Etude de la coordination gestes manuels/parole dans le cadre de la désignation

Benjamin Roustan

UNIVERSITÉ DE GRENOBLE

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE

Spécialité : **Ingénierie de la Cognition, de l'interaction, de l'Apprentissage et de la création**

Arrêté ministériel : 7 août 2006

Présentée par

ROUSTAN Benjamin

Thèse dirigée par **DOHEN Marion**
et codirigée par **SCHWARTZ Jean-Luc**

préparée au sein du **gipsa-lab**
et de l'**École Doctorale Ingéniérie pour la Santé et la Cognition et l'Environnement**

Étude de la coordination gestes manuels / parole dans le cadre de la désignation

Thèse soutenue publiquement le **10 Octobre 2012**,
devant le jury composé de :

Mme GUIDETTI Michèle
Pr Université de Toulouse le Mirail, Octogone, Rapporteur
Mme PELACHAUD Catherine
DR CNRS, TELECOM ParisTech, Rapporteur
Mme TELLIER Marion
MCF Université d'Aix-Marseille, LPL, Examinatrice
Mr COLLETTA Jean-Marc
Pr Université Stendhal (Grenoble III), LIDILEM, Examinateur
Mme DOHEN Marion
MCF Grenoble-INP, gipsa-lab, Directrice de thèse
Mr SCHWARTZ Jean-Luc
DR CNRS, gipsa-lab, Co-Directeur de thèse

UNIVERSITÉ DE GRENOBLE

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE

Spécialité : **Ingénierie de la Cognition, de l'interaction, de l'Apprentissage et de la création**

Arrêté ministériel : 7 août 2006

Présentée par

ROUSTAN Benjamin

Thèse dirigée par **DOHEN Marion**
et codirigée par **SCHWARTZ Jean-Luc**

préparée au sein du **gipsa-lab**
et de l'**École Doctorale Ingéniérie pour la Santé et la Cognition et l'Environnement**

Étude de la coordination gestes manuels / parole dans le cadre de la désignation

Thèse soutenue publiquement le **10 Octobre 2012**,
devant le jury composé de :

Mme GUIDETTI Michèle
Pr Université de Toulouse le Mirail, Octogone, Rapporteur
Mme PELACHAUD Catherine
DR CNRS, TELECOM ParisTech, Rapporteur
Mme TELLIER Marion
MCF Université d'Aix-Marseille, LPL, Examinatrice
Mr COLLETTA Jean-Marc
Pr Université Stendhal (Grenoble III), LIDILEM, Examinateur
Mme DOHEN Marion
MCF Grenoble-INP, gipsa-lab, Directrice de thèse
Mr SCHWARTZ Jean-Luc
DR CNRS, gipsa-lab, Co-Directeur de thèse

Résumé

Le travail présenté dans cette thèse vise à étudier la coordination entre gestes manuels et parole lors de la production d'énoncés multimodaux. Cette coordination est étudiée plus précisément dans le cadre de la désignation, réalisable à la fois dans la modalité manuelle et en parole. Les études présentées sont menées dans un environnement contrôlé de laboratoire afin d'obtenir des mesures précises et reproductibles. Un travail particulier de mise en place des protocoles a néanmoins permis de maintenir des tâche assez naturelles afin de ne pas induire des productions trop artificielles.

Les deux premières études s'intéressent à la production conjointe de gestes manuels et de parole contenant de la focalisation. Plusieurs types de gestes sont comparés (geste de pointage, geste de battement et geste d'appui sur un bouton). Il est montré que la production de focalisation attire le geste manuel quel que soit son type mais que l'attraction est plus « précise » et fine pour le pointage. Par ailleurs, l'apex du geste de pointage semble être cooccurent à une cible articulatoire plutôt qu'acoustique. La seconde étude manipule le lien de désignation le geste de pointage et la parole. Elle montre, en exhibant deux stratégies adoptées par les participants, la complexité des mécanismes mis en jeu dans cette coordination.

Finalement, une troisième étude s'intéresse à la coordination dans une tâche interactive et collaborative plus naturelle. Les résultats montrent une cooccurrence de la partie du geste qui montre avec l'information qui lui est complémentaire en parole. La perturbation de l'interaction par un bruit ambiant modifie les productions : la parole subit un effet Lombard classique et la production de gestes semble adaptater de la durée de la partie du geste qui montre à l'allongement de la parole.

Ce mémoire propose enfin une exploration des procédés d'annotation multimodaux mis en place pour l'annotation de tâches semi-contrôlées mais applicables à des cas plus généraux. Le manuscrit se conclut par une mise en perspective des résultats pour l'amélioration de certains modèles de production conjointe gestes manuels/parole et fournit quelques pistes utilisables dans le domaine des agents conversationnels ainsi que pour la détection de pathologies.

Mots-clefs

multimodalité, gestes manuels, parole, désignation, pointage, battement, Optotrak, video, interaction, coordination

---

Abstract

The work presented in this manuscript aims at describing the coordination between manual gestures and speech during the production of multimodal utterances. Since it is possible to designate using either manual gestures or speech (or both), the coordination is studied within this framework. All the experiments presented have been conducted in a lab-environment in order to obtain precise and reproducible data. Nevertheless, substantial thinking about experimental setups have been done so as to end up with rather natural tasks and try to avoid "fake" multimodal productions.

The first two studies shed light on the coordination when producing manual gestures and narrow-focused speech. Various types of manual gestures have been taken into account (pointing, beats, button-press). It is shown that focus production attracts manual gesture production whichever its type. Moreover, pointing gesture is more precisely and more consistently attracted than the other types of gestures. Interestingly, manual gestures seems to be coordinated with articulatory targets rather than acoustic ones. The second study modifies the "designation" link that unites pointing gesture and speech. It shows the great complexity of mechanisms underlying coordination by raising two strategies of coordination used by participants.

The last experiment uses a more natural interactive and collaborative task. Results show a cooccurrence in production between the part of manual gesture that shows and its complementary information found in speech. Moreover, hindering the interaction using noise has an influence on multimodal productions: a classical Lombard effect is observed for speech and the part of gesture that shows is lengthened by a duration close to the speech lengtening.

Finally, this manuscript presents an overview of the annotation process that was set up to label the aformentionned tasks. This annotation process can be generalized and thus constitutes an important part of the work. The conclusion put into perpective the results found in order to enhance existing models of manual gestures/speech coproduction and suggests applications in fields such as embodied conversational agents or pathology detection/rehabilitation.

Keywords

multimodality, manual gestures, speech, designation, pointing, beats, Optotrak, video, interaction, coordination

Remerciements

Cette thèse a été réalisée au sein du laboratoire gipsa-lab dans le département Parole et Cognition, au sein de l'équipe Parole, Cerveau, Multimodalité et Développement.

En premier lieu, je tiens à remercier toute l'équipe dans laquelle j'ai été intégré durant ces quatres années. Une équipe dynamique et pleine de curiosité dans laquelle il a été très agréable de travailler. En particulier, les nombreuses réunions d'équipe ont permis d'entretenir une grande réactivité scientifique ainsi que de faire évoluer l'équipe dans la bonne humeur. Une pensée toute particulière va bien évidemment à mes encadrants : merci à Marion pour toutes ces idées et ton enthousiasme, merci à Jean-Luc pour ton soutien important, ces riches discussions qu'on a pu avoir et ton éternel entrain.

Je tiens également à remercier tous les membres du jury qui se sont déplacés à Grenoble pour venir évaluer les travaux de cette thèse. Merci a Jean-Marc Colletta d'avoir accepté de présider le jury lors de la soutenance, merci aux rapporteurs Michèle Guidetti et Catherine Pelachaud pour leurs retours instructifs et constructifs. Enfin, merci à Marion Tellier d'être venue appuyer ce jury lors de la soutenance. Cette journée restera un bon souvenir.

Merci à tous les membres du laboratoire gipsa que j'ai pu rencontrer lors de mon séjour... Merci à toute l'équipe administrative qui nous rend la vie plus facile tous les jours, et en particulier à Nadine ! Merci pour ta disponibilité et tes "coups de gueule" par mail qui permettent au département de ne pas oublier certains fondamentaux ! Une pensée particulière va bien sûr à tous les collègues doctorants... Merci au Dr Amélie et à AudRey (vive R), de m'avoir soutenu dans les moments difficiles, de m'avoir battu au tennis et d'avoir participé aux *pauses café* (indispensables moments de détentes dans un labo). Merci aussi à Javier (mention spéciale : quasi-inconditionnel de la pause), Yo (pour sa bonne humeur, son grain de folie et sa passion du Japon), Nico (pareil, mais avec le ski de rando à la place du Japon), Olha (pour les sorties patin à glace :p), Thomas et Maëva (pour l'accompagnement à la piscine le midi et les bons conseils de bons chercheurs !) d'avoir partagé cet instant de détente ! Bien sûr, une petite pensée pour ceux "de Stendhal", en particulier Chloé, Rosario, Sandra, Krystyna (toujours Stendhal dans le coeur !), Hien, Maria, ... et à toutes celles et ceux que j'ai oublié !

La thèse a également été l'occasion de s'impliquer dans la réalisation de quelques "aventures" temporaires scientifiques... Merci à toutes les personnes qui se sont embarquées dans l'aventure ELIPSCE, tout n'a pas été facile, mais on a réussi, et la relève est assurée ! Merci en particulier à la présidente sans failles Mumu (maintenant Mymy apparement) qui a su *driver* de main de maître l'association pendant 3 ans ! Merci à tous les doctorant(e)s que j'ai rencontré par le biais de l'asso, en particulier toute la troupe du LPNC : Mathilde, Fleur, Lucie, Marcela, Alice et toute la clique... Une autre aventure a été l'organisation des RJCP, une réalisation compliquée, mais on a passé de super moments (et bien dansé autour de l'accordéon diatonique) : merci à Atef, Amélie, Hien, Mathilde, Rosario, Sandra et encore bravo pour l'organisation !

Finalement, la thèse n'a pas empêché de garder quelques activités en dehors du labo (et heureusement ! !). Merci aux copains du water-polo de la fac et aux entraîneurs, à bientôt dans les bassins. Merci aux collègues pompiers du Grésivaudan et de l'agglo qui m'ont aidé à me changer les idées lorsqu'il le fallait (Fred, je n'aurai pas de bonus pour la thèse, mais tu m'aura quand même bien supporté, merci LôLô et Seb') et avec qui on aura fait quelques bornes en basket !

Bien entendu, tout cela n'aurait pas été possible sans le support de la famille : merci Pôpa et merci Môman d'y avoir cru jusqu'au bout et de m'avoir soutenu tout le long (et c'était super de vous voir le jour de la soutenance ! ! j'espère que ce n'était pas trop incompréhensible !), merci au frère et à la sœur pour les encouragements à répétition également !

Enfin je tiens à remercier infiniment tous les amis grônoblois (bien que beaucoup soient partis). Que de bonnes sorties dans le coin. Merci a CC pour ses discussions macaroniques sur l'internet, aux Grumos d'envoyer autant de slip, à Edouardo et son Arduino (et sa passion pour la recherche !), Guiguizzo et sa vie de parisien addict au métro, El Chapón pour qui le RU et surtout les patates à l'eau était une grande passion (trop vite écourtée par la fin du doctorat...), Zrel et sa petite famille, la p'tite Mala et la grande LiLi... et à tous ceux que j'ai oublié !

Bon, vous l'aurez compris, après 24 mercis : MERCI à toutes et à tous !
et merci aux plus courageux qui liront le manuscrit ! :)

Table des matières

Table des figures

Liste des tableaux

Glossaire

$\lessapprox$	inférieur, s'approchant du seuil de significativité
apex	L'apex d'un geste manuel est l'instant où le geste atteint son "but". En particulier pour le geste de pointage, c'est l'instant où le bras ainsi que les doigts sont en extension.
coordination	Deux évènements sont coordonnés lorsque leur réalisation se produit à des instants séparés par un laps de temps constant au fil des productions.
Optotrak	Dispositif de suivi tridimensionnel de mouvements actifs utilisant trois caméras suivant l'évolution de diodes émettrices infrarouges
synchronie	Deux évènements sont synchrones lorsque leur réalisation se produit au même moment. C'est un cas particulier de coordination.

Acronymes

F_0	fréquence fondamentale
ACP	Analyse en Composantes Principales
ANOVA	analyse de la variance (ANalysis Of VAriance)
CV	cible articulatoire
EEG	electroencéphalographie
Foc	focalisation
Foc1	Focalisation sur le début de l'énoncé
Foc2	Focalisation sur la fin de l'énoncé
GP	Growth-Point
Int	intensité
IRMf	imagerie par résonnance magnétique fonctionnelle
MEG	magnétoencéphalographie
NAN	Not A Number
TMS	stimulation magnétique transcrânienne (transcranial magnetic stimulation)

Chapitre 1

Introduction

1.1 La parole est multimodale

1.1.1 Importance de la modalité visuelle dans la perception et la compréhension de la parole

Bien que l'intuition puisse faire penser que la parole est un objet purement unimodal constitué uniquement d'une suite d'évènements acoustiques, ceci est cependant clairement une conception trop minimaliste de la perception de la parole. Ainsi, comme le souligne Rosenblum [158], la perception de la parole est multimodale, une littérature conséquente prouve que les processus mis en jeu par le cerveau afin de traiter le signal de parole agissent à la fois sur le signal auditif mais également sur le signal visuel ainsi que sur le "ressenti" global de la situation d'interaction. Bien que cette affirmation puisse sembler déroutante pour beaucoup, elle devient assez évidente lorsqu'on évoque le cas des personnes malentendantes qui, malgré leur handicap, arrivent –parfois mieux que les normo-entendants– à percevoir la parole, en particulier dans des environnements bruités et donc peu propices à la communication par le biais de l'acoustique. Il en va de même lorsqu'une personne apprend une langue vivante, la lecture labiale est souvent utile afin d'aider à la perception/compréhension de la parole (voir par exemple, Sueyoshi et Hardison [170]) et l'utilisation massive des gestes manuels qui fait partie intégrante des cours de langue vivante a un effet sur la mémorisation (*cf.* les études de Tellier [173]). La faible performance du traitement auditif seul dans un environnement bruité est comblée par l'utilisation accrue de la modalité visuelle (*cf.* les expériences très connues de Sumby et Pollack [171] et plus récemment de Schwartz, Berthommier et Savariaux [164]) pour reconstituer le message transmis par l'interlocuteur : des indices visibles sur la face du locuteur permettent d'extraire de l'information sur l'articulation qui accompagne le signal acoustique en particulier, ceci grâce à l'examen des mouvements des articulateurs visibles (par exemple, les lèvres, la mâchoire, ...). Au niveau de l'importance relative des différentes parties de la face, des travaux ont montré l'importance de la zone des lèvres (pour les normo-entendants) parmi l'ensemble des indices visuels. Selon Benoit, Mohamadi et Kandel [13], la présentation de la zone des lèvres apporte en effet les deux-tiers de l'information fournie lors de la présentation du visage complet d'un locuteur. D'autres études ont montré l'intérêt des autres zones de la face –par exemple la mâchoire (voir Bateson et coll. [178]) ou la langue (voir Badin et coll. [6])– : la perception visuelle de la face d'un locuteur permet d'aider/guider/compléter la perception du signal acoustique de parole.

Ces études mettent en avant un rôle facilitateur de la perception visuelle sur la perception auditive de la parole, cependant, bien que la modalité visuelle ne soit pas indispensable (il est possible de communiquer sans se voir, par exemple, au téléphone), celle-ci joue un rôle très important lorsqu'elle est disponible. Le rôle de la modalité visuelle n'est pas simplement supplémentaire à celui de la modalité auditive mais bien complémentaire : il est possible que deux stimuli non congruents dans les deux modalités mènent à des percepts hybrides, comme présenté dans les travaux de McGurk et MacDonald [125] où la présentation d'un phonème /ba/ avec un visème représentant l'articulation du phonème /ga/ produit des percepts /da/ pour une grande partie des participants. L'étude de la façon dont les données visuelles (gestes articulatoires en particulier) et auditives sont fusionnées a fait l'objet d'un bon nombre d'études depuis les travaux pionniers de Summerfield [172] en la matière. En particulier, les travaux présentés par Robert-Ribes et coll. [155] fournissent différents modèles de fusion des

signaux audio-visuels possibles et montrent la nécessité d'une fusion multimodale dans un espace où *chaque modalité a sa tâche préférée* (l'espace "de fusion" n'est donc pas dominé par la modalité auditive).

1.1.2 L'apport des gestes dans la perception de la parole, un historique rapide...

Par delà les gestes articulatoires, le rôle des gestes non-articulatoires dans la communication n'est pas un résultat nouveau. En particulier, les gestes posturaux jouent un rôle dans la communication comme promu et démocratisé par Fast dans son best-seller "Body Language" [51]. Le cadre de ce manuscrit se réduit cependant aux "gestes" au sens de Kendon [96] : des actions caractéristiques d'une expressivité ouvertement *délibérée* (*actions that have the features of manifest deliberate expressiveness*) *i.e.* des configurations du corps qui sont réalisées dans un but principalement communicatif et, souvent en relation avec la parole.

L'intérêt des humains pour le "langage gestuel", bien qu'ayant connu un essor nouveau dans les recherches académiques récemment, n'est pas nouveau. Depuis l'Antiquité que les orateurs s'intéressent, tels des acteurs de théâtre, à leur gestualité afin de faire passer des messages le plus facilement possible aux foules d'auditeurs. Chez les Grecs, Aristote était parmi les fervents défenseurs de l'utilité des gestes dans la communication avec le public afin de transmettre des "sentiments" à celui-ci. Dès l'époque romaine, des ouvrages complets ont été consacrés à la gestualité pour les orateurs, c'est ainsi que Quintilien [149], un rhéteur romain, a consacré un volume complet de son œuvre *Institutio oratoria* (l'éducation de l'orateur) à la gestualité. Cette œuvre et les ouvrages de Cicéron ont été pendant de longues années parmi les seuls ouvrages traitant de la gestualité. Un regain d'intérêt pour les gestes peut se trouver dans la bibliographie au XVI^{e} siècle avec un livre portant sur l'art de la gestualité dans un cadre théâtral : *L'Arte de' Cenni* de Giovanni Bonifacio. A partir de cette époque, un regain d'intérêt périodique pour la gestualité est visible dans les publications. Les études se sont cependant souvent cantonnées à la caractérisation de la gestualité seule dans la communication alors que dans une grande majorité des cas, la gestualité accompagne la parole produite en parallèle. L'étude "moderne" des gestes et de leur relation avec la parole a rééellement commencé avec les études initiales de Birdwhistell [15] dans les années soixante puis les travaux séminaux de Kendon [93] et, plus tard McNeill [127].

Les études récentes ont montré l'intérêt de la gestualité manuelle dans l'accompagnement de la parole. Cependant, une multitude de questions restent ouvertes et sont le centre de débats depuis plus d'une vingtaine d'années. Globalement, trois questions fondamentales encadrent l'étude de la gestualité : *pourquoi*, *comment* et *quand* les personnes réalisent-elles des gestes ? La question *pourquoi* permet d'étudier sous un autre angle les recommandations des orateurs des siècles précédents : les gestes manuels sont-ils réellement produits pour le public ? Les deux autres questions seront abordées plus tard dans ce manuscrit.

> Remarque importante : Dans la suite du manuscrit, si la nature des *gestes* n'est pas spécifiée, on considèrera par convention que ce sont les gestes *manuels* auquel le manuscrit fait référence

1.1.3 Le rôle des gestes manuels dans le langage humain

1.1.3.1 Un débat de longue date...

Un débat qui secoue la communauté depuis de longues années porte sur l'utilité de la production de gestes manuels. Bien que l'utilité première d'un geste comme le geste "de pointage" (montrer avec son index) fasse l'unanimité (on montre pour son interlocuteur), il n'en va pas de même pour la plupart des autres gestes manuels produits lorsqu'on parle. Deux théories principales s'affrontent sur ce terrain.

