L'intégration bimodale de l'anticipation du flux vocalique dans le flux consonantique

Emilie Troille & Marie-Agnès Cathiard

Institut de la Communication Parlée, Université Stendhal-INPG, UMR 5009,
38040 Grenoble cedex 9, France
Mail : troille@icp.inpg.fr ; cathiard@icp.inpg.fr

ABSTRACT

It is well known that speech can be seen before it is heard: this has been repeatedly shown for the vowel rounding anticipatory gesture leading the sound (Cathiard [6]). In this study, the perception of French vowel [y] anticipatory coarticulation was tested throughout a voiced fricative consonant [z] with a gating paradigm. It was found that vowel auditory information, as carried by the noise of the fricative, was ahead of visual and even audiovisual information. Hence the time course of bimodal information in speech cannot be considered to display the same pattern whatever the timing of the coordination of speech gestures. As concerns vowel information only, consonantal coarticulation can carry earlier auditory information than the vowel itself, this depending of the structure of the stimulus. In our fricative-vowel case, it was obvious that the vowel building movement was audible throughout the fricative noise, whereas the changes in formant grouping occurred later.

1. INTRODUCTION

D'importantes questions demeurent concernant l'intégration des informations visuelles et auditives en dépit des apports significatifs à ce champ de recherche depuis une vingtaine d'années (cf. Schwartz [1] pour une revue). Dire que la parole est avant tout bimodale peut sembler aujourd'hui être une affirmation triviale, même si le poids de la contribution de l'une et l'autre des modalités n'est pas encore complètement déterminé. Audition et vision jouent-elles toujours en complémentarité, comme on a pu le conclure à partir de paradigmes contrastant une présentation audiovisuelle à une information monomodale (avec une information auditive soit dégradée par du bruit ou bien parfaitement audible mais sémantiquement difficile à comprendre) ?

Afin de mettre en évidence la contribution des informations auditives et visuelles présentées seules ou associées, nous avons retenu un cas classique en parole, celui de l'anticipation vocalique d'arrondissement, qui nous offre une situation dans laquelle l'audio pourra être testé sans dégradation par le bruit, tout en évitant les effets plafonds habituellement rencontrés dans les expériences qui ne dégradent précisément pas l'information auditive.

L'anticipation d'arrondissement vocalique a fait l'objet de nombreuses études tant au niveau articulatoire qu'au niveau perceptif. Il est ainsi établi que ce geste vocalique peut débuter plusieurs consonnes avant la voyelle arrondie cible (pour une synthèse concernant l'anglais cf. Perkell [2] ; pour le français cf. Abry et Lallouache [3]). La récupération auditive de cette anticipation a été mise en évidence en français par Benguerel et Adelman [4], et plus récemment par Ferbach-Hecker [5], cette dernière étude montrant que l'anticipation était maximale avec des consonnes intervocaliques fricatives plutôt qu'occlusives.

L'identification visuelle de l'anticipation labiale a été extensivement étudiée par Cathiard [6] au cours de pauses acoustiques silencieuses. Les séquences UHI et IHI étaient insérées dans la phrase « tu dis : UHI ise ? » [tydi#uiiz], avec une pause (#) longue de 460 ms. Des mesures articulatoires montraient que le geste d'arrondissement s'établissait au cours de la pause silencieuse de telle sorte que la position en protrusion de la lèvre supérieure et en constriction, caractéristique de la voyelle arrondie [y] à venir, soit établie dès le début acoustique de celle-ci. Les tests d'identification indiquaient une anticipation de la perception visuelle du [y] pouvant aller jusqu'à plus de 200 ms avant le début acoustique de la voyelle.

