle de production. In H. Méloni (Ed.), *Fondements et perspectives en traitement automatique de la parole*, 31-53, Aupelf Uref, 1996.

[8] H. McGurk and J. Mac Donald. Hearing lips and seeing voices. *Nature*, 264: 746-748, 1976.