Les gestes manuels (en particulier, de type iconique, *cf.* Section 1.2.1 pour une définition de ce type de gestes) peuvent, selon les théories, être utiles soit pour la personne qui les produit soit pour la personne qui les perçoit ou, de manière composites, utiles pour la compréhension (de l'interlocuteur) par leur rôle facilitateur dans la production de parole (du locuteur). Dans tous les cas, les théories s'accordent sur le fait que les gestes véhiculent, au moins en partie, de l'information complémentaire à la parole qu'ils accompagnent (voir par exemple les travaux de Goldin-Meadow [62]).

L'hypothèse communément admise pendant des années et celle encore soutenue par Kendon [92, 95] est que les gestes sont *un effort communicatif à part entière et ont un rôle direct à jouer dans ce type de processus* (p. 23) *i.e.*

les gestes manuels ont un rôle important à jouer dans la communication, mais leur rôle pour le locuteur ou pour l'interlocuteur n'est pas clairement défini. Cette vision est partagée par McNeill [128, p. 206] pour qui les gestes manuels permettent de rendre visible la structuration du discours en temps réel. Beaucoup d'études affluent dans cette direction en montrant que les gestes manuels permettent d'améliorer la compréhension du discours par les interlocuteurs, ces études prônent le rôle *communicatif* des gestes manuels. Typiquement, Goldin-Meadow et coll. [65, 60] montrent que la perception par des adultes (même naïfs) des gestes manuels des enfants leur permet de décrire le mode de raisonnement des enfants ; Cassel et coll. [27] montrent que les gestes manuels sont considérés comme de prime importance dans leur interprétation dans le flux d'information qui arrive à un interlocuteur puisque, par exemple, des gestes non congruents avec le flux de parole sont pris en compte lors de la construction de la représentation mentale de l'information transmise.

L'autre hypothèse selon laquelle les gestes joueraient un rôle *non communicatif* est défendue surtout par les travaux de Krauss et ses collègues (dont les études sont synthétisées dans [106]). Selon cette hypothèse, la production de gestes manuels aurait un rôle seulement pour le locuteur en participant à l'accès au lexique par l'intermédiaire d'un mécanisme d'amorçage entre modalités. Selon cette théorie, les humains produisent des gestes pour "faciliter" l'accès lexical et interdire aux personnes de faire des gestes rend cet accès plus complexe.

Finalement, des articles comme ceux de Bavelas [10], Özyürek [140], Driskell & Radtke [45] ou de Ruiter [37] adoptent une position intermédiaire en considérant les deux visions comme non fondamentalement opposées et pouvant cohabiter. Typiquement, ces articles avancent un rôle des gestes manuels à la fois directement dans la communication, par une amélioration de la compréhension de l'interlocuteur grâce aux gestes et dans l'accès au lexique pour le locuteur grâce à leur rôle cognitif (avancé par Krauss). On peut enfin considérer un couplage direct entre les deux théories "communicative" et "non communicative" en remarquant qu'il existe la possibilité d'une amélioration de la communication produite par une amélioration de la production de parole.

Ainsi, le rôle précis des gestes manuels reste flou dans la littérature actuelle. Deux rôles importants supplémentaires ont été assignés aux gestes manuels dans la littérature. Ces deux fonctions sont décrites ci-dessous.

1.1.3.2 Les gestes manuels dans la phylogénèse du langage

Un enjeu de débat important et récurrent porte sur la possibilité que les gestes manuels aient joué un rôle phylogénétique important comme rapporté dans les travaux de Gentilucci et Corballis [59]. En effet, des théories prônent une apparition des mots parlés par un appariement entre les sons et leur sens dans les mots (typiquement, les sons ouverts représenteraient quelque chose de large, mais les sons fermés quelque chose de petit). Cependant, cette théorie montre vite ses limites puisque la plupart des mots utilisés ne sont pas directement représentables par des sons (paradoxe de Rousseau : « La parole parait avoir été fort nécessaire pour établir l'usage de la parole »). D'autres comme Armstrong, Stokoe et Wilcox [5] mettent en avant une évolution du langage vers la communication parlée par une diminution progressive de la quantité d'information signée au fil du temps dans un système de communication multimodal initial utilisant à la fois les signes et la parole. Selon une théorie surtout défendue par Corballis lors des deux dernières décennies [34, 35], et bâtie sur des idées remontant au dix-huitième siècle (Condillac, Essai sur l'origine des connaissances humaines) et reprises par Hewes dans les années soixante-dix [72] : les gestes manuels ont pu mener au développement du langage. Parmi les animaux, seuls les humains semblent disposer d'un langage comportant la propriété de "générativité" (*i.e.* permettant un *"usage infini de moyens finis"* selon von Humboldt, 1792, en particulier, permettant la recursivité). Selon la théorie exposée par Corballis (*From hand to mouth* dans son article), le fait que l'*homo* se soit relevé a libéré les mains, lui permettant de les utiliser à des fins créatives (création d'objets, de peintures). Ces activités faisant intervenir une dimension sociale, les mains auraient alors servi à communiquer des idées en utilisant des "signes" manuels (représentées sous formes iconiques, ou de pantomimes). L'utilisation des mains pour communiquer aurait pu permettre l'apparition du langage : la pratique de la communication manuelle étant peu à peu démocratisée, celle-ci ferait ensuite partie de l'environnement et formerait une "boucle vertueuse" poussant à l'utilisation massive de celle-ci.

Des études plus récentes comme celle de Pollick et de Waal [146], portant sur les caractéristiques des gestes et des cris/mimiques des singes en captivité (chimpanzés et bonobos), ajoutent des arguments à cette théorie en

montrant une variation "culturelle" des formes de gestes entre les groupes de singes, ce qui semble moins vrai pour les cris (cependant, voir les études récentes par Lemasson et coll. [112] montrant une variation sociale dans l'organisation fine des cris au sein d'une même espèce de cercopithèques). Or l'indépendance au contexte est une caractéristique importante des langages "géneratifs", ce qui rajouterait du poids dans l'argument des gestes ayant joué un rôle d'amorce dans l'apparition du langage.

Un point épineux de la théorie de Corballis concerne la question du passage de la main à la bouche dans le langage. Selon Arbib [2, 4], un système d'imitation (reposant probablement sur un réseau de neurones miroirs) aurait mené à l'utilisation de gestes pantomimiques puis à l'émergence d'une forme conventionnalisée de ceux-ci qui aurait permis l'apparition d'une "proto-parole" [3] résultant d'une meilleure maîtrise du contrôle de l'appareil vocal. Selon cette théorie, production vocale et production de gestes manuels auraient évolué de concert en s'appuyant sur des mécanismes cérebraux communs, permettant, ici encore, l'apparition d'une "spirale vertueuse", la pratique amenant l'intégration de savoirs et encourageant à plus de pratique et donc à une géneration de nouvelles "configurations" plus importante.

1.1.3.3 Les gestes manuels dans le développement de la parole et de la pensée chez l'enfant et l'adulte

Le rôle des gestes manuels dans l'ontogénèse est sujet à moins de débats houleux dans la littérature car ce rôle semble plus consensuel. Bates et Dick [9], dans leur revue de la littérature, expriment assez bien la relation intime qui existe entre parole, geste et langage chez le jeune enfant. Susan Goldin-Meadow ainsi que les membres associés de son laboratoire ont publié de nombreux articles/ouvrages relatifs à ce rôle du geste manuel dans l'apprentissage. En particulier, les études récentes s'intéressent principalement au rôle des gestes manuels dans l'apprentissage (scolaire)/ la structuration de la pensée alors que des études antérieures avaient principalement pour but de montrer le rôle des gestes manuels dans l'acquisition de la parole.

Les travaux d'Iverson et Thelen [82] (voir Section 1.3.2.2 pour plus de détails) supposent un lien présent dès la naissance entre production de gestes manuels et production de gestes articulatoires et décrivent une co-évolution et un entraînement mutuel des deux systèmes de production dès les premières années de la vie des enfants (le reflexe de Babkin –ouverture de la bouche lorsqu'on presse sur la paume des mains du nourrisson– est présent dès la naissance).

Selon les études menées dans la première décennie des années 2000, les gestes manuels "pavent" la voie vers l'acquisition du langage (selon l'article au titre éponyme de Iverson et Goldin-Meadow [84]) chez l'enfant. En particulier, tout comme le langage est une capacité purement humaine, l'utilisation du geste de pointage afin de diriger l'attention d'autres humains semble être typiquement humaine comme indiqué par Kita [99] (mais voir les études de Call et Tomasello [24] ou, plus récemment, Russell et coll. [162] pour des contre-exemples chez les grands-singes en captivité). Ce geste de pointage semble avoir un rôle particulièrement important dans le développement du langage chez l'enfant dès 9 mois (voir Mathiot et coll. [119]). Ce geste est le geste le plus utilisé par les enfants en interaction avec leurs parents (*cf.* Guidetti [67]) et leur fréquence d'utilisation est en constante progression au cours des deux premières années. Un des rôles particulier joué par le geste de pointage est son importance dans l'émergence des premières combinaisons de deux mots : combinaisons qui n'apparaissent qu'après un certain laps de temps suite à l'apparition des premiers mots. Les études longitudinales (entre 10 et 24 mois, pour 10 enfants) rapportées dans les travaux de Butcher et Goldin-Meadow[21, 64] montrent le potentiel rôle geste de pointage dans l'apprentissage de la construction d'énoncés à deux mots chez l'enfant vers 18 mois. Typiquement, ces études montrent que l'enfant commence par prononcer des mots isolés puis devient capable de produire deux éléments d'une proposition au sein d'un seul acte communicatif multimodal par la suite : par exemple l'enfant dit *boire* et montre une bouteille pour signifier à l'autre qu'il veut boire le contenu de la bouteille. L'apparition de cette capacité de monstration gestuelle complémentant l'énoncé apparaît avant *et* prédit l'apparition de la capacité de former des énoncés à deux mots. Les évolutions observées respectivement dans les deux modalités semblent par ailleurs se poursuivre durant une grande partie de l'enfance, Colletta, Pellenq et Guidetti [33] ont en particulier montré une augmentation du nombre de gestes associée à une complexification progressive des énoncés associés chez des enfants agés de 6 et 10 ans (et chez les adultes).

Enfin, dans une suite logique des études présentées rapidement ci-dessus, l'équipe de Goldin-Meadow a mon-

tré l'importance des gestes dans l'apprentissage : la plupart du temps, les gestes manuels produits en même temps que la parole illustrent celle-ci. Il existe cependant des moments dans l'apprentissage où le geste illustre des concepts qui dépassent le cadre de ce qui est évoqué en parole. Ceci est typiquement le cas dans des tâches de conservation de Piaget –classiquement, tâches où on demande à un enfant si un même volume d'eau contenu dans deux récipients ayant des formes différentes est le même volume *i.e.* si la forme change le volume, comme [62]. Ces incongruences se rencontrent à tous les niveaux (que ce soit pour les enfants comme expliqué ci-dessus, pour les scolaires lors de l'apprentissage des équations [144] ou encore pour les adultes pour expliquer des problèmes complexes [55]) et sont en général précurseures d'un passage d'étape dans la compréhension du problème traité. Selon l'article de revue [63], les gestes ne signalent pas seulement le fait que l'apprenant est prêt à apprendre mais ils peuvent avoir un rôle à la fois direct (en changeant la manière d'apprendre) et indirect (en changeant par exemple la manière dont le savoir est enseigné) dans l'apprentissage.

1.2 La diversité des gestes

1.2.1 Une classification pour les gestes manuels

Comme mentionné ci-dessus, nombre de chercheurs ont considéré les gestes manuels comme étant intimement liés à la parole voire indissociables de celle-ci. Afin d'éviter les confusions et d'utiliser une dénomination commune pour les différents gestes manuels liés à la parole, une taxonomie des gestes a été proposée. Le problème de la multiplicité des termes utilisés pour désigner un seul et même "type" de geste rendant la comparaison d'études difficile a été soulevé par McNeill. Dans son livre de 1992 [128], il propose une taxonomie (aujourd'hui encore largement majoritaire dans la communauté) à partir des travaux précédents réalisés en particulier par Efron [46] et Ekman et Friesen [48] : quatre grandes catégories de gestes sont décrites. Il est à retenir que McNeill décrit uniquement les gestes produits avec la parole (les "gesticulations" présentées ci-après) et qu'il préfère parler de "dimensions" plutôt que de catégories de gestes, les différentes dimensions pouvant être mélangées au sein d'un même geste.

Les gestes **iconiques** représentent un objet concret ou un évènement et ont une relation forte avec le contenu sémantique de la parole qu'ils accompagnent. Les gestes **métaphoriques** sont similaires aux gestes iconiques mais représentent un concept abstrait. Deux exemples sont souvent utilisés dans la littérature et permettent d'illustrer les deux types iconique et métaphorique :

- iconique : parole *he grabs a big oak tree and bends it way back…*
  (il attrape un grand chêne et le tord vers l'arrière)
  main semble attraper quelque chose et le déplacer depuis l'avant jusqu'à son épaule
- métaphorique : parole *I wanted to ask you something*
  (Je voulais vous demander quelque chose)
  main représente une forme de "coupe"
  (la "coupe" représente le concept de question)

Les gestes **décitiques** sont les gestes de pointage (vers quelque chose, quelqu'un, de concret ou d'abstrait). Enfin les gestes de **battement** sont des gestes biphasiques constitués d'un mouvement rapide de la main (souvent de haut en bas) permettant de repérer un mot ou un énoncé comme étant significatif, leur sens propre n'est pas évident…

Les types de gestes étudiés par McNeill se limitent aux gestes qui accompagnent la parole (et qui la nécessitent le plus souvent pour être compris), cependant il existe d'autres gestes manuels qui sont compréhensibles sans être cooccurrents avec la parole. Typiquement, les personnes muettes ne parlent pas et produisent des gestes manuels : la langue des signes est un exemple de geste manuel non accompagné de parole. Ainsi, de manière plus générale, Kendon a proposé dès 1982 une classification des gestes (s'appuyant sur leur relation avec la parole) selon un continuum (appelé continuum de Kendon par McNeill en 1992 et présenté en Figure 1.1). Ce continuum contient quatre repères qui représentent quatre types de gestes qui font l'unanimité. Les "gesticulations" (qu'on appellera plus simplement gestes par la suite) sont les gestes (idiosynchrasiques) qui sont produits de manière spontanée avec la parole. Ce sont ces gestes qui sont décrits par McNeill (*cf.* ci-dessus). Un type de geste qui

n'apparaît pas toujours est le type constitué des gestes dits "*language-like*" : ce sont des gestes qui jouent un rôle grammatical dans la phrase qu'ils accompagnent (ils remplacent un mot dans la phrase), par exemple, "*passe moi le [geste de saisie qui tourne]*" pour demander un tournevis à quelqu'un. Ce type de geste n'est pas présent dans la description du continuum des gestes proposée par McNeill, dans certains articles, cette catégorie de gestes s'intercale entre les "gesticulations" et les "emblèmes". Les **pantomimes** sont des gestes qui ont un sens sans la parole mais qui ne sont pas codifiés, par exemple, un geste de l'index et du pouce "prolongeant le nez" pour répondre à la question "mais qui est Pinocchio ?" (typiquement : ces gestes sont très utilisés au théâtre)... Les **emblèmes** sont des gestes compréhensibles sans parole et codifiés, typiquement le geste représentant le mot "ok" formé en créant un cercle en joignant les bouts de l'index et du pouce et en étendant les autres doigts vers l'extérieur aux États-Unis (les emblèmes dépendent des cultures). Finalement, la langue des signes est le dernier type de geste énoncé par Kendon. Ces gestes et postures sont codifiées et constituent un système de communication linguistique complet.

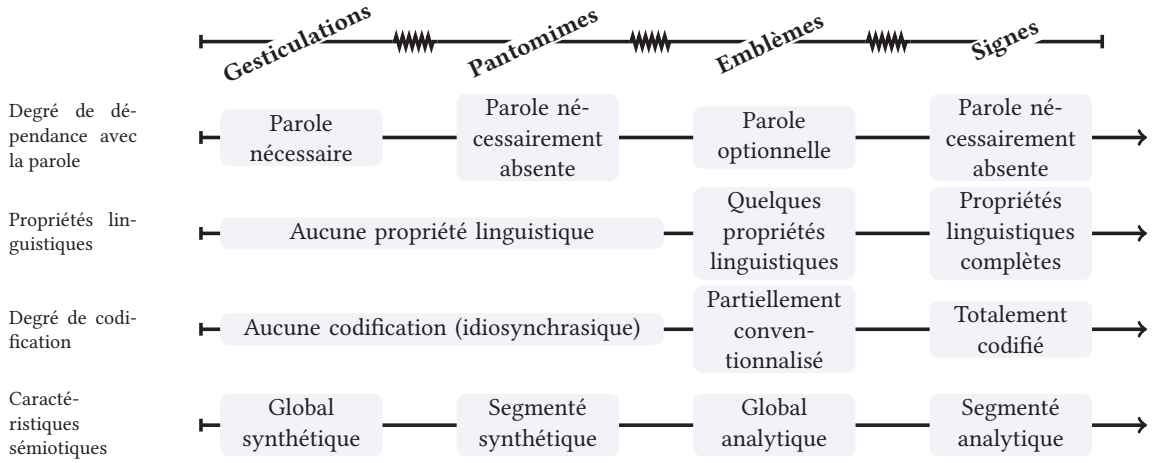


FIGURE 1.1 – Les différents continuums de Kendon tels que présentés par McNeill [129]

1.2.2 Les interprétations du continuum de Kendon

Le continuum présenté par Kendon et revu par McNeill peut s'interpréter de plusieurs façons... Chaque "lecture" du continuum représente une façon différente de classer les types de gestes et le classement dépend de la caractéristique observée sur les gestes. Ces différentes interprétations montrent que la dénomination de "continuum" est bien justifiable car pour chaque caractéristique l'évolution est "monotone" (sauf l'axe de degré de dépendance à la parole) selon le niveau de la caractéristique. Ces différentes lectures sont proposées par McNeill dans l'introduction de son ouvrage paru en 2000 [129]. Le premier continuum effectue un classement en fonction du degré de dépendance des gestes avec la parole (pour comprendre les gesticulations, la parole est absolument nécessaire ; pour les signes, la parole est inutile), le second en fonction de leur relation avec des propriétés linguistiques (aucune propriété linguistique pour les gesticulations et les pantomimes, quelques propriétés pour les emblèmes et des propriétés linguistiques complètes présentes pour la langue des signes), le troisième en fonction de leur degré de codification (les gesticulations sont idiosynchrasiques et donc non codifiées mais les signes sont une langue et sont donc totalement codifiés) et enfin le quatrième en fonction de leur caractéristiques sémiotiques (les gesticulations sont globales [1] et synthétiques [2] alors que les signes sont segmentés et analytiques).