Les études ayant directement comparé les modalités auditive et visuelle sont peu nombreuses. Escudier et al. [7], testant entre autres des séquences [zizy], concluaient à une avance du visuel sur l'audio (de 40 à 60 ms), qui est vraisemblablement, de notre point de vue, à relier à un geste de constriction pour [y] extrêmement précoce, soit dès la deuxième partie de la voyelle [i]. Dans une étude plus récente Roy [8] comparait la perception auditive et visuelle de sujets malentendants et normo-entendants pour des séquences [iCny] (Cn représentant un nombre variable de consonnes). En modalité auditive, la perception de la voyelle arrondie se situe au moment du relâchement consonantique lorsque l'intervalle n'est composé que d'occlusives, et à partir de l'inflexion de la limite inférieure du bruit de friction lorsque l'intervalle contient une fricative. Pour ce dernier cas, Roy constate que les sujets normo-entendants sont plus performants en modalité auditive qu'en modalité visuelle pour percevoir le trait d'arrondissement, mais sans que l'on puisse estimer précisément l'avance.

Dans cette étude, nous nous proposons de tester l'établissement de l'information vocalique à travers le flux consonantique sous 3 conditions de présentation : auditive, visuelle et audiovisuelle.

2. EXPERIENCE PERCEPTIVE

2.1. Enregistrement

Nous avons enregistré audiovisuellement, à 25 images/seconde, de face et de profil, un locuteur masculin français, prononçant en ordre aléatoire 12 répétitions des 2 séquences « t'as dit ZIZU ze ? » et « t'as dit ZIZI ze ? ». Ces phrases nous permettront d'explorer la transition de la voyelle [i] vers la voyelle [y] avec une consonne [z] intervocalique. Il est bien connu que cette fricative est perméable aux effets de coarticulation (Öhman [9]).

L'enregistrement a été réalisé en chambre sourde, au moyen du poste « Visage-Parole » de l'ICP (Lallouache [10]). Le son est échantillonné à 22050 Hz. Un maquillage en bleu des lèvres du sujet permettra par la suite l'incrustation d'un noir saturé, à l'aide d'un Chroma-Key, afin de permettre la détection des paramètres labiaux.

2.2. Analyse des données

Les images numérisées ont été détramées, afin d'obtenir une image toutes les 20 ms. La détection de 2 paramètres articulatoires permettant de caractériser le geste d'arrondissement, soient l'aire intérolabiale (S) et la protrusion de la lèvre supérieure (P1), est réalisée à l'aide du logiciel Tacle (Audouy [11]).

Nous avons réalisé, au niveau acoustique, un suivi de formants (F1 à F4) pour toutes nos séquences afin de trouver une réalisation [zizy] et [zizi] qui soient le plus semblables possibles, au moins en ce qui concerne le début de la transition [zi…] (fig. 1). Nous avons également suivi le bruit de friction de la consonne dans les deux transitions par une analyse LPC (en utilisant une fenêtre d'analyse large de 0.04 s de manière à privilégier l'effet de coarticulation avec la voyelle suivante, cf. Munson [12]). Dans la transition [zizy] retenue, on peut observer un abaissement du bruit de friction de la consonne [z] qui passe de 5653 Hz à 3293 Hz (alors qu'il reste stable dans la transition contrôle [zizi] autour d'une valeur moyenne de 6067 Hz), abaissement qui est à relier à la diminution de l'aire aux lèvres au cours du [z] par anticipation de l'arrondissement du [y] (fig. 2); on remarque aussi dès la fin du [z], une baisse du 3ème formant pour [zizy] jusqu'à 2400 Hz tandis qu'il reste aux alentours de 3000 Hz dans [zizi].

2.3. Montage des tests

Nous adoptons une technique de dévoilement progressif ou « gating » largement utilisée en perception de parole, afin de tester l'identification auditive, visuelle et audiovisuelle de l'arrondissement vocalique à travers la consonne. Nous délimitons un domaine d'exploration démarrant 40 ms avant la fin du [i] (soit, pour repère, à 1560 ms du début de la phrase) et se terminant à la fin de la consonne [z] (soit au temps de 1720 ms). Le pas du gating étant de 20 ms, nous obtenons 9 séquences tronquées, que nous répétons 10 fois pour ZIZU, auxquelles nous ajoutons 3 séquences pour ZIZI (les séquences tronquées respectivement à 1660, 1680 et 1720 ms, répétées 5 fois ; ces séquences permettent au sujet d'entendre et/ou voir de vraies séquences ZIZI sans pour autant alourdir la passation en proposant les 9 séquences tronquées de ZIZI).