1. Dans le langage, les éléments (les mots) sont associés pour créer un tout (une phrase), la direction est donc de l'élément vers le tout. Pour les gestes, au contraire, la direction est du tout à l'élément. (McNeill, 1992)

2. Dans le langage, la relation entre un mot et son sens est analytique : des mots différents sont attachés à des sens différents. Un geste, au contraire, peut combiner plusieurs sens à lui seul. (McNeill1992)

Un point important souligné lors de la description des continuums par McNeill est le fait que les gestes qui sont combinés le plus souvent à la parole (gesticulations) forment avec la parole des combinaisons non congruentes *i.e.* les gestes sont globaux (les parties du geste sont explicables par la compréhension du tout) et synthétiques (un geste seul peut avoir un sens qui est relatif à tout un énoncé, aussi complexe soit-il), non codifiés et n'ont aucune propriété qui pourraient les rapprocher d'un langage alors que la parole qui les accompagne est toujours segmentée (la compréhension du tout se fait par la compréhension des différentes parties), analytique (les fonctions sémantiques sont réparties sur toute la longueur de l'énoncé), codifiée et appartenant à une langue (utilisant donc des propriétés linguistiques fixées). Par ailleurs, hormis leur rôle complémentaire dans les continuums évoqués ci-dessus, les gestes apportent le plus souvent des précisions sur des points implicites dans la parole. Enfin, un rôle important des gestes coverbaux est leur implication dans la cohésion du discours, les gestes permettent d'évoquer dans la durée certains points importants du discours qui ne sont mentionnés que sporadiquement dans celui-ci. Globalement, geste et parole expriment les mêmes idées mais vues sous des angles différents…

1.2.3 Les phases gestuelles selon Kendon

Kendon propose également un formalisme afin de décrire le décours temporel des gestes. Le formalisme développé par Kendon repose ici sur les études vidéos qu'il a menées ; selon lui, pour chaque *niveau d'organisation* observable en parole, on observe une façon de gesticuler correspondante. Ainsi, de l'organisation hiérarchique de la parole, il tire une organisation semblable pour les gestes manuels. Kendon nomme unité gestuelle (*gesture-unit/g-unit*) le laps de temps séparant deux repos des bras/mains. Au sein de chaque unité gestuelle viennent se loger plusieurs (éventuellement, une unique) phrases gestuelles (*gestural phrase*) et chaque phrase gestuelle est décomposable en plusieurs phases gestuelles. Au sein d'une phrase gestuelle, seule une phase est obligatoire, cette phase est nommée *stroke*, c'est au sein de cette phase que le sens du geste manuel est exprimé. Le stroke est encadré par des phases de gestes optionnelles : avant le stroke, la phase de préparation mène la main depuis la position de repos jusqu'à la position de début de stroke, cette phase est possiblement suivie d'une tenue pré-stroke (*pre-stroke hold*) qui correspond à un "temps mort" sans mouvement de la main, avant l'exécution du stroke ; après le stroke, on trouve optionnellement de la même façon une tenue post-stroke (*post-stroke hold*) et une phase de retraction (retour à une position de repos). Selon Kendon, le *stroke* est le laps de temps saillant (*peak*) du geste, dans la suite, on nommera *apex* l'instant où le geste atteint son "but". Cet instant qui correspond à la fin du *stroke* pour le geste de pointage et est atteint lorsque le bras et la main/les doigts sont au maximum de leurs extensions respectives.

Le formalisme mis en place par Kendon et repris par McNeill quelques années plus tard est réellement une brique de base dans l'étude des gestes coverbaux : aucun "langage commun" n'était utilisé auparavant, ce formalisme largement adopté a permis une clarification des différentes études les unes par rapport aux autres.

1.3 Contexte théorique

Cette section présente le cadre théorique dans lequel s'inscrit le travail présenté dans ce manuscrit. Il est important de comprendre l'intérêt de l'étude de la gestualité coverbale. Un des intérêts principaux est la compréhension des mécanismes internes qui lient ces deux modalités. Globalement, deux lignées de théories expliquant le fonctionnement de ces mécanismes internes s'opposent dans la littérature : certaines études prônent un mode de fonctionnement dans lequel l'information est séparée en deux "flux" (l'un pour la génération de gestes, l'autre pour la génération de parole) dès le départ, chaque flux subissant des traitements différents (et éventuellement communiquant entre eux) pour arriver à une production multimodale geste/parole ; d'autres études avancent un traitement unique de l'information menant à l'extériorisation sous les formes gestuelle et en parole. La théorie de McNeill est un bon cadre théorique pour montrer les enjeux d'une telle étude, celle-ci sera expliquée rapidement dans une première sous-partie. Une notion qui a son importance dans le manuscrit est la prosodie (rhythme, intonation, phrasé de la parole), quelques études princeps ont également mis en avant le rôle de la prosodie dans la coordination, ces études seront présentées ici.

1.3.1 Le modèle de McNeill

Dans leur article récapitulatif écrit en 2008 [131], McNeill et ses collaborateurs résument globalement la théorie défendue par McNeill depuis les années 80. Les différents travaux de McNeill s'appuient sur des expériences de restitution d'histoire. Globalement, la plupart des études utilisées filment un participant qui essaie de restituer un épisode du dessin animé Titi et Grosminet (*Tweety and Sylvester*) à un autre participant qui n'a pas vu l'épisode. L'échange est filmé et une analyse vidéo est réalisée par la suite (d'autres protocoles sont également utilisés, mais celui-ci est le plus commun dans les études de McNeill). Dans toutes ses études, McNeill s'intéresse aux gestes co-verbaux, aux "gesticulations" du continuum de Kendon, c'est de ces gestes-là dont il est question ci-dessous. La théorie mise en place s'inscrit dans la continuité de ce qui a été mis en place par Slobin dans le modèle du *thinking-for-speaking* [167] : lorsqu'un acte de parole est entrepris, la pensée est organisée dans le but de la production de la parole (et en particulier, se structure de façon conforme à la langue utilisée pour la parole).

Selon McNeill, les gestes manuels et la parole sont deux modes sémiotiques qui émanent d'un processus mental commun qui s'extériorise le plus souvent par une "synchronie" entre production de parole et production de gestes manuels. Ces deux modes sémiotiques sont à la fois co-expressifs (ils contribuent de concert à faire émerger un sens de la production et participent au dynamisme communicatif du discours) et différents (pour des raisons exposées ci-avant, en particulier, les gestes sont globaux et synthétiques alors que la parole est segmentée et analytique, par exemple). Ce sont les différences entre les deux modalités qui les rendent non-redondantes et la perception du tout permet l'émergence d'un seul et unique symbole. Cette dualité est possiblement assimilable à la "double essence" du langage dont parlait Saussure au siècle dernier : deux modes sémiotiques différents sont cooccurrents.

Un point central dans l'argumentation de McNeill est l'existence d'une unité minimale d'un code combinant l'imagerie et le langage. Cette unité minimale s'appelle le Growth-Point (GP). L'existence d'un GP se traduit par une synchronie entre geste et parole et une coexpressivité de ces deux modalités lors de la production de parole. Les deux modes sémiotiques présents simultanément dans un GP le rendent "instable". Or l'instabilité joue un rôle moteur dans le modèle du *thinking-for-speaking* [130] et permet ici l'extériorisation de l'idée : lors de la production, un GP est "développé" (*unpacked*) et les deux formes parlée (linguistique) et gestuelle (imagée) rentrent en interaction. Une structure grammaticale est mise en place afin d'accueillir la partie linguistique, cette structure grammaticale permet de limiter l'instabilité, de remodeler la pensée et de faire émerger la forme finale geste manuel/parole. La grammaire joue donc ici un rôle important dans la stabilisation du processus. La résolution du problème d'instabilité venant des deux modes sémiotiques présents dans le GP a pour but de trouver une façon efficace de représenter l'information dans les deux modalités. Il est à noter que ce processus est, selon McNeill, réalisé de façon totalement parallèle dans les deux modalités (les traitements interagissent entre eux), aucune modalité n'est traitée avant l'autre.

Une étude de cas souvent citée qui permet de justifier cette théorie est le cas de IW [126]. Ce patient a subi une désafférentation de tout le corps sous le niveau du cou à l'âge de 19 ans. Suite à cette intervention, le patient n'a plus du tout de proprioception mais a quand même réussi à rétablir son contrôle moteur en utilisant la cognition et sa vision. Lorsqu'on empêche à IW de voir ses mains, il ne peut pas réaliser d'actions avec ses mains. Par contre, même lorsqu'il ne voit pas ses mains, celui-ci arrive parfaitement à réaliser des gestes cooccurrents avec la parole (gestes bien formés et en synchronie avec la parole) et à adapter la vitesse de ses gestes à la vitesse de la parole… Ce comportement peut être expliqué avec les briques théoriques mises en place par McNeill et sa théorie du GP en montrant le statut particulier des gestes coverbaux et met bien en avant la relation entre pensée/parole/gestes manuels.

Typiquement, l'étude des productions issues de l'extériorisation des GP (synchronie entre deux modalités sémiotiquement différentes) peut permettre de "remonter" à des étapes antérieures de son "développement" (*i.e.* avant l'extériorisation) et prouve un traitement unifié de l'idée (*i.e.* l'origine commune aux réalisations dans les deux modalités) résultant en une production synchrone lors de l'extériorisation.

1.3.2 Modèles de production conjointe geste manuel/parole

Comme montré ci-dessus, les études ayant pour but de caractériser les liens entre parole et gestes manuels sont assez nombreuses et regorgent de résultats qu'il est intéressant de synthétiser par la mise au point de modèles de production. Comme l'argumente de Ruiter [38] dans sa thèse, l'approche par modèles type "traitement de l'information" est une approche utile car elle permet de mettre au clair le fonctionnement de différents "modules" et de leurs connexions : elle permet de créer un idéal scientifique de la façon dont fonctionne une entité réelle. Confronter ces modèles à la réalité permet de les valider ou les falsifier et donc de les améliorer. Le grand avantage de ces modèles est donc de permettre une compréhension précise de chaque élément représenté dans le modèle, et alors de s'affranchir d'un *homunculus* qui règlerait les points problématiques. . . La plupart des modèles présentés ci-dessous sont de ce type (bien que plus ou moins détaillés et donc "falsifiables" selon les modèles).

Une revue de littérature des modèles de production de la parole et de gestes coverbaux est proposée ci-dessous. Globalement, trois visions de la production conjointe de gestes manuels et de parole s'opposent [82]. Dans tous les cas, deux "flux" (parole, geste) sont produits par les modèles, les différences entre les trois visions viennent de l'instant auquel les deux flux se séparent totalement effectivement.

- Selon une première vision (soutenue par exemple par Butterworth et Beattie, Levelt, Hadar [22, 114, 69]), les systèmes de production de geste et de parole sont deux entités distinctes : les seuls liens qui existent entre les deux processus sont totalement unidirectionnels (de la parole vers le geste). Selon cette vue, le geste sert d'ersatz de parole lorsque celle-ci est indisponible, ainsi selon ces modèles la parole peut avoir une influence sur les gestes mais l'inverse est totalement exclu. Peu de modèles s'inscrivent dans cette vision.
- Une vision moins modulaire, où les gestes ont un simple rôle d'aide à l'accès au lexique, est défendue par quelques chercheurs, suite à une théorie exposée par Krauss. Dans ces travaux [106, 152, 115], une interaction (bidirectionnelle) entre production de parole et production de gestes manuels existe mais est très localisée dans la planification des deux flux. Typiquement, les travaux de Krauss proposent un effet facilitateur de la production de gestes lors de l'accès au lexique grâce à un amorçage multimodal des entrées du lexique. Dans ces modèles, les gestes interagissent avec la parole simplement afin de fournir une "aide" pour celle-ci. Seuls quelques modèles sont inspirés de cette vue (voir Section 1.3.2.4).
- Une dernière vision s'appuie directement sur la théorie mise en place par McNeill. *Mutatis mutandis*, dans ces modèles (typiquement,ceux de Kita et Özyürek ou Hostetter et Alibali [100, 78]) geste et parole sont issus d'un processus similaire et évoluent à partir d'une même représentation interne de façon conjointe et totalement interactive. Selon cette vision, le geste influe sur la parole et *vice et versa*. C'est généralement cette vision qui est adoptée dans les modèles présentés ci-après.

La plupart des modèles présentés ci-après ne concernent que les gestes "coverbaux" *i.e. a priori* seulement les "gesticulations" au sens de Kendon ainsi qu'éventuellement les emblèmes.

1.3.2.1 Le modèle de Tuite : les pulsations rythmiques (1993)

Ce modèle ainsi que celui d'Iverson et Thelen présenté ci-après ancrent leurs théories sous-jacentes dans l'entraînement mutuel de plusieurs systèmes et reprennent des idées liées aux systèmes dynamiques [90].

Ce modèle (présenté en 1993 dans [177]) s'appuie fortement sur la théorie de production de McNeill selon laquelle la représentation interne servant de base à la production conjointe de geste et parole vient d'une image et d'un symbole de parole intérieure générés de façon conjointe. Par ailleurs, les études de Kendon ont montré des liens temporels entre la production de gestes manuels et celle de parole (règle de synchronie phonologique) et l'analyse des données proposée par Tuite met en avant une régularité dans les instants de production des gestes, qui selon lui, pourraient traduire l'existence de "pulsations kinésiques" (ou pulsations rythmiques) sous-tendant cette régularité. Ces pulsations rythmiques se retrouveraient soit dans la production de gestes de battements, soit, lorsqu'il n'y a pas de gesticulation, dans des gestes rythmiques de la tête, des pieds, des paupières, . . .

Selon Tuite, ces pulsations rythmiques se traduisent gestuellement par la réalisation de *strokes* et en parole par des pics d'intonation (ces deux évènements n'étant pas strictement synchrones car la production de parole requiert des étapes plus complexes dans sa préparation).

Une représentation schématique du modèle est proposée en Figure 1.2. Ce modèle ne décrit pas en détail le fonctionnement de chaque entité : selon l'auteur le but est de poser un modèle de départ sur lequel se baser par la suite... La caractéristique de ce modèle réside principalement dans l'existence d'une "horloge interne" (les pulsations rythmiques) qui impose une synchronie entre les flux de gestes manuels et de parole.

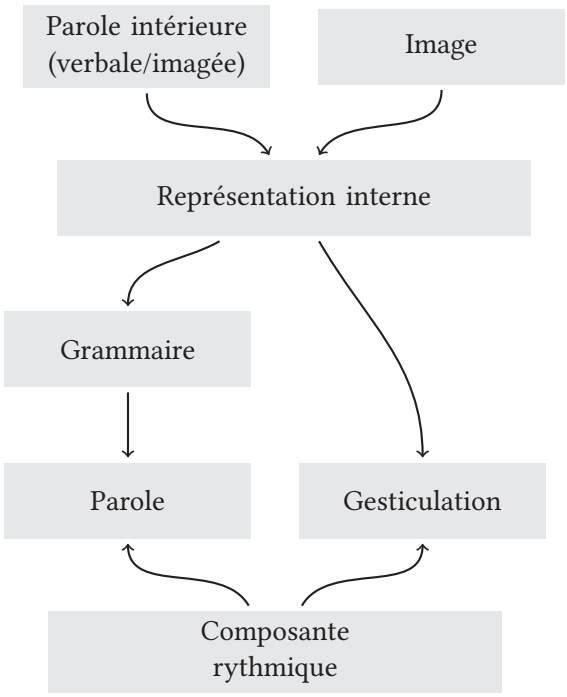


FIGURE 1.2 – Modèle de Tuite : les pulsations rythmiques (d'après [177])

1.3.2.2 Le modèle développemental d'Iverson et Thelen (1999)

Ce modèle présenté dans [82] est surtout un modèle développemental qui décrit la mise en place de la coordination gestes manuels/parole chez l'enfant. Ce modèle s'appuie très largement sur la théorie des systèmes dynamiques pour le contrôle moteur selon laquelle plusieurs acteurs coopèrent afin de former des motifs. Dans le cadre de cette théorie, plusieurs sous-systèmes fonctionnent au sein d'un système global et chaque sous-système réalise un "comportement". Deux concepts sont fondamentaux dans ces théories : les seuils et les forces d'activation. Un seuil pour un comportement est reflété par sa facilité d'exécution : pour un comportement donné, un seuil d'activation "faible" indique une action facilement réalisable (et donc plus souvent réalisée, *a priori*). La force d'activation d'un comportement représente la "force" du comportement une fois le seuil d'activation dépassé, typiquement, les comportements stables, maîtrisés ont une grande force d'activation par rapport aux "nouveaux" comportements. Une hypothèse importante est que, pour que deux systèmes se "couplent" (*i.e.* fonctionnent "de concert", en formant des motifs), ceux-ci doivent avoir des hauts niveaux d'activation (*i.e.* deux comportements non maîtrisés ne peuvent pas être couplés).

L'évolution des systèmes de production orale et manuelle menant à des productions geste manuel/parole synchrones serait jalonné par quatre étapes principales chez le jeune enfant. Dès la naissance et pendant les (environ) dix-huit premiers mois, les deux systèmes évoluent jusqu'à permettre à l'enfant de produire des "énoncés" très proches de ce que réalise l'adulte (au moins au niveau de la synchronie entre les deux modalités). Les quatre étapes sont résumées ci-dessous :

1. Les systèmes de production de voix et de gestes manuels sont déjà (faiblement) couplés à la naissance. Le réflexe de Babkin et l'exploration orale d'objets vers deux mois montrent un couplage de l'activité manuelle

et orale. Ces comportements sont largement observés chez les enfants et indiquent un seuil bas ainsi qu'un haute activation pour l'activité conjointe geste manuel/voix.

2. Le contrôle à la fois des articulateurs et des mouvements des mains progresse rapidement et vers trois ou quatre mois, des mouvements rythmiques sont observés à la fois dans le système bras/main et dans le système vocal/oral. Autour de six/huit mois, la quantité de mouvements rythmiques du système bras/main augmente sensiblement et, à un instant très proche, on aperçoit l'apparition du babillage chez le jeune enfant. Ainsi, le "babillage vocal" et le "babillage manuel" (par exemple, le contact rythmique de l'index droit sur la paume de la main gauche) semblent émerger d'une interaction/d'un entraînement mutuel entre les deux "oscillateurs" bras/main et vocal/oral. Il est possible, en particulier, que l'oscillateur bras/main, une fois entraîné (seuil bas et activation relativement haute), entraîne/facilite l'oscillation de l'oscillateur vocal/oral.

3. Autour de neuf mois, les jeunes enfants changent leurs productions à la fois gestuelles et orales. Petit à petit, les mouvements rythmiques des mains laissent place à des gestes plus communicatifs (pointages déclaratifs, impératifs) et le babillage vocal laisse place à des productions plus communicatives également, ressemblant à des mots... Ici encore, bien que les deux modalités voient leurs changements arriver à des instants proches dans le temps, la modalité verbale tend à changer avec un petit temps de retard sur la modalité gestuelle. Ceci est possiblement dû au fait que le seuil pour la communication dans la modalité gestuelle est plus bas que pour la modalité vocale chez le jeune enfant, cette modalité est donc préférée (bien que les deux modalités restent liées).

4. Vers un an et demi, l'enfant est dans une phase d'apprentissage de nouveau mots. Cette phase demande beaucoup de concentration et implique beaucoup de pratique. Cette pratique intense permet au seuil lié à la parole de baisser puis au niveau d'activation de monter en flèche. Le couplage avec l'oscillateur lié à la gestualité manuelle mène à un entraînement de l'oscillateur bras/main par l'oscillateur vocal/oral puisque celui-ci est fortement sollicité lors de l'acquisition de nouveaux mots. Une fois cet entraînement réalisé, les deux oscillateurs s'entraînent mutuellement : des gestes liés à des mots sont produits, de manière synchrone... et ce tout au long de la vie.

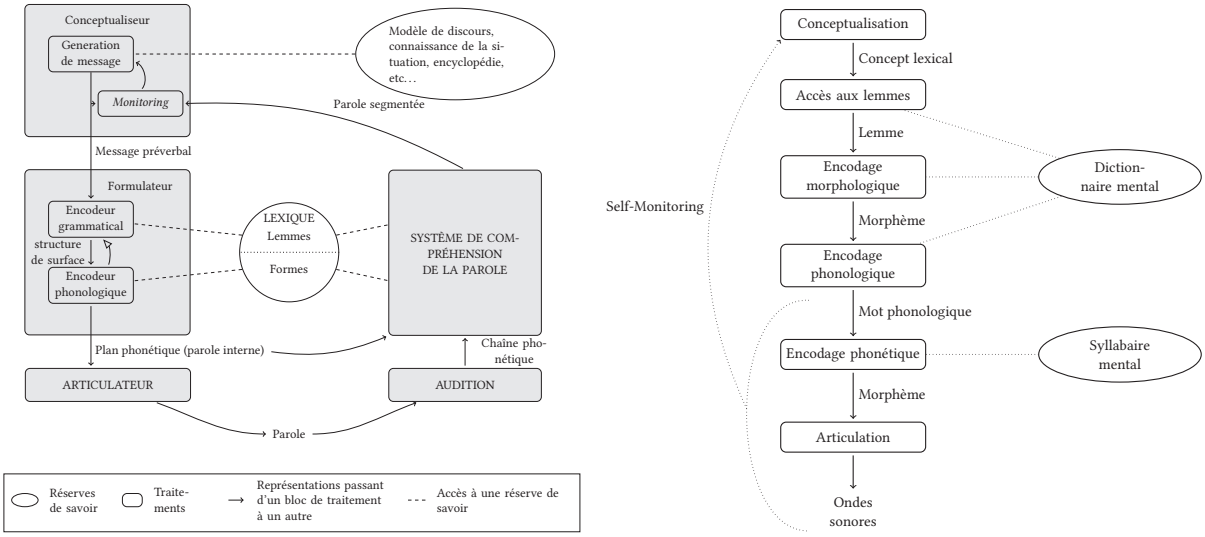
Ce modèle est le seul à prendre en compte l'établissement du duo geste manuel/parole depuis la naissance jusqu'à l'âge adulte. La présentation de ce modèle (et du précédent) à partir de la théorie des systèmes dynamiques est intéressante puisqu'elle propose une approche qui est totalement exclue des modèles présentés ci-dessous avec une grande importance du rythme dans la mise en place des deux systèmes gestes manuels et parole. Selon la vue présentée ici (et ci-dessus), le rythme, et la synchronie des deux modalités qui en découle sont des briques de base à la fois pour la mise en place et pour le fonctionnement des systèmes de production.

1.3.2.3 Le modèle de production de parole de Levelt (1989)

Dans son livre paru en 1989 [115], Levelt présente un modèle de production de la parole représenté en Figure 1.3. Ce modèle n'inclut aucune composante manuelle. Cependant, il est ici présenté car il servira de base pour plusieurs modèles qui suivent. Dans le schéma proposé en Figure 1.3a (modèle *blueprint* de Levelt), les boîtes rectangulaires représentent les traitements et les ellipses des "réserves de savoir" (*i.e.* où sont stockées les représentations).

Différentes étapes sont proposées par Levelt :

– la **conceptualisation** consiste en la création d'une "intention communicative", la sélection des informations à exprimer, l'ordonnancement de ces informations, la conservation d'une trace de ce qui vient d'être dit et au *monitoring* de la production en cours. La conceptualisation est le fruit de deux types de connaissances : des connaissances procédurales (représentées par les rectangles) et des connaissances déclaratives (connaissance du monde, de la situation etc... stockées dans la mémoire à long terme).
  Le message préverbal est la représentation qui sort du conceptualiseur, représentation *a priori* propositionnelle, synthétisant ce que doit contenir la parole (une liste de spécifications sémantiques).



(a) Modèle de production de la parole de Levelt (b) Production d'un mot selon Levelt

FIGURE 1.3 – Modèle de production de la parole et du mot selon Levelt (d'après [115])

- le **formulateur** transforme une structure conceptuelle en une structure linguistique. Il est décomposé en deux sous-étapes qui interagissent afin de former le plan articulatoire (aussi appelé parole interne), représentation d'une séquence de mouvements articulatoires à réaliser, qui constitue la représentation de sortie de l'étape de formulation. L'encodage grammatical consiste en l'accès aux lemmes du lexique (symboles abstraits représentant des mots comme des entités sémantiques/syntaxiques) et en la construction de la syntaxe. Un lemme contient le sens des objets qu'il représente ainsi que les informations syntaxiques sur ces objets. Un lemme est activé lorsque son sens correspond à une partie du message préverbal, une fois celui-ci sélectionné, il met à disposition de l'encodeur grammatical la syntaxe qui lui est applicable, l'encodeur active alors ses procédures de construction syntaxique, le résultat de ces process est la "structure de surface".
  L'encodeur phonologique transforme la structure de surface en un plan articulatoire en utilisant les représentations phonologiques et morphologiques dans le lexique, des traits diacritiques ainsi que l'information phrasale permettent d'adapter quelques propriétés de la parole interne générée.
  Il est à noter que le feedback fourni par l'interprétation du plan articulatoire par le système de compréhension de la parole permet une action sur de potentielles erreurs avant même que celles-ci soient prononcées.
- enfin l'**articulateur** exécute le plan articulatoire fourni par le formulateur, produisant de la parole audible.

1.3.2.4 L'hypothèse d'accès au lexique de Krauss (1995-2000)

Une série d'articles (plus d'une dizaine) a été publiée par Robert M. Krauss et plusieurs collaborateurs autour d'une théorie prônant un rôle communicatif marginal des gestes coverbaux, laissant la place à un rôle important de ceux-ci dans l'accès au lexique. Dans un article [104] reprenant les résultats des études précédentes, le modèle mis au point par Krauss et Hadar [105] et ayant légèrement évolué est présenté. Ce modèle représenté en Figure 1.4 s'appuie sur l'architecture proposée par Levelt [115] pour la production de la parole (en grisé sur la Figure 1.4) sur laquelle vient se greffer une partie de traitement de l'information menant à la production de gestes manuels.

Dans ce modèle, les auteurs distinguent donc *deux* systèmes de production qui fonctionnent en parallèle et en interaction. Les gestes considérés sont les gestes dits "lexicaux" (gestes "coverbaux", gesticulations de Kendon). Selon les auteurs, plusieurs formats de représentation (au minimum : propositionnel et spatial) sont utilisés dans la mémoire afin de représenter les différents concepts et l'activation d'un concept en mémoire active les concepts

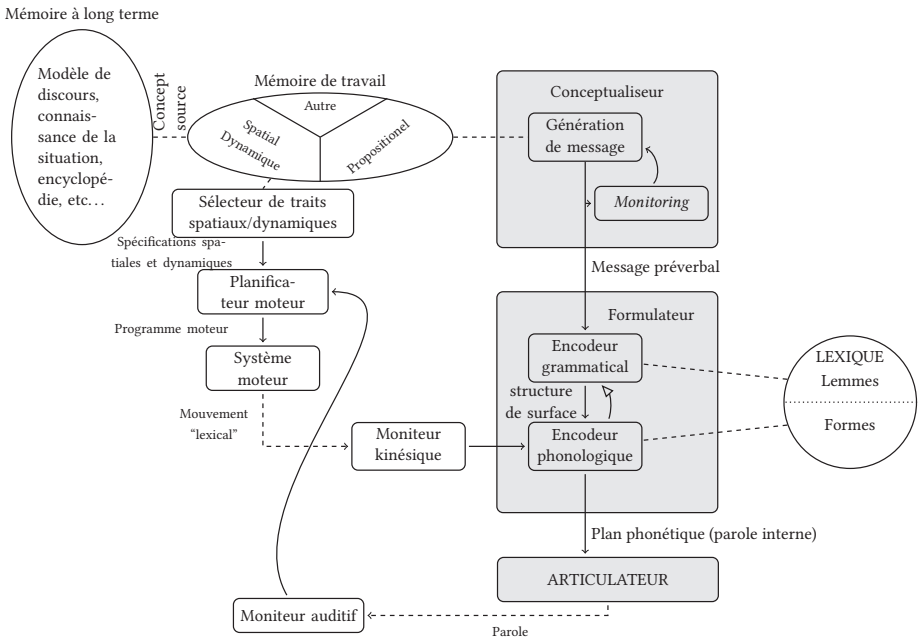


FIGURE 1.4 – Modèle de production gestes manuels/parole selon Krauss, Chen et Gottesman (d'après [104])

liés (et ce, dans tous les formats de représentation). Deux hypothèses centrales sont avancées pour mettre en place le modèle : lorsque le geste et la parole représentent des concepts non identiques, ce qui est représenté par le geste ne fait pas partie de l'intention communicative (par exemple, pour un locuteur parlant d'un "gros gâteau" et faisant un geste représentant la forme du gâteau –rond–, le concept "rond" ne fait pas partie de l'intention communicative) et les gestes lexicaux sont issus de concepts utilisant un format de représentation spatial.

Ainsi, pour un concept source donné, ce concept peut-être représenté par des traits spatiaux et/ou propositionnels, certains traits spatiaux sont sélectionnés par un "sélectionneur de traits spatiaux/dynamiques" et sont transformés en un ensemble de caractéristiques spatiales/dynamiques à leur tour transformées en des programmes moteurs par le planificateur moteur. Un exemple fourni dans [107] est celui du mot "tourbillon". Les traits spatiaux liés à ce mot sont typiquement une taille, une forme et des éléments dynamiques tels que la trajectoire, la vitesse de révolution… les traits propositionnels sont *a priori* tout ce qui n'est pas spatial/dynamique (le tourbillon est composé d'eau, il peut happer des élements à l'intérieur etc…

De manière similaire, des traits propositionnels sont sélectionnés par le conceptualiseur et sont passés à l'encodeur grammatical. Lors de l'exécution du geste par le système moteur, celle-ci est contrôlée par le "moniteur kinétique" qui agit sur l'encodeur phonologique par effet d'amorçage : les traits spatiaux sélectionnés (et extériorisés sous forme de gestes manuels) permettent de faciliter l'accès lexical pour les traits propositionnels qui leur sont directement associés. Le type d'encodage des différents traits représentant le concept source étant *a priori* variable selon les traits, il est possible que le geste lexical contienne des traits non exprimés par la parole. Dans ce cas, aucun effet facilitateur n'est possible (pas d'effet d'amorçage). L'exécution de l'articulation est surveillée par un moniteur articulatoire qui prévient le planificateur de gestes de la fin d'énonciation d'un concept afin de permettre la fin du geste.

Quelques points importants du modèle sont discutés par les auteurs de l'article, en particulier, la décomposition des concepts en "traits" spatiaux/propositionnels. D'autres modèles (typiquement [128, 38]) font l'hypothèse de gestes représentés de manière holistique (sous forme d'images) dans un "dictionnaire" de gestes. Cette vision n'est *a priori* pas viable pour les auteurs puisqu'elle nécessiterait un processus "intelligent" qui puisse abstraire les traits spécifiques d'un objet/d'un concept et le transformer en un ensemble de gestes, par ailleurs, les images étant spécifiques, elles ne peuvent pas être généralisées (une image d'un gâteau rond ne peut pas se généraliser en une image d'un gâteau carré). Enfin il n'y a pas de façon d'apprendre une image d'un concept abstrait (cas des gestes métaphoriques). Ainsi, le concept de "dictionnaire de gestes" semble peu probable aux auteurs.

Bien que les auteurs soient très virulents vis-à-vis des travaux utilisant des “dictionnaires de gestes”, aucun contre-exemple invalidant cette théorie n’est proposé. Dans une grande majorité des cas, les critiques faites aux travaux sont que les auteurs de cet article (Krauss, Chen, Gottesman) ne “connaissent pas de données permettant de valider” les autres modèles. . . cet argument est clairement insuffisant pour invalider les autres théories.

1.3.2.5 Le Sketch Model (hypothèse de maintien de l’image) de de Ruiter (1998)

Le modèle mis au point par Jan-Peter de Ruiter [38] prend le parti de gestes communicatifs. Selon lui, la vision de Krauss n’est pas incompatible avec des gestes communicatifs : les gestes peuvent avoir plusieurs rôles. Selon de Ruiter, le but premier des gestes est communicatif, mais il se peut que ce but ne soit pas atteint (éventuellement, dans une majorité des cas) ; par ailleurs, le fait que les humains fassent des mouvements bracchio-manuels même en l’absence d’interlocuteur (typiquement, au téléphone), pourrait venir du fait que la production des gestes coverbaux est tellement pratiquée qu’elle en devient automatique (mais voir [83] pour des données sur les enfants non-voyants). Toujours selon l’auteur, et contrairement aux hypothèses de Krauss, la production de gestes manuels lors d’une difficulté d’accès au lexique est réalisée grâce au feedback de la parole qui revient dans le conceptualiseur : celui-ci, constatant un problème dans la parole, décide de représenter plus d’information sous forme gestuelle afin d’améliorer la communication.

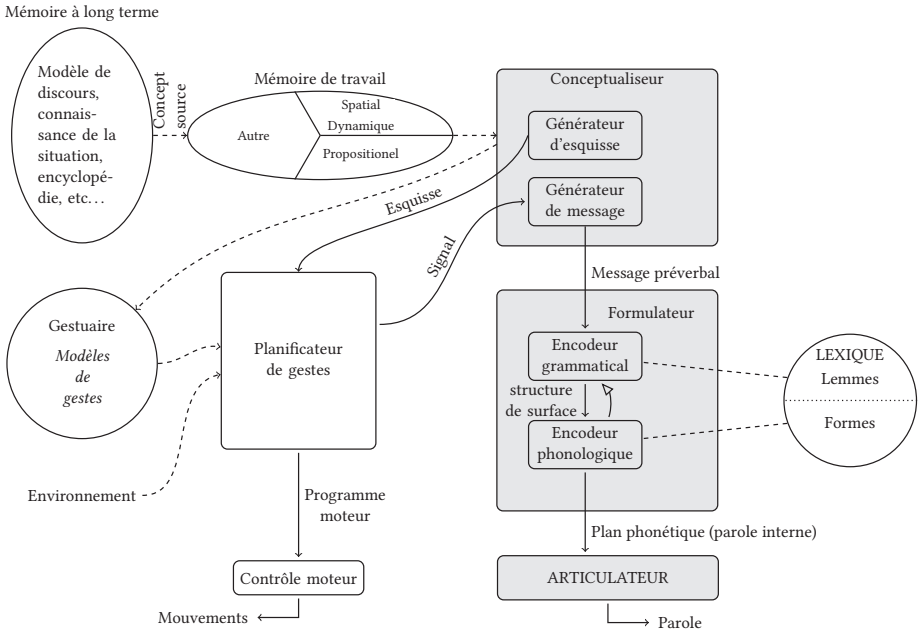


Figure 1.5 – Modèle Sketch (d’après [38])

Ce modèle présenté en Figure 1.5 s’appuie, comme celui présenté précédemment, sur les travaux de Levelt. Comme dans le modèle précédent, le modèle *sketch* laisse le modèle de Levelt inchangé et vient se greffer sur celui-ci au niveau du conceptualiseur (cette étape étant la seule ayant accès aux informations de type spatial dans la mémoire et n’étant pas cantonnée à des représentations propositionnelles). Ici le conceptualiseur est légèrement modifié pour envoyer à la fois un message contenant des représentations (images) de mouvements appelé esquisse (“*sketch*”) à un module de géneration des gestes manuels et un message préverbal comme dans le modèle de Levelt. Globalement, un rôle supplémentaire est associé au conceptualiseur qui fonctionne selon la devise “un geste vaut mieux qu’un long discours” (ce qui est en particulier vrai pour les concepts spatiaux ou dynamiques), dans le modèle de de Ruiter, le conceptualiseur décide que ce qui est difficilement encodable sous forme de parole doit être encodé sous forme de gestes (lorsque cela est plus simple).

Selon le type de geste coverbal à réaliser, l’esquisse qui est envoyée au planificateur de gestes (“*gesture planner*”) est accompagnée d’un modèle typique de geste pour les gestes déictiques et les emblèmes récupérés dans un

gestuaire, ou dictionnaire de gestes. Pour les autres gestes, l'esquisse est générée directement à partir du concept à illustrer. Le planificateur de gestes doit ensuite construire un programme moteur à faire exécuter par le système moteur (points similaires avec le modèle de Krauss). Le planificateur permet, en outre, la résolution de conflits (réalisation d'un geste manuel alors que les mains sont utilisées), l'enchaînement/fusion des gestes, la prise en compte de l'environnement (par exemple : obstacle sur la trajectoire du geste).

Au niveau temporel, ce modèle permet de faire des prédictions sur les productions relatives de gestes manuels et de parole. Afin de prendre en compte ces résultats, une hypothèse est posée (les gestes sont plus rapides à générer que la parole) et des liens entre les différents processus sont tracés : le planificateur de gestes, suite à la réception de l'esquisse, réalise le début du geste (*pre-stroke*) et attend jusqu'à ce que le message préverbal soit généré –lien entre le générateur de messages et le planificateur de gestes–, le *stroke* est ensuite exécuté, lorsque le message est terminé, le conceptualiseur envoie un signal-stop au planificateur de gestes pour que le *stroke* se termine : c'est le *post-stroke*. Enfin, si une synchronisation au niveau du mot est observée dans le cas du pointage, ceci est expliqué dans les premières versions du modèle (au début du manuscrit) par un signal de synchronisation entre le module d'exécution motrice et l'encodeur phonologique : lorsque cette synchronisation est nécessaire, l'encodeur phonologique s'arrête et attend un signal *go* envoyé par le module d'exécution motrice lorsque celui-ci connaît le temps d'exécution du geste jusqu'à l'apex. Cependant, ce signal qui était présent dans les premières versions du modèle a disparu dans les versions suivantes, laissant place à une coordination issue d'une interaction en amont de l'exécution.

Selon de Ruiter, une des seules limitations de son modèle (au moment de sa publication) est le fait que les gestes de battements ne sont pas intégrés au modèle (et ce, pour une raison de bibliographie insuffisante). Par ailleurs, celui-ci permet d'expliquer plusieurs comportements (pourquoi certaines productions ne contiennent pas de gestes coverbaux, synchronisation avec les groupes verbaux ou nominaux plutôt qu'avec les verbes/noms, non-influence des représentations du formulateur sur le déroulement des gestes).

Ce modèle diffère bien entendu du modèle ci-dessus, en particulier à cause du fait qu'il se rapproche "un peu plus" de la théorie du Growth-Point : la façon dont une intention communicative est représentée à la fois de manière vocale et gestuelle est décidée au sein d'un seul et même processus (le conceptualiseur). Dans le cas du modèle précédent, la sélection de la composante gestuelle est faite de manière indépendante de la sélection de la composante vocale. Par ailleurs, le modèle de de Ruiter est assez clair sur les signaux permettant la synchronisation des deux modalités et sur le fonctionnement des différents processus (en particulier, la sélection des contenus des gestes manuels et de la parole), ce qui permet de faire des prédictions et de les tester.

1.3.2.6 Le modèle d'interface de Kita et Özyürek (2003)

Une autre hypothèse pour la génération des gestes est proposée par Kita et Özyürek : l'hypothèse d'interface, cette hypothèse permet d'expliquer une influence à la fois de la parole sur le geste et du geste sur la parole. Dans ce modèle publié en 2003 [100], basé sur l'hypothèse d'encapsulation de l'information défendue par Kita [98] et les études d'Özyürek portant sur les différences de gestes liées à des différences linguistiques [140] (et plus tard, [141]), l'hypothèse mise en avant pour la génération de gestes coverbaux est qu'il existe une représentation à l'interface de la parole et de la "pensée spatiale". Cette représentation est de type spatio-motrice et est structurée en fonction des contraintes linguistiques. Ce modèle permet d'expliquer des résultats observés sur des participants turcs, japonais et anglais qui ne représentent pas de la même façon les mêmes trajectoires que ce soit en parole ou gestuellement. Le modèle propose ainsi l'existence d'une "ressemblance" entre la façon dont sont structurés le geste et la parole associée.

Par exemple (exemples tirés de l'article) : il existe un verbe pour "se balancer" en anglais (*to swing*), alors qu'en turc ou japonais, seuls des verbes comme "sauter" ou "voler" existent ; il existe un verbe rapidement accessible pour "descendre en roulant" (*roll down*) en anglais, mais ce n'est pas le cas en turc et en japonais (ou en français) où manière et trajectoire de l'action sont encodées par deux clauses. Lors de la production conjointe de gestes et parole représentant une action de balancier, la majorité des participants japonais/turcs n'utilisent pas de mouvement de balancier dans leurs gestes alors que les participants anglais le font ; de même lorsqu'on demande de décrire une action consistant en une descente réalisée en roulant, les participants turcs et japonais utilisent

deux gestes distincts pour encoder manière et trajectoire de l'action alors que les participants anglais encodent tout en un même geste. Ainsi, les gestes produits exhibent une structuration de l'information similaire à ce qui est observé en parole. Ce modèle s'appuie en partie sur les idées de Slobin [167] et du *thinking-for-speaking* : la mise en place des gestes est contrainte par la perspective de la parole. Ainsi, ce modèle reprend la conception conjointe des gestes et de la parole telle qu'elle avait déjà été évoquée par McNeill et Duncan [130].

Le modèle présenté ci-dessous (*cf.* Figure 1.6) s'appuie ici encore sur le modèle de production de la parole de Levelt (et le modifie) mais ne permet pas, contrairement au modèle précédent, d'expliquer/prédire des comportements temporels entre les productions de gestes et de parole.