Le découpage audio du domaine perceptif toutes les 20 ms est réalisé sous Praat, à l'aide d'un Script. Les différentes séquences créées commencent toutes au début de « t'as dit » et se terminent aux différents temps de la zone de *gating*. Pour le test audiovisuel, nous ajoutons sous Adobe Première Pro 1.5., les images détramées. Pour le test visuel, nous montons les images avec seulement le début de la séquence audio, soit « t'as dit », pour servir d'amorce attentionnelle au sujet. Les tests sont ensuite insérés dans le logiciel Multimédia Toolbook., qui permet le tri aléatoire des séquences présentées aux sujets.

2.4. Sujets

26 sujets de langue maternelle française, sans déficience ni auditive ni visuelle (ou corrigée) sont testés. Aucun n'est familiarisé avec ce type de test, ou n'a de connaissance particulière en lecture labiale. Le groupe est composé de 2 hommes et de 24 femmes, de 20 à 25 ans (21 ans et 3 mois en moyenne). Ils commencent tous par le test audiovisuel, suivi du test audio, et terminent par le test visuel : nous avons retenu cet ordre car nous souhaitions recueillir en premier l'identification audiovisuelle sans qu'il y ait eu d'entraînement préalable en condition unimodale. La tâche des sujets consiste à d'identifier la voyelle finale [i] ou [y] de chaque séquence tronquée.

3. RESULTATS

3.1. Résultats par modalités

Nous présentons les courbes d'identification, pour tous les sujets confondus, obtenues pour [zizy].

(1) L'identification auditive augmente rapidement et de façon très nette entre 1620 et 1640 ms (fig. 3). La voyelle [y] est identifiée à plus de 80% à 1640 ms, soit clairement dans la consonne [z] qui précède, puisque le début effectif de cette voyelle est situé à notre dernier temps de gating, soit à 1720 ms. Les identifications individuelles sont relativement regroupées autour de la courbe moyenne, la frontière à 50% d'identification [y] variant de 1610 à 1640 ms.

(2) L'identification visuelle évolue très lentement (fig. 4). La frontière à 50% correspond au temps 1680 et l'identification du [y] à plus de 80% n'est atteinte que 20 ms avant le début acoustique de la voyelle. On observe un empan de variation des identifications individuelles à 50% de 1640 à 1710 ms, soit 70 ms, ce qui reflète un comportement perceptif visuel variable selon le sujet.

(3) Les résultats du test audiovisuel montrent que la voyelle arrondie est perçue à plus de 80 % à 1680 ms en condition bimodale (fig. 5). A 1660 ms, [y] est déjà perçu à 76 %. La dispersion autour de la courbe moyennée est intermédiaire par rapport à celles observées pour les 2 conditions unimodales, avec un empan à 50% d'environ 50 ms (de 1620 à 1670 ms).

3.2. Comparaison des modalités

La figure 6 présente les courbes d'identification moyennées pour les 3 conditions (auditive A, audiovisuelle AV et visuelle V). L'identification auditive de la voyelle [y] apparaît être la plus précoce. Elle est suivie de l'identification audiovisuelle, puis de l'identification visuelle. L'analyse des données individuelles indique que cet ordre est observé chez 24 sujets sur 26.

Nous avons réalisé une analyse Probit (Finney [13]) pour extraire les frontières et les pentes pour chacun des sujets dans les trois conditions. Nous réalisons, sur les frontières ainsi que sur les pentes, une analyse de la variance à un facteur intra-sujets, soit la condition de présentation (A, AV et V). En ce qui concerne les frontières, après correction de l'homogénéité des variance (Greenhouse-Geisser), la condition de présentation a un effet significatif : $F(2,38)=172.42$ avec $p=0.000$, avec A > AV > V. Ainsi, la frontière moyenne à 50% est à 1629 ms en audio, 1649 ms en audiovisuel et 1673 ms en visuel.