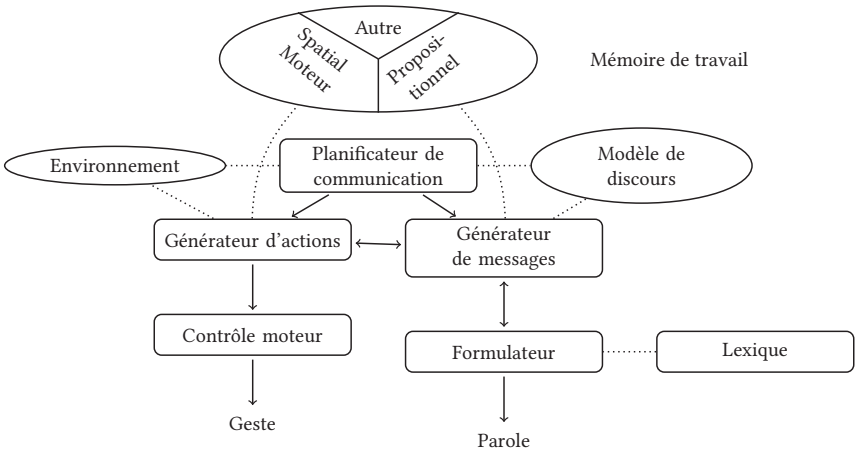


FIGURE 1.6 – Modèle des liens geste/parole selon l'hypothèse d'interface (d'après [100])

Dans le modèle proposé par les auteurs, le conceptualiseur est scindé en deux parties : le planificateur de communication et le générateur de messages. Ces deux entités matérialisent respectivement les deux processus mentionnés par Levelt : le *macro-planning* (génération de l'intention communicative *i.e.* que veut-on exprimer, dans quel ordre…) et le *micro-planning* (génération d'une proposition en prenant en compte le but de celle-ci et le contexte dans lequel elle va être exprimée). Le planificateur de communication choisit quelles sont les modalités à utiliser (sans déterminer précisément le contenu de chaque modalité) puis cette intention communicative est envoyée au générateur de messages et d'actions. Pour la production de la parole, le principe de fonctionnement est similaire au modèle de Levelt, mais il existe une communication bidirectionnelle entre le formulateur et le générateur de messages : le formulateur essaie, pour chaque intention communicative, de produire une verbalisation encapsulée en une unité, selon le résultat de cette étape (une unité possible ou non), un message est renvoyé au générateur de messages. Si la verbalisation en une unité n'est pas possible, le générateur de messages découpe l'intention en plusieurs sous-intentions et essaie de les soumettre au formulateur pour trouver une verbalisation en une unité. Ce dialogue entre générateur de messages et formulateur est constamment transmis au générateur d'actions pour que celui-ci puisse mettre au point le contenu du geste en fonction de l'encodage décidé en parole (*i.e.* la structure de celle-ci), de l'environnement (contraintes physiques) et de l'intention communicative. De façon symétrique, le générateur d'actions informe le générateur de message de la structuration de l'information qu'il lui est possible de produire : l'hypothèse de Kita sur l'encapsulation des données [98] selon laquelle les gestes aident à la segmentation des images mentales pour faciliter la verbalisation est modélisée par le lien entre le générateur d'actions et le générateur de messages. L'échange d'informations tel qu'il est décrit ci-dessus a lieu jusqu'à ce qu'une forme facilement verbalisable soit trouvée : un état d'équilibre est atteint. Une fois cet équilibre atteint, le message peut être verbalisé et une représentation spatio-motrice envoyée aux processus de contrôle moteur pour l'exécution du geste manuel.

Quelques cas particuliers sont explicables par ce modèle en lui imposant des contraintes. En particulier, cette modélisation permet d'expliquer que les gestes soient parfois différents de ce qui est exprimé en parole de deux façons : soit cela est dû au fait que les deux générateurs peuvent sélectionner l'information à représenter dans leur modalité de manière autonome ; soit cela est dû au fait que le planificateur de communication a assigné deux *buts*

distincts aux deux générateurs (l'équilibre sera alors trouvé lorsque les deux buts distincts seront atteints). Par ailleurs, la connexion du planificateur de communication au modèle de discours lui permet de spécifier les parties importantes du discours dans le but de mener à bien celui-ci. Finalement, pour une intention communicative, tous les passages dans le formulateur (sauf le dernier) ne sont pas extériorisés en parole, il se peut cependant, dans certains cas rares, que la structure du geste ne corresponde pas à l'encapsulation utilisée en parole : ceci peut être dû à un "seuil de convergence des structures" qui, lorsqu'il est dépassé, déclenche la production de gestes/parole. Dans le cas de gestes non structurés de la même façon que la parole, une variation de ce seuil pourrait être la cause de la production non correspondante.

Ce modèle a été mis à l'épreuve face à une possibilité de falsification en 2007 par les auteurs du même modèle [101] en faisant l'hypothèse que l'effet de la linguistique sur le geste ne soit pas une interaction en ligne entre ce qui est représenté dans le geste et ce qui est exprimé sous forme parlée mais plutôt que cela soit le résultat de schémas conceptuels propres à la langue utilisée. Cette hypothèse semble non verifiée et les auteurs restent donc sur l'hypothèse d'interfaces pour laquelle les locuteurs coordonnent les représentations linguistiques et gestuelles de façon dynamique et interactive.

Les trois modèles présentés ci-dessus se basent tous sur le modèle de production de la parole de Levelt en le modifiant peu (ou pas, pour Krauss) et en ajoutant une partie communiquant avec la production de la parole qui permet la production de gestes. Globalement, les différents modèles varient dans le degré d'interaction qu'il y a entre les deux "branches" (geste/parole) du modèle (faible pour Krauss, assez faible pour de Ruiter et assez important pour Kita et Özyürek) et dans l'endroit où sont sélectionnés le contenu effectif du geste manuel et son évolution (sélection en amont du conceptualiseur pour Krauss, en sortie du conceptualiseur pour de Ruiter et sélection dans le conceptualiseur avec un contenu qui varie au cours du temps –en fonction des résultats du formulateur– dans le modèle de Kita et Özyürek).

Un dernier modèle, reposant sur l'*embodiment*, est maintenant présenté : ce modèle ne respecte plus le formalisme "traitement de l'information" mais permet une approche novatrice sur les mécanismes possiblement sous-jacents à la production de gestes coverbaux.

1.3.2.7 Le framework GSA (*gestures as simulated action – gestes simulant l'action*) (2008)

Les bases du modèle proposé par Hostetter et Alibali en 2008 [78] renvoient aux théories de l'incorporation (*embodiment*). Ces théories prévoient en particulier que la perception et l'action sont fortement liées et s'influencent mutuellement. Les résultats typiques de ces théories sont, par exemple, que la vision d'un objet semble "amorcer" les actions en lien avec la saisie de cet objet et que, inversement, la saisie d'un objet permet de modifier l'environnement qui nous entoure et d'acquérir de nouvelles informations perceptives sur le geste de saisie : selon Gibson "Nous devons percevoir pour nous déplacer, mais nous devons également nous déplacer pour percevoir". La découverte des neurones miroirs et l'évocation d'une possible existence de tels neurones chez l'humain [153] ont été un argument de plus pour étayer de telles théories. Au niveau du langage, l'incorporation se traduit par le fait que le contenu sémantique des objets linguistiques est lié à la perception du monde qui nous entoure, ceci permet de justifier en partie des expériences (par exemple [71]) montrant que la lecture/l'écoute de phrases relatant des actions activent les zones motrices correspondantes du cerveau. De manière inverse, ces théories prédisent une activation des représentations sensorimotrices lors de la production. Au niveau des gestes, les modèles ci-dessus et la théorie de McNeill mettent en avant la nature globale et synthétique des gestes et prédisent une génération de gestes basée sur l'imagerie mentale (décomposable en deux sous-parties : l'imagerie mentale visuelle qui permet de se souvenir de faits passés et l'imagerie mentale motrice qui permet (par simulation, ou émulation) de "revivre" des actions passées). Selon le cadre théorique évoqué, la production de gestes serait sous-tendue par une activation des représentations sensori-motrices dans l'imagerie mentale.

Le modèle (visible en Figure 1.7) mis en place par les auteures affirme que les gestes sont issus des simulations motrices et perceptuelles de l'imagerie mentale engendrées lors de la production de parole (et donc d'objets linguistiques, pleins de sens). Le fonctionnement du *framework* GSA repose sur des notions similaires à celles proposées par Iverson & Thelen dans leur modèle. Selon ce modèle, l'extériorisation d'un geste se fait selon les variations de trois facteurs :

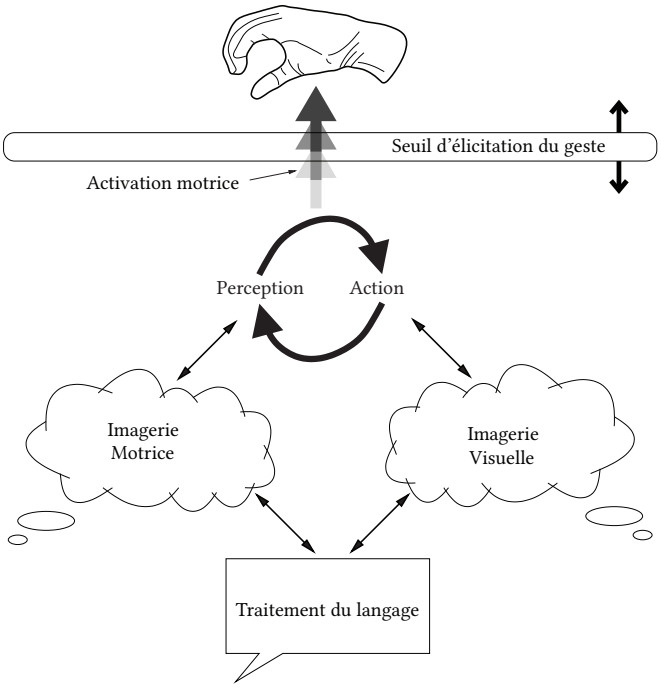


FIGURE 1.7 – Gesture-as-Simulated-Action framework (d'après [78])

- l'énergie d'activation de l'action simulée : la "force" de l'activation dépend de la force de la simulation utilisée lors du traitement linguistique, de l'âge, de la "clarté" du percept simulé, de la concrétude/abstraction des pensées utilisées pour le langage.
- le seuil d'élicitation des gestes : le seuil au delà duquel la production effective d'un geste est effectuée peut varier en fonction des expériences passées, des croyances, de facteurs neurologiques (connectivité entre les zones prémotrices et motrices par exemple), de facteurs cognitifs (désir de restreindre la production de gestes), de facteurs sociaux (utilité des gestes pour le but communicatif).
- l'utilisation simultanée du système moteur de la parole : si les articulateurs sont sollicités, des activations se propagent du cortex prémoteur (lié à la parole) vers le système moteur (lié à la parole). Il est possible que ceci rende plus simple la transmission d'un message entre le cortex prémoteur et le cortex moteur pour d'autres commandes motrices (en particulier : pour les gestes manuels).

Ici le modèle présenté ne s'appuie pas sur les travaux de Levelt sur la production de la parole : le but du modèle est d'expliquer la provenance des gestes plutôt que la production coordonnée de gestes et de parole (en particulier, les processus mis en jeu lors de la production de parole sont hors du cadre de l'article). Cependant, le modèle implique la production de la parole d'une façon qui se rapproche de la vision unifiée de McNeill selon laquelle geste et parole sont deux composantes d'un seul et même système cognitif et communicatif. Ici, la parole est obtenue par une boucle de simulations de perceptions/actions, mécanisme *a priori* absent de la théorie de Levelt.

Ce cadre théorique a été (tout comme le modèle de Kita et Özyürek) soumis à quelques tests de falsification par les auteures du modèle [79] en comparant le taux de gestes lors de la description de tâches motrices *vs* de simulation visuelle. Les essais de falsification n'ont pas abouti et appuient donc le cadre théorique mis en place initialement.

1.3.2.8 Synthèse sur les modèles

Les différents modèles exposés ci-dessus constituent une base d'explication pour les mécanismes sous-jacents à la production de parole et de gestes coverbaux. Globalement, tous les modèles ont pour but d'expliquer (au moins) la production de gestes au sens de "gesticulation" (de Kendon) –mis à part de Ruiter qui écarte les gestes

de battement de son modèle.

Comme présenté ci-dessus, les modèles diffèrent non seulement par les traitements qu'ils font subir aux différentes informations circulant en leur sein mais aussi et surtout par des hypothèses/implications théoriques importantes. Les points importants différenciant ces modèles sont résumés dans une revue de littérature proposée par Hostetter et Alibali [78] et concernent les problématiques suivantes :

- Les représentations sous-tendant les gestes sont-elles des images mentales (purement visuo-spatiales) ?
  Selon le modèle Sketch, le modèle d'interface et le modèle des gestes simulant l'action, oui. La réponse est négative pour le modèle d'hypothèse d'accès au lexique (des *traits* élémentaires permettent la mise au point de gestes) et pour la théorie du Growth-Point et le modèle des pulsations rythmiques (l'imagerie n'est pas purement visuo-spatiale). Enfin la question n'est pas traitée par le modèle d'entraînement dynamique.
- Les facteurs linguistiques ont-ils une influence sur les gestes produits ?
  Selon le modèle des pulsations rythmiques, la théorie du Growth-Point, le modèle d'interface et le modèle des gestes simulant l'action oui. La réponse est négative pour pour le modèle Sketch et le modèle d'hypothèse d'accès au lexique (deux "branches" distinctes dans le modèle avec peu d'interactions/feedback). La question n'est pas traitée par le modèle d'entraînement dynamique puisque l'étude s'intéresse surtout à des périodes avant l'apparition de la parole.
- Les gestes sont-ils porteurs d'information pour l'interlocuteur ?
  Oui pour la théorie du Growth-Point et le modèle d'entraînement dynamique. Ce n'est pas leur rôle premier selon le modèle d'hypothèse d'accès au lexique et le modèle des pulsations rythmiques. Enfin, la question n'est pas centrale dans les autres études (modèle Sketch, modèle d'interface, modèle des gestes simulant l'action).
- Le locuteur utilise-t-il consciemment le potentiel rôle communicatif des gestes ?
  Oui pour le modèle Sketch, le modèle d'interface, la théorie du Growth-Point, le modèle des gestes simulant l'action et le modèle d'entraînement dynamique et *a priori* non pour le modèle des pulsations rythmiques et le modèle d'hypothèse d'accès au lexique.
- Par quel processus la production de gestes facilite-t-elle la production de parole ?
  Selon le modèle d'hypothèse d'accès au lexique, l'aide à la parole a lieu en facilitant l'accès au lexique par un processus d'amorçage cross-modal. Selon le modèle Sketch, la théorie du Growth-Point et le modèle des pulsations rythmiques, les gestes aident la parole simplement en facilitant la communication (ce qui n'est pas nécessaire à la parole peut faire l'objet d'un encodage gestuel). Selon le modèle d'interface, les gestes permettent de faciliter la segmentation des unités lexicales à utiliser. Selon le modèle d'entraînement dynamique, les gestes sont nécessaires au développement de la parole et aident celle-ci par un entraînement mutuel des oscillateurs. Pour le modèle des gestes simulant l'action, les gestes permettent de faciliter la parole par le biais d'une sollicitation des imageries mentales.
- La production de gestes/parole se fait-elle dans un système unifié ou dans deux systèmes ?
  Selon la théorie du Growth-Point, le modèle des pulsations rythmiques et le modèle des gestes simulant l'action, un seul système unifié prend en charge la génération de la parole et des gestes manuels associés. Selon le modèle d'entraînement dynamique, deux systèmes sont présents à la naissance et des liens les unissent, ces liens se renforcent (et permettent la fusion des deux systèmes ?) par la suite. Enfin, selon le modèle Sketch, modèle d'hypothèse d'accès au lexique et modèle d'interface, deux systèmes interagissent au niveau du conceptualiseur pour le modèle Sketch et le modèle d'interface et au niveau du formulateur pour le modèle d'hypothèse d'accès au lexique.

Ces modèles sont pour la plupart (le *sketch-model* mis à part) totalement descriptifs et s'intéressent principalement aux relations pragmatiques/sémantiques entre les deux modalités. Ainsi, ils ne permettent pas de tirer des prédictions quantitatives concernant la coordination entre gestes manuels et parole au niveau temporel... mais ils fournissent des bases solides et variées, *a priori* perfectibles au niveau du déroulement temporel des évènements.

Les modèles présentés ci-dessus diffèrent sur la nature et la force du lien geste manuel/parole. Comme le souligne McNeill, l'étude de la coordination geste manuel/parole (en production) pourrait donner des information afin d'affiner ces modèles. Un angle d'attaque intéressant est l'étude de la coordination autour de la prosodie puisque celle-ci semble avoir un rôle à jouer dans la coordination entre les deux modalités.

1.3.3 Quelle importance de la prosodie dans la coordination gestes manuels/parole ?

Selon la théorie exposée ci-dessus, la coordination entre gestes et parole peut rendre compte de processus internes. Bien que McNeill ne mentionne que quelques points du "développement" du GP, hormis la grammaire, il est probable que d'autres phénomènes entrent en jeu dans le développement des GP et dans l'explication de la coordination temporelle des gestes manuels et de la parole.

On appelle prosodie l'étude des phénomènes supra-segmentaux : variations de l'accentuation et de l'intonation (variation de la hauteur –fréquence fondamentale–, de la durée, du rythme et de l'intensité) de la parole. Selon certains auteurs, il est probable que la prosodie joue un rôle important dans la synchronisation des productions de gestes manuels et de parole. De façon assez globale, une "règle" faisant la quasi-unanimité parmi les chercheurs unit la prosodie et la production de gestes ; cette règle énoncée par Kendon [94], baptisée "Règle de synchronie phonologique" par McNeill [128], se traduit par le fait que le *stroke* d'un geste co-verbal précède ou se termine au pic phonologique de la syllabe accentuée mais ne la suit jamais.

Le lien entre les gestes manuels et la prosodie n'est pas un objet d'étude très récent puisque des études réalisées dans les années trente au sein de l'Union Soviétique par Dobrogaev (et rapportées par Kendon [94]) montrent une voix sans mélodie, ni accentuations lorsqu'on empêche les participants de bouger leurs mains, tandis qu'une autre étude des années cinquante, par Birdwhistell [15] montre une corrélation entre la mélodie de la parole et les mouvements des paupières. Cependant ces deux études sont très peu détaillées sur leur mode opératoire. De manière plus récente, Kendon [94] rapporte des corrélations entre les mouvements des mains et de l'intonation au sein d'une phrase. De même, Bolinger [17] trouve une évolution parallèle des mouvements de la face (et généralise cette conclusion aux mouvements manuels) et du contour de fréquence fondamentale des productions vocales (*i.e.* lorsque la fréquence fondamentale de la parole augmente, les mouvements de la face sont dirigés vers le haut et inversement). Quelques études plus récentes (par exemple Cavé et coll. [28]) ont aussi noté des parallèles entre les mouvements des paupières et la courbe de fréquence fondamentale de la parole (en français pour l'exemple précédent), Loehr [116] trouve un regroupement des évènements gestuels et intonatifs et, plus précisément, trouve un alignement temporel entre les instants d'apex des gestes et les pics de fréquence fondamentale. D'autres chercheurs ont des interprétations plus prudentes des résultats comme McClave [123, 124] qui, bien que défendant l'idée d'une coordination entre gestes de battement et courbe de fréquence fondamentale, prône le fait que "coordonner la direction de la fréquence fondamentale et des gestes manuels est une option pour les locuteurs, mais n'est pas biologiquement fixé". Cette façon de voir les choses semble être concordante avec des travaux récents [80] en chinois qui trouvent une covariation des directions de la fréquence fondamentale, de l'intensité et des battements proche du niveau du hasard.

D'autres études se sont intéressées à la rythmicité conjointe de la parole et des gestes manuels. Cet aspect de la prosodie a surtout été mentionné en relation avec les gestes de type battement (suivant l'idée de McNeill [128] selon laquelle les gestes de battements "bougent avec la pulsation rythmique de la parole") comme ayant un rôle dans la synchronisation des deux modalités. Peu d'études ont été menées dans ces directions. Byers (rapporté par McClave [124]), suite à deux études s'intéressant aux rythmes des gestes et de la parole par rapport aux rythmes mesurés en electroencéphalographie (EEG), pense que les rythmes observables sur les signaux de parole et des gestes sont la manifestation d'un rythme biologique sous-jacent commun. De même, Tuite [177] pense que les parties pertinentes des gestes sont construites à partir de "pulsations rythmiques" internes qui se répartissent de manière régulière sur le signal de parole. Finalement, le modèle présenté par Iverson & Thelen [82] s'appuie sur des hypothèses de rythmicité pour prédire le développement conjoint de la parole et des gestes manuels. Ici encore McClave [123] reste prudente sur l'implication de la rythmicité dans la coordination gestes manuels/parole ("les gestes de battement sont organisés en des patrons rythmiques et ces patrons ne sont pas dépendants de la parole"). Plus récemment, les travaux de thèse de Hsieh [80] reprennent ces résultats et montrent des résultats similaires (il existe bien un rythme dans la production de gestes –de battement–, mais ceux-ci ne sont pas forcément cooccurrents avec les syllabes accentuées). Enfin, la plupart des travaux étudiant la coordination gestes manuels/parole du point de vue des systèmes dynamiques (voir par exemple Rusiewicz ou Rochet-Capellan [161, 157]), s'appuient le plus souvent sur des idées sous-jacentes (oscillateurs) qui mènent à des hypothèses sur des fréquences de résonance permettant de lier les deux modalités.

On voit bien par l'intermédiaire de ce panorama rapide des études s'étant intéressées de près à la coordination geste manuel/parole et à la prosodie que l'implication de la prosodie dans la coordination peut être forte (mais reste variable selon les études) et que les hypothèses quant à son rôle précis dans la coordination sont multiples (hypothèses sur la mélodie, sur le rythme) sans être toujours incompatibles. La prosodie sera un des points d'intérêt des études présentées par la suite dans ce manuscrit. Un rôle particulier de la prosodie sera particulièrement mis en lumière : son utilité dans la désignation.

1.3.4 Le cadre de la désignation

Un autre point central par la suite est la désignation ou deixis (dans son sens étymologique de monstration). Toutes les contributions de ce travail se placent dans ce contexte. Selon Kita [99, p.133], la deixis se rapporte aux caractéristiques linguistiques ou expressions linguistiques qui mettent en relation les énoncés avec les circonstances de l'espace et du temps dans lesquels ils ont lieu. Lorsqu'une discussion est engagée et contient de la deixis, celle-ci est le plus souvent accompagnée de geste(s) de pointage(s), car des actions non linguistiques sont nécessaires afin de lier réellement la deixis à des références spatiales ou temporelles. Dans le cadre de ce manuscrit, on considérera que la deixis concerne à la fois la désignation linguistique *et* la désignation qui l'accompagne (mouvement, manuel ou non). Dans tous les cas, le but de la deixis est d'attirer l'attention de l'interlocuteur sur un élément/groupement d'éléments qui devient alors centre de *l'attention partagée*. Par la suite, on utilisera indifféremment les termes désignation et deixis.

Le cadre théorique de la désignation est intéressant dans le cadre des études présentées dans ce manuscrit, en effet, la deixis est, par nature, multimodale. Il est possible de désigner à la fois vocalement et manuellement et ces deux modalités coopèrent dans leur rôle de désignation (voir Gonseth et coll. [66] qui montrent une coopération entre ces deux modalités en comparant des tâches simples –désignation en parole seule ou geste manuel seul– *vs* désignation multimodale).

1.3.4.1 La désignation gestuelle et le geste de pointage

La désignation gestuelle, réalisée par le geste de pointage (ou de pointer), est un procédé existant dans toutes les cultures afin d'établir une attention conjointe entre les interlocuteurs (voir Kita [99]). La forme du geste peut varier selon les cultures (extension des doigts –typiquement l'index–, de la main complète, orientation de la main, extension du bras, ...) mais celle-ci est toujours réalisée afin d'indiquer un axe (matérialisé par un doigt, la main complète,...). Il existe certaines cultures où la désignation ne se fait pas manuellement mais avec les lèvres par exemple (voir Enfield [49] pour de tels exemples), mais ces cas sont minoritaires et hors du cadre de ce manuscrit. Tout comme la deixis verbale, la désignation gestuelle a un rôle primordial dans l'acquisition du langage chez l'enfant (*cf.* Butcher et Goldin-Meadow [21]), le pointage manuel étant "l'outil" le plus utilisé par le bébé pour attirer l'attention de ses interlocuteurs.

Finalement, le geste de pointage sert principalement deux buts : indiquer un objet ou une direction et coordonner le focus attentionnel de plusieurs interlocuteurs.

Dans la suite, seul le geste de pointage consistant en une extension (plus ou moins complète selon les tâches) du bras et l'extension de l'index (seul) sera considéré...

1.3.4.2 La désignation verbale, son lien avec la prosodie

La fonction de monstration est bien entendu réalisable dans la modalité parole. La deixis verbale a été montrée comme étant une pierre d'angle de l'acquisition du langage chez l'enfant. Maîtrisée de manière successive à la deixis manuelle (*cf.* ci-dessus), l'apparition de son utilisation de concert avec celle-ci permet de prédire certains stades de l'acquisition (voir Section 1.1.3.3). Les processus permettant de montrer verbalement se subdivisent en deux catégories principales : la focalisation et les démonstratifs.

La focalisation La focalisation consiste à mettre en avant une partie d'un énoncé afin de la rendre saillante, de la qualifier comme véhiculant l'information essentielle contenue dans l'énoncé. Dans son ouvrage de linguistique,

Nølke [137] présente la focalisation (ou "focus" ou "foyer") comme une notion complexe et souvent mal définie (selon les cas, certains voient ce concept comme un concept syntaxique, d'autres comme un concept prosodique uniquement). Selon son interprétation, la focalisation est un phénomène sémantique qui peut se manifester de plusieurs façons (prosodique, syntaxique *cf.* Berthoud [14]). Ce phénomène se définit comme permettant "d'attirer l'attention sur un élément" et "d'attirer l'attention sur le rôle qu'il joue par rapport aux autres éléments de son contexte" (p.128-129).

La focalisation prosodique s'effectue en modifiant les indices prosodiques de la partie à focaliser d'un énoncé (production d'un contour intonatif spécifique portant sur le mot à focaliser et allongement en français par exemple) : par exemple "**Bobby** m'a apporté un ballon". La focalisation syntaxique s'obtient par l'intermédiaire d'un processus d'extraction syntaxique obtenu en faisant précéder d'un présentatif et suivre d'une conjonction de coordination (qui, que, ...) la partie à focaliser : par exemple "C'est Bobby qui m'a apporté un ballon".

La focalisation sert principalement deux buts (qui mènent à deux "types" de focalisation) : la transmission de nouvelles informations (focalisation informative) et la mise en avant de différences de deux constituants placés de manière similaire dans un énoncé (focalisation contrastive). Ces deux types de focalisation sont décrits en détail par exemple par Di Cristo [39]. Selon les travaux de la littérature, ces deux types de focalisation sont différents non seulement par leur rôle pragmatique mais également par leur contour intonatif (un pic de fréquence fondamentale sur la syllabe accentuée pour la focalisation informative et une fréquence fondamentale minimale suivie d'un pic sur la syllabe accentuée pour la focalisation contrastive). Dans les travaux présentés dans ce manuscrit, l'objet d'intérêt sera principalement la focalisation contrastive et les constituants focalisés auront donc des schémas tonaux de type "vallée-pic" *a priori* : une chute de fréquence fondamentale suivie d'une montée de celle-ci.

Un exemple de chacun de ces moyens d'expression de la focalisation est donné en Table 1.1.

	Focalisation informative	Focalisation contrastive
Contexte	Qui a tué Davey Moore ?	Est-ce le destin qui a tué Davey Moore ?
Deixis syntaxique	**C'est** Sugar **qui** l'a tué.	Non, **c'est** Sugar **qui** l'a tué
Deixis prosodique	**SUGAR** l'a tué.	Non, **SUGAR** l'a tué

TABLE 1.1 – Les différents types/moyens d'expression de la focalisation

Les démonstratifs Les démonstratifs sont des expressions linguistiques (en petit nombre), le plus souvent définies comme des mots déictiques spatiaux indiquant l'emplacement d'un objet à référencer par rapport à un centre déictique. Selon Diessel [41], outre leur fonction d'indication d'emplacement, les démonstratifs servent en particulier à coordonner l'attention conjointe entre plusieurs interlocuteurs (ils sont très liés au geste de pointage dont un des rôles principaux est de guider l'attention partagée des interlocuteurs). Tout comme pour la focalisation, les démonstratifs peuvent avoir un rôle informatif ou contrastif (selon les langues, ces deux fonctions sont distinctes formellement, voir Kuntay et Özyürek [109] pour le turc par exemple).

Dans la suite, on considérera comme démonstratif les adjectifs (ce, cet, cette, ces) et pronoms (celui, celle, ceux, celles, cela, ça, ce) démonstratifs ainsi que les adverbes de lieu (ici, là).

Il est à noter que, *a priori*, les deux processus de focalisation et les démonstratifs ne sont pas liés : les démonstratifs ne sont pas toujours "focalisés" (marqués comme prosodiquement saillants).

1.4 Aspects temporels et fonctionnels de la coordination gestes manuels/parole

Une littérature extensive traite des interactions entre gestes manuels et parole (et de la coordination entre ces deux modalités). Il a été montré précédemment que la plupart des modèles proposés dans la littérature prédisent des relations de sens entre les gestes et la parole produits en même temps. Il est évident que tous ces modèles ont

été construits afin de prendre en compte les résultats d'expériences réelles. Cette sous-section rassemble quelques articles qui semblent d'intérêt pour expliquer à la fois le caractère indéniable de ces connexions entre modalités mais également pour décrire ces relations.

Dans la suite, une courte introduction sur les études portant sur les populations pathologiques sera présentée. Ceci permet de montrer de façon claire la coordination qui existe entre les deux modalités. Par delà ces études sur des populations pathologiques, une multitude d'études à la fois comportementales et en neuroimagerie ont été menées que ce soit en production ou en perception : dans la suite, une rapide revue de bibliographie sera effectuée. Enfin, une description plus détaillée des études ayant servi de base pour la mise au point des expériences de ce travail sera effectuée.

1.4.1 Preuves d'une connexion entre les deux modalités : dysfonctionnements et pathologies

Quelques études (assez rares) sur les populations pathologiques permettent d'exposer de manière claire des connexions profondes entre deux effecteurs/fonctions/zones du cerveau/...

Une étude menée par Iverson et Goldin-Meadow [83] sur les aveugles congénitaux montre de manière assez flagrante la connexion profonde qui existe entre les gestes manuels et la parole. Après avoir comparé les productions de douze aveugles de naissance avec douze participants contrôle lors d'une tâche de narration, les auteures ont montré que le taux de gestes produits était le même dans les deux populations testées et que les gestes étaient produits à des taux similaires lorsque les participants savaient qu'ils parlaient à des personnes aveugles. Ceci montre que la production de gestes manuels fait partie intégrante du processus de production de parole (puisque les aveugles de naissance n'ont jamais vu de gestes manuels, ceci ne peut pas être dû à un apprentissage). Par ailleurs, des études ont été menées sur des populations présentant un handicap sur une des deux modalités.

Typiquement, les études menées par Hadar et ses collaborateurs [70] sur les aphasiques (de différents types : sémantique –difficulté de dénomination des objets–, conceptuel –difficulté de compréhension des phrases–, phonologique –erreurs phonologiques lors de la répétition/production de parole–) effectuant une tâche de description montrent que des caractéristiques pathologiques légèrement différentes affectent de manière différente les productions de gestes associés à la parole : les aphasiques sémantiques produisent plus de gestes iconiques que les participants contrôles, ce qui n'est pas le cas pour les autres types d'aphasie qui ont un taux de gestes équivalents aux contrôles... Des études similaires ont été menées sur des cas d'aphasie "de conduction" (*cf.* Cocks et coll. [31]) et mènent à des conclusions similaires. De manière peut-être plus démonstrative, des études sur les personnes souffrant de bégaiement (voir les études de l'équipe de Mayberry et Jaques [122, 121]) ont montré que, lors de phases de bégaiement, les gestes coverbaux "suspendent" leur exécution jusqu'au retour d'une parole fluide. Globalement, l'étude des aphasiques permet d'affirmer qu'une atteinte des capacités de parole peut avoir une influence directe sur les gestes associés. De manière similaire, mais plus inhabituelle, les productions vocales de populations apraxiques des membres ont été étudiées (voir par exemple Barrett et coll. [8]), et il a également été montré une influence des difficultés de produire des mouvements sur la fluence/pertinence verbale. De même les données cliniques rapportent des déficits de production de parole souvent associée au stade post-AVC dans les cas des apraxies des membres comme rapporté par Agnew et coll. [1].

Ces quelques études montrent assez clairement qu'il existe des liens profonds entre les gestes manuels et la parole. Plus particulièrement, les handicaps portant sur la parole mènent à des difficultés pour la réalisation de gestes manuels coverbaux, et de manière symétrique, des handicaps sur le mouvement des bras/mains ont une influence sur la parole. Les données sur les personnes aveugles sont un résultat très classique mais qui illustre de manière non équivoque la relation entre les deux modalités. Ci-dessous, un tour d'horizon des études portant sur cette relation est proposé, il est à mentionner directement le fait qu'une grande majorité de ces études décrivent les liens entre gestes manuels et parole de façon qualitative. Quelques rares études quantitatives sont décrites ultérieurement. Globalement, la plupart des études présentées s'intéresse aux gestes coverbaux et le plus souvent, aux gestes iconiques (et éventuellement au pointage). Quelques études prennent en compte les gestes de type battement et les pantomimes, ceci est mentionné si tel est le cas.

1.4.2 Coordination en production

Dans la suite du manuscrit, il sera souvent fait une distinction entre coordination des productions et synchronie des productions (bien que ces deux termes aient pu être utilisés de façon interchangeable dans la littérature). La synchronie entre deux évènements est ici considérée comme une cooccurrence assez stricte entre ces deux évènements (*i.e.* les deux évènements interviennnent de façon simultanée, avec une différence temporelle proche de zéro). La synchronie est un cas particulier de coordination temporelle : deux évènements sont coordonnés temporellement si leur différence temporelle est constante au fil des productions étudiées.

1.4.2.1 La modulation multimodale des productions gestes manuels/parole

Comme mentionné dans la présentation générale des gestes en Section 1.1.3, les rôles attribués aux gestes sont le résultat d'études menées en production de parole accompagnée de gestes. Krauss et ses collègues ont mis au point une multitude d'expériences de description de scènes au cours desquelles des hypothèses portant sur l'aide à l'accès au lexique par les gestes étaient testées. Ainsi, typiquement, dans [152], les expérimentateurs cherchent à étudier l'influence de la restriction des mouvements manuels et de la production de mots plus "complexes" sur la fluence verbale. Les résultats montrent qu'empêcher les participants de gesticuler a des effets similaires à leur imposer l'utilisation de mots plus complexes et que les participants produisent largement plus (trois fois environ) de gestes lorsqu'ils évoquent des concepts spatiaux. La plupart des résultats de Krauss allant dans ce sens sont synthétisés dans [106]. De la même façon, c'est en utilisant des tâches de description que d'autres équipes ont avancé leurs théories. Wesp et coll. [179] ont fait décrire des tableaux à des participants, le tableau étant visible ou non ; ils trouvent un taux de gestes plus important dans le cas où le tableau n'est plus visible lors de la description (résultat aussi rapporté dans Morsella et Krauss [134]), ce qu'ils interprètent comme un rôle du geste dans le maintien de concepts non lexicaux (spatiaux) en mémoire. Özyürek [140] montre dans une tâche de description de dessins animés (protocole cher à McNeill) que l'environnement et en particulier l'emplacement des interlocuteurs sont susceptibles de modifier les gestes produits pour décrire une même situation, ceci s'accompagnant de changements pertinents dans la parole. En utilisant des protocoles similaires Kita et Özyürek [100] concluent à un rôle des gestes dans l'organisation de la pensée en unités lexicalisables. Enfin certaines théories prêchent pour un allégement de la charge cognitive par les gestes : Goldin-Meadow et collaborateurs [61] ont utilisé une tâche dans laquelle les participants devaient successivement mémoriser une séquence de lettres, résoudre un problème mathématique (gestes manuels interdits ou autorisés selon la condition expérimentale) et restituer la séquence mémorisée. Les participants ayant eu la possibilité de bouger les mains ont de meilleurs taux de mémorisation, ce qui peut s'expliquer, selon les auteurs, par un rôle d'allègement de la charge cognitive lors de la production de gestes manuels.

Bien entendu, l'étude des apports de la production de gestes ne représente qu'une partie des recherches effectuées sur les gestes coverbaux : ces hypothèses sont pour la plupart non incompatibles les unes avec les autres mais sont sources de conflits passionnés entre les différents partisans de chaque théorie. D'autres études plus consensuelles s'intéressent aux caractéristiques "qualitatives" de la production de gestes associés avec la parole : ces études ne prennent partie pour aucune hypothèse spécifique quant au rôle des gestes (ou du moins, ne les mettent pas en avant) mais décrivent l'influence mutuelle de chaque modalité sur l'autre. Plusieurs études portent sur la quantité de gestes produits dans différentes situations. Ainsi, Hadar et collègues [70] montrent que lors d'une description d'objets ayant des formes complexes, la quantité de gestes produits est plus grande pour les patients aphasiques ayant des difficultés lexicales. Krauss et coll. [108] ainsi que Bavelas et coll. [11] remarquent une augmentation de la quantité de gestes produits en configuration face à face (par rapport à des conditions où les interlocuteurs ne se voient pas) et montrent une influence de la visibilité sur la forme des gestes. Jacobs et Garnham [85] montrent également un effet de l'attention des interlocuteurs (plus de gestes manuels sont produits lorsque le/les interlocuteurs sont attentifs) et le fait que le taux de gestes produits ne varie pas si on narre plusieurs fois la même histoire. De manière indépendante, de (rares) études ont étudié l'influence de chaque modalité sur le contenu de l'autre, ainsi, Wolff et Guttstein [185] ont montré dès les années soixante-dix un effet de la production de gestes sur le contenu de la parole qui l'accompagne (dans l'expérience, les chercheurs ont imposé aux participants de réaliser des gestes de type circulaire ou linéaire ; la parole associée étant ensuite

annotée par des juges qui arrivent à retrouver le type de geste effectué simplement à partir du son). Au niveau du signal, il a été montré l'influence dans les deux modalités de la congruence des gestes avec la parole qu'ils accompagnent (*i.e.* le fait qu'ils fassent référence à un même concept) : pour les gestes de pointage, la congruence a une influence sur la dynamique des gestes –plus rapide en condition congruente, moins rapide en condition incongruente– et sur le spectre de la voix selon Chieffi, Secchi et Gentilucci [29] $-F_2$ plus faible en condition incongruente. Enfin, une étude sur les gestes symboliques de Barbieri et coll. [7] montre un faible effet de la congruence sur le spectre de la parole $-F_2$ augmente en condition congruente.