L'étude des pentes montre également un effet de la condition : $F(2,50)=7.451$; $P<0.001$, avec cette fois l'homogénéité des variances vérifiée sans correction (test de Sphéricité de Mauchly). On ne trouve pas de différence de pente entre la condition auditive et la condition audiovisuelle ($F<1$). On peut donc associer les pentes de l'audio et de l'audiovisuel pour les comparer au visuel. On obtient cette fois une différence significative : $F(1,25)=11.18$; $P<0.003$. La bascule d'identification est ainsi comparable en audio et en audiovisuel, mais plus lente en identification visuelle.

3.3. Relation perception-production

La comparaison des identifications auditive et audiovisuelle au suivi du bruit de friction de [z] (fig. 7) montre que la perception de la voyelle [y] augmente dès que le mouvement de bruit de friction commence à descendre. On note un retard de 20 ms de l'identification audiovisuelle sur l'identification auditive, [y] étant perçu auditivement lorsque la fréquence de [z] passe en dessous de 4000 Hz contre 4500 Hz en audiovisuel.

En vue de face, le paramètre visuel le plus facile à suivre étant l'aire aux lèvres (fig. 8), nous le comparons au suivi de l'identification visuelle et audiovisuelle [y]. [y] est ainsi identifié lorsque l'aire aux lèvres passe en dessous de 1 cm² en condition audiovisuelle : l'identification passe de 23% lorsque S est à 1,12 cm² au temps 1640 ms à 76% lorsqu'elle est à 0,85 cm² au temps 1660 ms. En revanche, la bascule est plus tardive et plus lente pour l'identification visuelle : elle s'établit au temps 1680 ms, lorsque l'aire au lèvres est inférieure à 0,62 cm². Il semble qu'en vision seule, les sujets suivent plutôt l'évolution de la protrusion de la lèvre supérieure (fig. 9), comme s'ils attendaient une forme aux lèvres à la fois bien arrondie et protruse pour être sûrs de l'identification de la voyelle.

En résumé, en modalité auditive, les sujets utilisent au maximum les informations acoustiques pour identifier [y] en suivant l'abaissement du bruit de friction de la consonne [z] comme un indice d'arrondissement de la voyelle à venir, tandis qu'en condition visuelle, les indices articulatoires ne leur permettent pas d'être aussi précoces. L'abaissement du bruit de friction est bien dû au mouvement de constriction labiale du [y], et il semble qu'un petit mouvement audio suffise aux sujets pour commencer à identifier [y], alors qu'il leur faut attendre un seuil de constriction parfois inférieur au cm² pour identifier visuellement la voyelle. Dans la condition audiovisuelle, il semble que les sujets se soient laissés influencer par les deux modalités à la fois.

4. DISCUSSION

Notre objectif était de tester la perception uni- et bimodale du flux vocalique en présence d'un flux consonantique continu.

(1) Les résultas en audio seul mettent en évidence une perception précoce de la voyelle arrondie dans la consonne qui la précède, puisque le [y] est perçu dès la fin du premier tiers de la consonne (avance de près de 90 ms dans une consonne de 120 ms). L'identification audiovisuelle suit avec un retard de 20 ms, tandis que la perception visuelle est la moins précoce, à moins de 50 ms du début acoustique de la voyelle. Concernant l'avance de la vision sur l'audition défendue par Cathiard [7], rappelons qu'elle se produisait pour une voyelle produite à l'initiale, sans consonne, et avec une structure prosodique démarcative (pause silencieuse) : dans ce cas précis où le flux vocalique est perceptible isolément, la vision est forcément gagnante puisque l'audio est délivré de manière naturelle beaucoup plus tardivement. Dans le cas de notre séquence [zizy], les informations articulatoires (constriction labiale) et acoustiques (bruit de friction) sont parfaitement synchrones ; de plus, le flux vocalique se trouve en présence d'un flux consonantique. Dans ce cas précis, l'auditeur piste la voyelle aussitôt que possible à travers la consonne en s'appuyant sur l'évolution des zones d'énergie depuis l'abaissement du bruit de friction de la consonne jusqu'au 3ème formant de la voyelle.