Ici encore, les arguments sont assez convaincants en ce qui concerne une réelle interaction entre les gestes et la parole, plutôt que deux systèmes évoluant indépendamment l'un de l'autre. Il existe d'après la littérature une véritable coordination entre les deux modalités : l'intégration de la vision/écoute d'une scène impliquant des gestes manuels et de la parole semble être du ressort d'aires impliquées dans l'intégration multimodale au niveau cortical. Comme démontré dans une étude en EEG de Habets et coll. présentée ultérieurement [68], une certaine concordance temporelle est requise entre les deux modes afin que l'intégration puisse se faire correctement. Une étude précise des relations temporelles entre gestes manuels et parole permettrait de se prononcer sur les possibles évènements de la parole et des gestes qui sont "intégrés" ensemble... Les études portant sur des mesures temporelles précises sont très peu nombreuses, et seront pour la plupart détaillées dans la section suivante. La règle de synchronie phonologique de McNeill semble être une règle respectée dans la plupart des études (le *stroke* est toujours réalisé en même temps ou avant la syllabe ou le mot accentué –mais jamais après). Cependant, cette relation temporelle forte est parfois réfutée, ainsi de Ruiter et Wilkins [159] invalident de façon catégorique cette relation en Arrernte (langage parlé par le peuple du même nom en Australie), les gestes étant produits *après* leur affilié lexical. La plupart des autres études portent sur l'anglais et valident cette règle, cependant la "synchronie" semble assez fragile... Par exemple, concernant les gestes de battement, McClave [123] ne trouve pas de synchronie entre le *stroke* de ces gestes et la syllabe accentuée ni aucune régularité dans les durées temporelles séparant ces deux évènements (et de conclure que l'instant d'occurrence des battements est imprévisible à partir de la position des accents). Nobe [136], quant à lui, montre clairement que les strokes de ces gestes interviennent toujours avant ou avec les pics de fréquence fondamentale/d'intensité de la parole (le même résultat pour le pic de f_0 est trouvé par Leonard & Cummins [113] : de manière plus précise, c'est le pic de vitesse du stroke qui est aligné avec le pic de f_0). D'autres études montrent que cette règle de synchronie est fragile et sujette à des ajustements : typiquement Morrel-Samuels et Krauss [133] indiquaient une relation de proportionnalité entre le laps de temps séparant les onsets gestuel et verbal et l'inverse de la familiarité d'un mot pour un participant, indiquant par le même biais la forte variabilité dans les données d'alignement temporel.

De manière un peu plus précise, quelques études décrites ci-dessous ont mis au point des protocoles permettant l'étude précise des relations temporelles entre la production de gestes manuels et de parole. Dès les années 80, des études mesurant précisément temporellement les différentes phases des gestes ainsi que les évènements de la parole ont vu le jour...

1.4.2.2 Production conjointe de parole et de gestes : les mécanismes sous-tendant la production

Daniel Holender et son équipe [73] ont mené en 1980 des travaux préliminaires sur la réalisation d'une double tâche lors de la présentation d'un stimulus. Cette étude reprend quelques travaux de Theios (1973) qui proposait des modèles de traitement des informations après perception de stimuli.

Dans cette étude, Holender présente aux participants des lettres sur un écran. Les participants doivent soit appuyer sur un bouton correspondant à la lettre présentée, soit nommer –en français– la lettre présentée, soit effectuer simultanément ces deux tâches. Les mesures –mesures temporelles d'*onset* vocal et gestuel– permettent de montrer que dans les tâches simples la réponse vocale est plus rapide que la réponse gestuelle et que dans la tâche double, le temps de réponse vocal est plus long que le temps de réponse gestuel (et ce, d'un laps de temps constant). L'équipe conclut que cet effet de détérioration du temps de réponse vocal soutient l'hypothèse d'une interférence dans les traitements internes des stimuli responsables de la production de gestes manuels et de parole.

Les résultats sont en accord avec les théories soutenant l'utilisation compétitive de capacités communes pour

le traitement des informations : les modalités doivent s'adapter à la limitation des capacités de traitement des informations.

L'étude de Holender ne s'inscrit pas totalement dans le cadre posé (plus tard) par McNeill, où la coordination est une coordination non pas entre organes moteurs mais entre les contenus dans les deux modalités : dans l'étude de Holender, le mouvement manuel d'appui sur un bouton n'a aucun lien communicatif avec la parole produite (il s'apparente –à la limite– plus à un geste rythmique…). Cependant, comme cela a été présenté précédemment, c'est assez naturellement que des études ont émergé en ce qui concerne l'étude de la coordination temporelle entre le geste de pointage et la parole associée et plus particulièrement, dans le cadre de la désignation, du fait du lien particulier qui unit les deux modalités dans ce cadre, comme présenté en Section 1.3.4. Un des travaux considéré comme pierre d'angle dans la qualification des interactions geste bracchio-manuel / parole a été réalisé par l'équipe néerlandaise de William "Pim" Levelt dont les expériences de 1985 [114] ont permis de poser des bases solides pour les études suivantes. Dans son article de 1985, Levelt compare deux manières de modéliser la planification et l'exécution motrice. Une vision dite "interactive" qui considère que les paramètres temporels du geste et de la parole sont liés de manière permanente car les deux systèmes communiquent entre eux au cours de la préparation et de l'exécution motrice. Une autre vision, dite "balistique", considère au contraire que les deux systèmes évoluent de manières totalement indépendantes durant la phase d'exécution, les paramètres temporels de cette phase ayant été préetablis durant la phase de préparation – durant laquelle les deux systèmes communiquent –. Les conclusions vont dans le sens d'une vision hybride. Dans les expériences menées ici, les participants sont assis devant un bouton poussoir. Après avoir appuyé sur le bouton, une des quatre lampes situées dans la salle s'allume, les participants doivent alors désigner soit vocalement –en hollandais : *dat lampje*/cette lampe–, soit gestuellement par un geste de pointage, soit vocalement et gestuellement, la lampe qui s'allume. Les mesures réalisées ici sont axées sur des mesures temporelles d'*onset* et d'apex du geste de pointage et sur les instants de la parole.

Dans une première expérience, Levelt et ses collègues ont étudié l'effet de la distance sur la synchronisation voix/geste. Ils trouvent un effet significatif de la distance sur cette synchronisation, l'apex du geste étant précédé par la parole par une durée plus longue lorsque la cible visuelle se trouve plus loin. L'étude de l'effet de la distance leur permet d'affirmer que les temps de début de geste et de début de prononciation de la phrase covarient de manière significative. Une seconde expérience cherche dans quel "sens" se fait l'adaptation des productions. Les mesures montrent que le temps nécessaire pour atteindre l'apex du geste n'est pas affecté par la production de parole alors que le temps avant l'initiation du geste l'est. Pour ce qui est de la parole, l'onset de la parole est assez largement retardé par la présence de gestes manuels. Ainsi, selon les auteurs, le système de production de la parole s'adapte au système de production du geste, la seule influence de la parole sur le geste intervenant, par ailleurs, dans la phase de préparation du geste. Ceci conforte selon les auteurs une hypothèse selon laquelle la planification des productions dans les deux modalités serait interactive (et effectuée par un *pool* de ressources communes) puis l'exécution serait modulaire. L'influence de l'augmentation du nombre de cibles possibles (à la fois le nombre de lampes et le nombre de mots-cibles pour les désigner) est étudié dans une troisième expérience. Les résultats montrent que les instants de début du geste de pointage sont influencés par le nombre d'alternatives possibles à la fois dans l'espace et en parole *mais la durée séparant l'apex et l'onset ne varie pas.* En parole, l'onset n'est influencé que par le nombre de possibilités dans l'espace et par la production concurrente de gestes. Ainsi une théorie balistique semble probable pour l'exécution de gestes (mais pas pour la parole). Enfin, une dernière expérience essaie de mesurer le temps pendant lequel l'exécution des gestes manuels semble influer sur la planification de la parole. Ceci est réalisé en perturbant le mouvement de pointage juste après son onset. Les résultats semblent montrer que la planification de la parole n'est plus influencée par les modifications de l'exécution du geste si celles-ci ont lieu moins de 350 ms (en fait, un temps inconnu entre 300 et 370 ms) avant l'onset de la parole. Les conclusions générales de l'article recoupent en partie les conclusions trouvées dans le cadre du geste d'appui sur un bouton, avec quelques spécificités du geste de pointage. Les chercheurs notent que, comme dans le cas de la tâche presse-bouton, la parole s'adapte au geste mais le geste n'est que marginalement affecté par la parole. Par ailleurs, les deux systèmes semblent se comporter comme concurrents pour l'accès à une quantité de ressources limitées nécessaire au traitement des données dans les instants précédant le geste. Enfin, les deux systèmes, sont en relation jusqu'au départ du geste. Une fois le geste lancé, celui-ci a un comportement

balistique. Il en est de même pour la parole qui opère balistiquement à partir de 300 à 370ms avant son début.

L'équipe de Feyereisen a conduit une étude en 1997 [52] dont le but avoué était d'apporter quelques précisions sur les travaux de Levelt et coll. (interactions/interférences dans la planification, généralisation à d'autres types de gestes). Feyereisen propose à ses participants, assis devant un écran d'ordinateur, de désigner (vocalement, en français belge, et/ou gestuellement) un signal apparaissant sur cet écran. L'équipe mesure les temps d'*onset* des gestes et de la voix.

Deux expériences permettent de retrouver les résultats de Levelt pour le geste de pointage (tâche : désigner une croix sur l'écran par un geste de pointage, en prononçant « ti » ou « ta ») et d'étendre ces résultats aux gestes iconiques (tâche : représenter les formes \, ◇, #, et ○ par des gestes iconiques représentant ces formes en prononçant « barre », « blanc », « bloc », « boule »). En utilisant une mise en place différente de Levelt et coll. –utilisation de l'amorçage, donnant des indices sur la réponse à produire avant que le stimulus apparaisse–, Feyereisen reproduit par ailleurs les résultats de l'équipe hollandaise concernant le lieu d'interférence dans la planification : il semble que les deux systèmes interagissent jusqu'au point où les mouvements (manuels ou de parole) sont initiés puis le mouvement est balistique. Une dernière expérience permet de voir quel modèle d'élaboration des réponses multimodales est le plus probable (selon l'un, le traitement des données gestuelles et de parole se fait par deux systèmes distincts qui interagissent entre eux, selon un autre modèle, un seul système traite les deux modalités). Ceci est réalisé en imposant une charge cognitive forte sur la sélection à effectuer en parole (ou en geste). Cette expérience ne parvient pas, selon les auteurs, à démontrer une dissociation entre les systèmes de sélection de réponse de chaque modalité.

Cette étude a permis de généraliser les travaux de Levelt à des cas plus complexes et dans un paradigme expérimental différent et va dans le sens d'un traitement interactif des stimuli dans les deux modalités comme le montre la dernière expérience. Les études présentées jusqu'ici se sont surtout intéressées aux mécanismes sous-jacents à la sélection de réponses multimodales. Les études suivantes s'intéressent principalement aux caractéristiques temporelles des productions dans les deux modalités. Il est évident que ces deux types d'études sont importantes les unes pour les autres puisque seule leur combinaison permet de bien comprendre le fonctionnement de concert des deux modalités.

1.4.2.3 Relations temporelles dans la production de gestes et de parole

Dans le cadre de sa thèse [38], de Ruiter s'est intéressé à l'organisation temporelle du geste de pointage en relation avec la parole cooccurrente. Dans les deux expériences portant sur la coordination temporelle parole/gestes manuels, les mouvements sont suivis grâce à un capteur ultrasons qui capte les émissions d'un *buzzer* fixé sur l'index droit des participants.

Dans une première expérience, de Ruiter fait asseoir des participants devant quatre diodes surplombées par des images installées sur une table (la Figure 1.8 représente son dispositif expérimental, dispositif proche de ceux utilisés dans les autres travaux). La tâche consiste pour les participants, à pointer vers l'image sous laquelle la diode s'allume tout en nommant cette image (par exemple, pointer vers une image d'appareil photo et dire *de camera* –l'appareil photo en hollandais). Les résultats montrent que la parole s'adapte au geste –*i.e.* si le geste commence plus tard, la parole aussi. Au niveau de la cooccurrence temporelle, l'auteur montre une cooccurrence assez fine (et régulière) entre l'apex du geste et le début du nom de l'objet désigné (59 ms d'écart en moyenne) mais aucune influence de la position de l'accentuation lexicale (les mots sont accentués sur le début *vs* la fin) sur les instants de production des gestes.

Dans une seconde étude, de Ruiter s'intéresse à l'influence de la position de la focalisation contrastive sur la production de gestes associée. La tâche est globalement similaire mais un mot de la phrase prononcé contient de la focalisation prosodique contrastive (par exemple : *de groen krokodil* –le crocodile vert– focalisé soit sur *groen* soit sur *krokodil*). L'auteur trouve un résultat intéressant : la position de la focalisation au sein de l'énoncé modifie l'instant de début de geste *i.e.* si la focalisation porte sur le début de phrase, le geste commence plus tôt et se termine plus tôt. Contrairement aux résultats précédents, la synchronisation entre l'apex et la syllabe accentuée du discours n'est pas observée –l'auteur pense que ceci est dû à un dispositif trop contrôlé– mais une corrélation importante ($r = 0,61$) est observée entre le début de l'apex et l'onset de la syllabe accentuée.

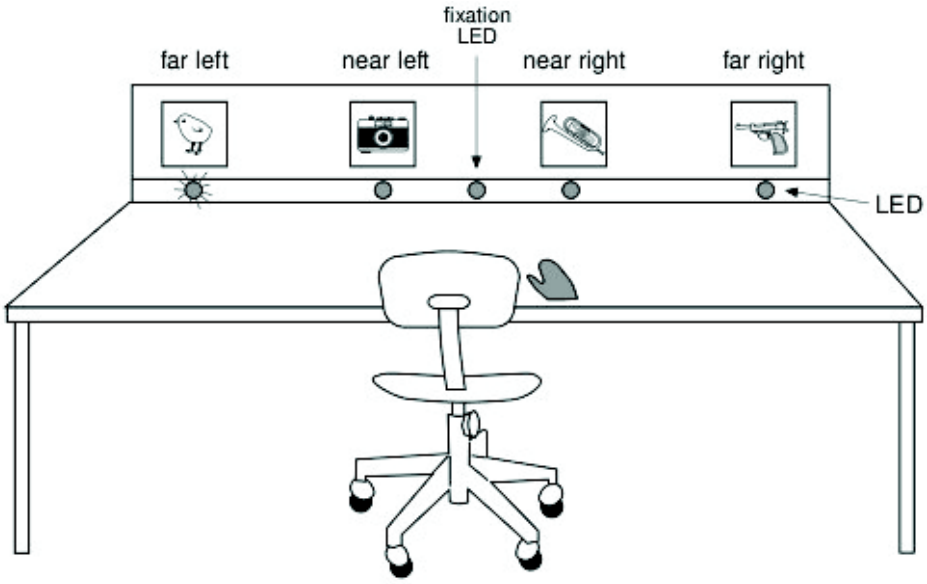


FIGURE 1.8 – Dispositif expérimental des études de de Ruiter (d'après [38])

Ce dernier résultat est important pour les expériences présentées dans ce manuscrit : l'apex du geste de pointage (dans la seconde expérience de de Ruiter) covarie avec l'onset de la syllabe accentuée (*i.e.* portant un pic de fréquence fondamentale) mais les deux évènements ne sont pas synchrones.

Dans son manuscrit de thèse, Dan Loehr [116] étudie de manière plus précise la coordination entre l'intonation (en particulier, les variations de fréquence fondamentale) et la production de gestes. L'étude menée par Loehr s'appuie sur des enregistrements audio/vidéo de diades/triades d'amis discutant. Les données temporelles sont les principales données d'intérêt pour Loehr et celles-ci sont extraites de l'annotation manuelle conjointe des signaux acoustiques et vidéo. Au niveau acoustique, l'annotation suit les recommandations du le système d'annotation ToBI (en particulier les "tons" regroupent accents de *pitch*, accents de syntagme et tons de frontière). Au niveau gestuel, une annotation des mouvements des mains et de la tête est effectuée sur la video.

Les résultats s'intéressant à la proximité temporelle de deux évènements, Loehr s'est heurté à la définition de cette notion (*i.e.* quand peut-on dire que deux évènements geste/parole sont *proches* temporellement ?). Pour chaque évènement temporel gestuel annoté, l'auteur a calculé le laps de temps le séparant du ton le plus proche. L'auteur considère finalement que deux évènements sont proches s'ils sont séparés par une durée inférieure à un écart-type de la distribution constituée par toutes ces différences temporelles (soit 275 ms).

Son étude propose se bases sur les ouvertures proposées par Yerian [186] dans sa dissertation sur une revue de l'état de l'art des études portant sur l'intonation et les gestes manuels. Les résultats sont présentés en cinq points : ① l'hypothèse de Bolinger (*i.e.* les mouvements du corps suivent les mouvements –verticaux– de la courbe de fréquence fondamentale de la voix) n'est pas vérifiée (ni pour les mains, ni pour la tête) ; ② les apex des gestes s'alignent temporellement avec un accent de fréquence fondamentale (moyenne par participant < 100 ms) ; ③ pas de relation entre les types de gestes utilisés (iconiques, emblèmes, déictiques ...) et les types d'unités intonatives qui leur sont associées (focalisation, accent, ...) dans les "alignements" temporels ; ④ au niveau pragmatique, l'intonation de la parole et les gestes semblent corrélés (cette corrélation n'est pas tout le temps vraie, mais intervient de manière claire parfois, par exemple pour signifier l'emphase, le contraste) ; ⑤ partant d'une paraphrase de Duncan selon laquelle "tout est battement", Loehr étudie l'hypothèse de rythmes corporels (initialement intonation/gestes puis s'étendant rapidement aux mouvements de la tête, des yeux...) permettant d'étayer une synchronisation d'évènements. Selon lui, les différents "instruments" du corps (les mains, les articulateurs, ...) sont liés rythmiquement mais leur relation n'est pas uniforme dans le temps (typiquement, chaque modalité semble suivre une organisation rythmique propre *mais* ces organisations sont liées entre elles : les frontières des évènements gestuels et les frontières dans le flux de parole s'attirent, bien que les évènements

n'aient pas forcément les mêmes rythmes dans les deux modalités) ; par ailleurs chaque instrument semble avoir un "tempo" qui lui est propre.

Il est à noter que les résultats concernant la cooccurrence des apex gestuels avec les pics de fréquence fondamentale est un résultat qui a été confirmé par l'étude vidéo lors de débats publics réalisée par Jannedy et Mendoza-Denton en 2005 [86] qui trouvent que lorsqu'il y a un apex gestuel, celui-ci est systématiquement associé à un pic de fréquence fondamentale (mais la relation inverse ne tient pas : il y a plus d'accents prosodiques que d'apex de gestes).

Finalement, les résultats de Loehr corroborrent les conclusions de Kendon et McClave et complètent les points avancés par McNeill en ajoutant le concept d'intonation (déjà abordé par McClave) aux études sur la coordination geste/parole. Par le lien (temporel, sémantique, pragmatique) montré entre intonation et gestualité, Loehr affirme que, l'intonation étant liée à l'intention communicative, le geste est communicatif. Par ailleurs, cette étude met en exergue l'existence de rythmes dans la parole/les gestes en situation écologique de discussion informelle.

Dans une étude menée par A. Rochet-Capellan [156], la tâche utilisée imposait aux participants –de langue maternelle portugaise brésilienne– de pointer un personnage sur un écran lors de la prononciation du nom de ce personnage ou de réaliser l'une des deux réponses indépendamment. Le mot associé au dessin comprenait deux syllabes (par exemple : *"papa"*), chacune d'elle pouvant être la cible d'une accentuation ou non. Cette étude est la seule des études mentionnées ci-dessus utilisant un système de capture de mouvement précis (Optotrak). Ceci a permis d'ajouter un paramètre d'intérêt dans les mesures : le mouvement de la mâchoire, qui est donc la mesure d'un mouvement articulatoire, ce qui est nouveau par rapport aux études précédentes qui ne prenaient en compte que les indices acoustiques et temporels. Par ailleurs, l'utilisation d'un tel système a permis d'obtenir des mesures précises à la fois temporelles et d'amplitude quant aux déplacements de la main et de la mâchoire. Les résultats montrent que le lancement (resp. la fin) du geste est toujours effectué avant (resp. après) le début (resp. la fin) du mouvement de la mâchoire associé à la syllabe accentuée. Par ailleurs, l'apex du geste est en général attiré par la syllabe accentuée. Plus précisément, l'apex du geste de pointé est aligné avec le moment où la mâchoire atteint sa cible (point le plus bas dans sa trajectoire) si l'accent est sur la première syllabe, l'apex du geste est synchrone avec le point de début d'articulation de la seconde syllabe (point le plus haut de la mâchoire) dans le cas d'une accentuation de la seconde syllabe. Ainsi, dans le cadre de cette expérience, les systèmes ont l'air de s'adapter afin de permettre à la partie accentuée d'être exécutée en même temps que la tenue du geste de pointé (il n'y a pas de synchronisation stricte entre l'apex du geste et la prononciation de la partie accentuée). En particulier, le système de production de la parole est présenté, comme dans l'étude de Levelt et coll., comme s'adaptant au timing du système gestuel, permettant alors à celui-ci d'avoir un comportement – de l'initiation du mouvement jusqu'à l'apex du geste – indépendant de la partie du mot accentuée. La chercheuse pose plusieurs idées d'études à mener par la suite, elle met en particulier en avant le fait que les tâches utilisées dans beaucoup d'études (présentées ci-dessus) ne sont pas très "naturelles" (ne permettent pas une vraie focalisation, utilisent parfois des non-mots, ne permettent pas l'étude de phrases complètes). Le travail mené par Laurent [111] a repris cette expérience en utilisant une tâche un peu plus "naturelle" (en utilisant une phrase utilisée couramment). Les résultats trouvés sont globalement similaires : ① le mouvement de la main commence toujours avant la syllabe accentuée ② le mouvement de la main se termine toujours après la prononciation de la syllabe accentuée.