(2) Nous démontrons également que l'identification audiovisuelle se situe temporellement entre les deux présentations unimodales. On pourrait rapprocher ce résultat de certaines interprétations de l'illusion McGurk qui défendent un percept audiovisuel juste intermédiaire entre les stimuli audio et visuel. Mais n'oublions pas que nous ne sommes pas, dans cette expérience, dans le cas d'une information conflictuelle, mais dans le cas d'un flux congruent, qui s'avère récupérable plus précocement dans une modalité que dans l'autre. On pourrait aussi penser que notre résultat en audiovisuel contredit la règle de la supériorité de l'information bimodale sur l'information unimodale : cette règle est en réalité valable en situation bruitée, ce qui n'est pas non plus notre cas. Nous sommes donc en présence d'une situation où, avec un son parfaitement audible et un flux visuel non contradictoire, on observe néanmoins un certain retard (20 ms en moyenne) de la perception audiovisuelle sur la perception auditive.

En conclusion, nous avons pu tester l'établissement de la perception de l'anticipation vocalique : l'information auditive vocalique, portée par le bruit de la fricative, est en avance sur l'information visuelle et même sur l'information audiovisuelle. Ainsi, l'évolution temporelle de l'information bimodale en parole dépend fortement du timing de la coordination des gestes articulatoires. En ce qui concerne l'information vocalique seule, la coarticulation consonantique peut la délivrer plus précocement auditivement que visuellement : dans le cas étudié ici le mouvement d'établissement de la voyelle est audible dans le bruit de friction de la fricative tandis que les changements dans les groupements formantiques se produisent plus tardivement.

BIBLIOGRAPHIE

[1] J.-L. Schwartz. La parole multisensorielle : Plaidoyer, problèmes, perspective. *XXVème Journées d'Etudes sur la Parole, pp. 11-18, Fès, Maroc, 19-21 avril 2004.*

[2] J. S. Perkell. Testing theories of speech production : implications of some detailed analyses of variable articulatory data. In W.J. Hardcastle & A. Marchal (Eds), *Speech Production and Speech Modelling*, pp. 263-288, K.A.P., London, 1989.

[3] C. Abry. & T. Lallouache. Le MEM : un modèle d'anticipation paramétrable par locuteur, données sur l'arrondissement en français. *Bulletin de la Communication Parlée N°3*, 85-99, 1995.

[4] A.-P. Benguerel & S. Adelman. Perception of Coarticulated Lip Rounding. *Phonetica*, 33, 113-126, 1976.

[5] V. Ferbach-Hecker. La perception auditive de l'anticipation des gestes vocaliques en français. *Thèse de Sciences du Langage, Strasbourg*, 2002.

[6] M.-A. Cathiard. La perception visuelle de l'anticipation des gestes vocaliques : cohérence des évènements audibles et visibles dans le flux de la parole. *Thèse de Psychologie Cognitive*, Grenoble, 1994.

[7] P. Escudier, C. Benoit & T. Lallouache. Identification visuelle de stimuli associés à l'opposition /i/-/y/ : étude statistique. *Acte du Premier Congrès d'Acoustique, Lyon, 10-13 Avril, Suppl. au Journal de Physique*, 2, 541-544, 1990.

[8] J.-P. Roy. Etude de la perception des gestes anticipatoires d'arrondissement par les sourds et les malentendants. *Thèse de Sciences du Langage, Strasbourg*, 2004.

[9] S. E. G. Öhmann. Numerical model of coarticulation. *Journal of the Acoustical Society of America, 41(2)*, 310-320, 1967.

[10] M.-T. Lallouache. Un poste "Visage-parole" couleur. Acquisition et traitement automatique des contours des lèvres. *Thèse de l'ENSERG, Spécialité: Signal Image Parole*, Grenoble, 1991.

[11] M. Audouy. Logiciel de traitement d'images vidéo pour la détermination de mouvements des lèvres. Projet de fin d'études, option génie logiciel, ENSIMA Grenoble, 2000.

[12] B. Munson. Variability in /s/ Production in Children and Adults : Evidence from Dynamic Measures of Spectral Mean. *Journal of Speech, Language and Hearing Research, 47(1)*, 58-69, 2004.

[13] D.-J. Finney. Probit Analysis. Cambridge University Press, (3ème édition) 1971

.

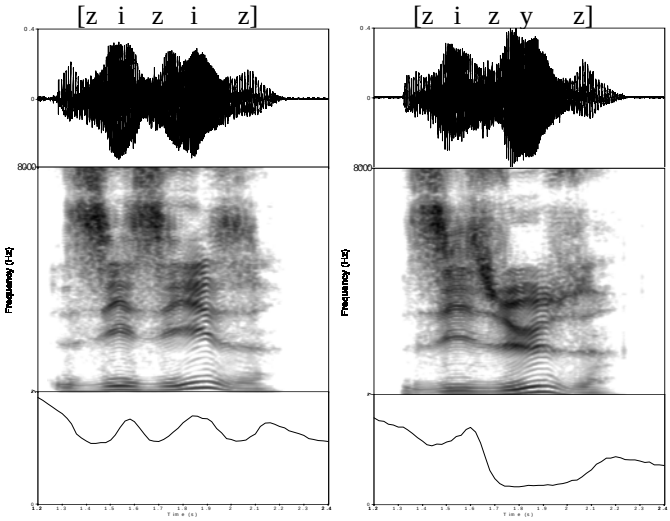

Fig. 1 : signal acoustique, spectrogramme (échelle de fréquence de 0 à 8KHz) et aire aux lèvres (de 0 à 2 cm²) pour [ziziz] (à gauche) et [zizyz] (à droite).

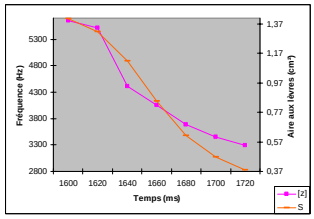

Fig. 2 : suivis de la fréquence du bruit de friction et de l'aire aux lèvres.

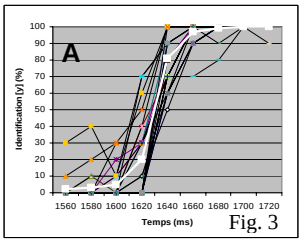

Fig. 3

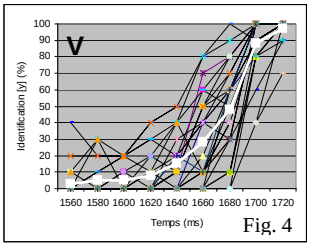

Fig. 4

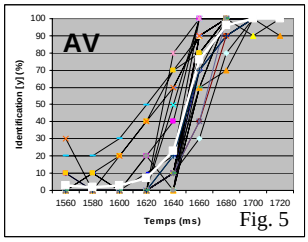

Fig. 5

Fig. 3, 4 et 5 : comparaison des courbes auditive, visuelle et audiovisuelle globales à la distribution des courbes individuelles.

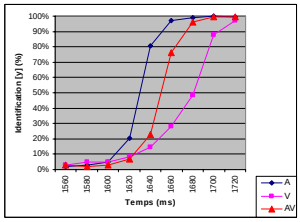

Fig. 6 : comparaison des courbes d'identification moyennées auditive, visuelle et audiovisuelle.

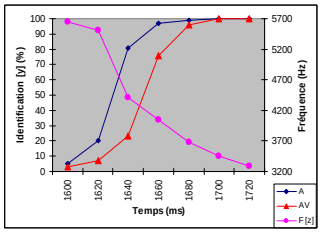

Fig. 7 : comparaison des courbes d'identification auditive et audiovisuelle au suivi du bruit de friction de [z].

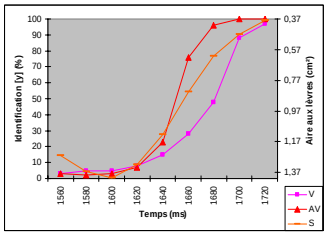

Fig. 8 : comparaison de l'aire aux lèvres (échelle inversée) aux courbes d'identification visuelle et audiovisuelle.

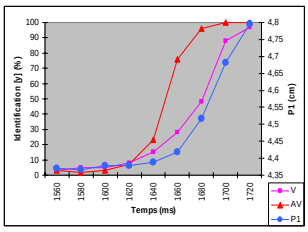

Fig. 9 : comparaison de la protrusion aux courbes d'identification visuelle et audiovisuelle.