Ces études montrent qu'il semble exister un lien fort entre la main et la mâchoire lors du pointage. Elles fournissent par ailleurs des éléments pour une coordination se faisant plutôt entre le geste bracchio-manuel et le geste articulatoire qu'entre le geste bracchio-manuel et le son produit.

Enfin, le mémoire de licence écrit par Wilmes [183], supervisée par Ebert & Rieser, étudie le marquage de la focalisation par les gestes manuels. Son étude vidéo s'appuie sur une interaction du corpus SaGA [117] – comportant les vidéos de 25 diades en interaction s'exprimant à propos de concepts spatiaux (indications de directions, description de scènes) et contenant environ 5000 gestes (majoritairement iconiques). Sur la partie étudiée (20 minutes d'une interaction entre deux personnes), Wilmes a annoté les pics de fréquence fondamentale et les parties focalisées (les gestes étant déjà annotés sur le corpus initial). Les parties sont jugées comme focalisées selon si la structure informative du discours : si le contenu apporte une information nouvelle au discours ou s'il contraste avec une information précédemment donnée.

Les résultats de son analyse montrent que le stroke (apex) des gestes "est aligné" avec les pics de fréquence

fondamentale principaux (en moyenne, les apex précèdent les pics de F_0 de 362 ms ($\sigma = 553$ ms) et la fin des strokes suit la fin des "pics" de F_0 de 529 ms ($\sigma = 1247$ ms). Par ailleurs, dans le cas des focalisations informatives, le début des gestes (onset) précède la partie de la parole focalisée de 300 ms. Dans le cas de la focalisation contrastive, c'est l'apex (stroke) des gestes qui est le mieux aligné avec le début de la focalisation (9 ms d'écart en moyenne).

Selon l'auteure, les conclusions sont en désaccord avec les résultats de Loehr en ce sens que les alignements principaux entre geste et parole ne sont pas trouvés entre les instants annotés sur le geste et les indices prosodiques de la focalisation (pics de fréquence fondamentale) mais plutôt entre le geste et l'onset des parties focalisées du discours. Cependant, il ne faut pas perdre de vue que le but de ce travail est moins général que celui de Loehr et les résultats plus "génériques" de Loehr ne sont *a priori* pas en désaccord avec les résultats de l'étude de Wilmes.

Un point intéressant ici est la définition de l'alignement temporel entre deux évènements (selon Loehr, deux évènements étaient considérés comme alignés s'ils étaient produits à une distance temporelle inférieure à un écart-type de la distribution de toutes les distances temporelles considérées) : selon Wilmes, deux évènements peuvent être considérés comme alignés si la variance de la distribution constituée par les différences temporelles est faible, et ce, même si la moyenne de cette distribution n'est pas nulle. Cette dualité dans la qualification des alignements (proximité temporelle / régularité de la proximité) sera reprise dans la suite du manuscrit.

Les études présentées précédemment ne mettaient en avant qu'une partie des interactions qu'on peut trouver entre les deux systèmes de production du geste et de la parole. Ainsi, outre le fait que les deux systèmes semblent être liés de manière temporelle (synchronisation/alignement d'évènements), certaines études ont été menées afin de montrer qu'il existait également une co-influence des systèmes au niveau de l'amplitude des gestes articulatoires produits. C'est ainsi qu'en particulier, une équipe hollandaise s'est récemment attelée à l'étude du geste de battement. Ce type d'approche d'étude des influences d'un système sur l'autre au niveau de l'amplitude est assez nouveau dans les interactions geste/parole.

1.4.2.4 Le geste de battement, les mesures en amplitude

L'étude publiée en 2007 par Emiel Krahmer et Marc Swerts [103] s'intéresse au cas particulier des gestes de battement –gestes rapides montants/descendants de parties du corps– (battement de la main, de la tête ou encore des sourcils) dans leur relation avec le discours et plus particulièrement avec l'emphase. Dans cette étude en langue hollandaise, les mesures d'amplitudes réalisées sont des mesures acoustiques. La première expérience de cet article étudie l'influence des gestes de battement sur les paramètres acoustiques du discours produit simultanément. Les participants ont pour consigne d'effectuer un geste de battement en parallèle d'un mot précis (focalisé ou non) : il y a des conditions congruentes (mot focalisé + battement) et des conditions non congruentes (mot non focalisé + battement) – par exemple *Amanda gaat naar* MALTA (Amanda va à Malte) avec un battement cooccurrent avec "Malta" est congruent, AMANDA *gaat naar Malta* avec un battement cooccurrent avec "Malta" est incongruent. En plus de confirmer le fait que la focalisation (vocale) entraîne un allongement du son focalisé, une énergie plus forte sur le mot focalisé, et une augmentation de la fréquence fondamentale, les chercheurs trouvent des effets significatifs des gestes de battement. En particulier, des effets sont trouvés sur l'allongement du mot, sur l'augmentation du second formant et une tendance est observée sur l'augmentation du troisième formant, une chose particulièrement intéressante étant le fait que l'effet observé sur l'allongement et la modification sur le second formant sont semblables aux effets observés pour l'accentuation vocale. Deux autres expériences permettent de mieux comprendre l'effet perceptif apporté par les gestes de battement. Au final, trois points sont avancés comme points importants dans l'article. ① la production parallèle d'un geste —n'importe lequel des 3 gestes précités— et d'un mot augmente la durée du mot et on peut observer un effet sur le deuxième formant et une tendance vers un effet sur le troisième formant —le deuxième formant subit une réduction lors de l'accentuation en hollandais et en anglais— du mot émis : on observe un renforcement acoustique de la focalisation –augmentation de la proéminence du mot– ② la perception d'un battement augmente la proéminence perçue du mot et réduit la proéminence perçue des mots l'entourant. Les deux chercheurs concluent leur article en avançant qu'il semble que l'activité musculaire requise dans l'exécution de gestes de type battement apporte une activité

musculaire accrue pour l'articulation (ce qui est cohérent avec les théories sur le mouvement des articulateurs de Rizzolatti et Arbib [154] qui insistaient sur le fait qu'il était possible que le contrôle moteur des bras et des articulateurs vocaux soient gouvernés par un mécanisme unique).

Cette section a montré de manière assez claire qu'il existait des liens entre la production de gestes manuels et la production de parole cooccurrente. En particulier, les études se sont intéressées aux liens temporels et en amplitude des deux modalités mais également aux contenus "sémantiques" exprimés dans les deux modalités. Les mécanismes de production et de perception ont souvent été considérés comme étant proches l'un de l'autre. La perception des gestes manuels coverbaux n'échappe pas à cette constatation : quelques études présentées ci-après dénotent, tout comme en production, des liens assez forts entre perception de parole et perception de gestes manuels associés à cette parole.

1.4.3 Le lien geste manuel/parole en perception

Comme le constatent Kelly et Özyürek [89], il existe assez peu d'études qui ont été menées sur les liens gestes manuels/parole en perception. Une bonne partie des études comportementales en perception s'intéressent à la valeur communicative des gestes. L'intérêt des gestes manuels dans la perception d'un message multimodal semble résider principalement dans les faits que : les gestes accompagnant des énoncés contenant des informations spatiales/motrices améliorent la perception de descriptions (voir Driskell et Radtke [45]), que les gestes manuels apportent de l'information qui n'est pas contenue dans la parole (gestes "supplémentaires", *mismatches*) (*cf.* les études de Chu et Goldin-Meadow ainsi que McNeill [19, 128]), ou bien que ceux-ci fournissent des indices supplémentaires pour aider à la compréhension dans des conditions de communication difficiles (par exemple, en LV2, les gestes, et en particulier, les gestes manuels, améliorent la compréhension selon Sueyoshi et Hardison [170]). L'amélioration de la communication peut également être le résultat d'une meilleure fluidité de la production vocale du locuteur (comme évoqué en 1.4.3 dans Krauss, Chen et Gottesman [104]) ou en menant à des interactions où les participants se sentent plus "impliqués" (voir Maricchiolo et coll. [118]). Enfin, un dernier point soulevé par quelques rares études est l'intérêt possible de la perception de gestes coverbaux sur la mémoire à long terme (en particulier, chez les enfants *cf.* Church et coll. [30]).

Dans sa revue de bibliographie, Hostetter [76] tente, par le biais d'une méta-analyse d'articles, de comparer et de synthétiser les résultats selon neuf axes (seule une partie de ces axes est ici présentée), la question englobant ces neuf axes étant « *quand* les gestes communiquent-ils ? ». La méta-analyse regroupe 63 études portant sur la perception de la parole et des gestes coverbaux. Les résultats de l'étude sont synthétisés ci-dessous :

- *Un message de parole accompagné de gestes manuels est-il mieux compris ?* La méta-analyse montre un effet (de taille moyenne) de la production de gestes sur la compréhension. Ceci est en particulier vrai pour les gestes iconiques qui véhiculent une sémantique propre comme soutenu par Beatie et Shovelton [12].
- *Les gestes sont-ils plus communicatifs lorsqu'ils accompagnent de la parole à propos de concepts spatiaux/moteurs ?* Ici encore, il apparaît que les gestes sont surtout efficaces lorsqu'ils accompagnent de la parole portant sur des concepts spatiaux/moteurs plutôt qu'abstraits. Ceci indique donc une scission entre les gestes métaphoriques et les gestes iconiques dans leur rôle dans la communication. Ce résultat est en particulier illustré dans l'article de Feyereisen [53] qui a comparé le bénéfice dans la mémorisation des gestes représentatifs et non-représentatifs et celui de So et coll. [168] qui ont étudié la différence de performance de mémorisation lors de la production de gestes de battement et de gestes iconiques : le type de geste influe sur la bonne réception/rémanence du message chez l'interlocuteur.
- *Les gestes sont-ils plus communicatifs lorsqu'ils sont supplémentaires à la parole ?* Le résultat est assez clair, les gestes supplémentaires sont clairement plus "communicatifs" que les gestes exprimant les mêmes idées qu'en parole. En particulier, les études menées en EEG par Holle et Gunter [74] donnent une interprétation possible de ce résultat selon laquelle le rôle des gestes non congruents serait l'aide à la compréhension immédiate alors que les gestes redondants aideraient à la mémorisation…
- *Est-ce que les gestes sont plus "efficaces" pour les auditeurs qui ne maîtrisent pas la modalité verbale ?* L'étude des situations d'enseignement ou de communication avec les enfants montre clairement un impact beaucoup plus important de l'utilisation des gestes manuels coverbaux lorsque la parole n'est pas maîtrisée

par la personne à qui s'adresse la parole. En particulier, les gestes semblent aider à retenir l'attention des enfants et aider à la mémorisation comme montré par Hostetter et Alibali [77].

- *Est-ce que l'amélioration de la communication est un effet de bord d'une meilleure fluence verbale ?* Les études ne semblent pas corroborer cette vision : en production (*cf.* ci-dessous), les gestes sont considérés comme aidant à la fluence verbale par plusieurs biais (amorçage multimodal facilitant l'accès au lexique selon Krauss [106], organisation de l'information selon Kita et Özyürek [101], maintien en mémoire de concepts selon Wesp et coll. [179] ou encore allégement cognitif selon Goldin-Meadow [62]), mais cette meilleure fluence verbale ne semble pas bénéficier spécifiquement aux interlocuteurs : les gestes améliorent la compréhension de manière indépendante du fait qu'ils facilitent la fluence verbale.
- *Quel est le rôle des gestes lorsqu'ils n'apportent que peu d'information (contexte non spatial/moteur) ?* Il se peut que les gestes jouent un rôle dans la capture de l'attention (un interlocuteur est possiblement plus enclin à écouter une personne qui gesticule qu'une personne immobile) mais aucune étude ne porte sur ce point de vue, c'est donc un axe de recherche à compléter.
- *Les gestes manuels ont-ils un rôle facilitateur en plus des gestes faciaux ?* Quelques études comme celles de Sueyoshi et Hardison [170] montrent un rôle facilitateur supplémentaire des gestes manuel dans la compréhension par delà l'effet facilitateur de la vision de la face de l'interlocuteur.

Cette étude synthétise globalement les résultats trouvés en perception de la parole dans les études comportementales.

1.4.4 Coordinations corticales : les preuves d'une coordination en imagerie cérebrale

Beaucoup d'autres études ont été menées plus récemment pour étudier la coordination (surtout pour la perception) en imagerie cérebrale. Les principales technologies utilisées dans les études mentionnées ci-dessous sont l'imagerie par résonnance magnétique fonctionnelle (IRMf), l'EEG, la stimulation magnétique transcrânienne (transcranial magnetic stimulation) (TMS), la magnétoencéphalographie (MEG). Quelques-uns des résultats sont présentés ci-dessous.

Agnes Floel et son équipe [54] ont montré, en utilisant un protocole réunissant TMS et EMG une activation du cortex moteur (orofacial et) manuel lors de tâches de production/perception linguistiques : les gestes manuels ont également leur place dans le traitement des informations linguistiques. De façon légèrement différente, Pulvermüller et collègues [148] ont montré en 2005 l'interaction entre les systèmes du langage et de l'action lors d'une tâche de décision lexicale. L'étude en TMS a montré en particulier que la stimulation des aires du cortex moteur gauche chez des droitiers (donc, majoritairement latéralisés à gauche pour le langage) liées à la main (resp. au pied) mènent à des décisions linguistiques plus rapides lorsque le stimulus porte sur une action réalisée avec la main (resp. le pied). Ces travaux sont repris dans le travail de synthèse réalisé par Willems & Hagoort en 2007 [181] : le cortex moteur est activé lors de la perception de parole et le cortex prémoteur est activé lors de la perception de mots liés à l'action. Tout cela va dans le sens d'une cognition dans laquelle le traitement du langage serait ancré dans l'action corporelle.

Ces résultats sont valables de manière générale mais mettent en avant principalement un lien entre action et langage dans le cerveau. Une question d'intérêt dans le cadre de ce manuscrit est la "coopération" entre les traitements des percepts de gestes manuels et les traitements des percepts de parole dans le cerveau. Beaucoup d'études se sont intéressées aux traitements sémantiques qui sont associés aux gestes manuels. En particulier, des études en EEG ont permis d'affirmer qu'un traitement semblable au traitement sémantique de la parole existait lors de la perception des gestes manuels. C'est le cas de l'étude menée par Özyürek, Willems, Kita et Hagoort en 2007 [142] où les participants perçoivent des énoncés de parole accompagnée d'un geste iconique lié (ou non) au verbe de la phrase. Les auteurs comparent quatre conditions dans lesquelles le geste et la parole sont congruents (le geste représente le verbe qui s'intègre bien dans le contexte de la phrase), le geste s'intègre bien dans le contexte de la parole mais pas le verbe, le verbe s'intègre bien dans la parole mais pas le geste qui l'accompagne ou ni le verbe ni le geste ne s'intègrent bien dans la parole mais leurs "sémantiques" sont identiques. Les résultats montrent que toutes les discordances (parole seule, geste seul, parole et geste) mènent à des modulations de l'onde N400 similaires (à la fois au niveau de l'amplitude et de la latence de l'onde) : la

nature de l'intégration sémantique des deux modalités est très similaire et cette intégration a lieu en parallèle, probablement en zone frontale inférieure d'après l'estimation des sources. L'étude IRM menée par la même équipe [182] (*cf.* Figure 1.9) propose une intégration multimodale dans le sulcus frontal inférieur (et plus précisément dans la zone BA45). *A priori* l'intégration multimodale ne peut avoir lieu que si les deux modalités arrivent de manière assez "synchrone" aux zones d'intégration : si l'onde N400 –qui intervient 400 ms après la diffusion du stimulus– représente l'intégration sémantique, il semble assez évident qu'une partie de l'intégration sémantique se fera sans prendre en compte les deux modalités si les deux modalités sont désynchronisées de plus de 400 ms. L'atout principal de l'EEG étant sa précision temporelle, une étude menée par Habets et collaborateurs [68] a permis de montrer que l'intégration sémantique multimodale n'était possible que lorsque les deux signaux étaient perçus assez proches temporellement (*i.e.* moins de 360 ms d'écart entre les onsets du geste et de la parole associée).

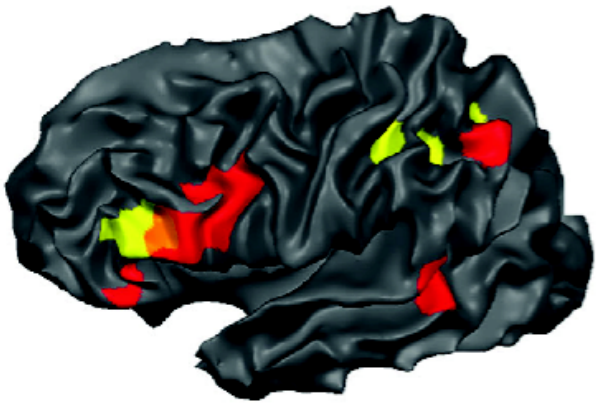


FIGURE 1.9 – Activations significatives lors de la perception de parole et geste (iconique) dans un énoncé, les incongruences sont dues au contexte. Pour les conditions "incongruence" parole *vs* congruent (en rouge), "incongruence" geste *vs* congruent (en jaune), en orange, les aires de recouvrement (BA45). Aucune activation dans l'hémisphère droit (d'après [182])

Nombre d'études ont par ailleurs utilisé l'IRMf afin de localiser plus précisément les zones du cerveau prenant en charge la perception multimodale de parole accompagnée de gestes. Globalement, toutes les études en IRMf s'accordent sur des rôles forts du lobe temporal et du lobe frontal dans l'intégration sémantique des gestes. La plupart des études portant sur les gestes iconiques, en particulier les travaux de Dick et coll. [40], montrent une implication des différentes aires en fonction de l'apport sémantique des gestes iconiques et prônent une implication du sulcus temporal supérieur auquel s'ajoute un réseau cortical distribué dans le traitement de l'information sémantique des gestes. Ceci est corroboré par plusieurs articles issus d'autres groupes de recherche. Ainsi, Holler et ses collègues [75] avancent un rôle important des sulcus/gyrus temporal supérieur (bilatéralement) dans l'intégration multimodale ainsi que du gyrus frontal inférieur (pour l'intégration sémantique) et du pars triangularis. Des résultats similaires sont obtenus par Skipper et coll. [166] (en insistant sur le rôle du pars triangularis dans la recherche sémantique et du pars opercularis par rapport à l'attention). Finalement, des résultats équivalents sont montrés pour d'autres types de gestes : Hubbard et coll. [81] pour les gestes de battement trouvent que le gyrus/sulcus temporal supérieur est activé lors de la perception conjointe de gestes de battements et de parole (mais pas lors de la production de battements seuls, et de manière beaucoup plus intense qu'en parole seule), de même Kircher et coll. [97] montrent une implication des mêmes régions pour les gestes métaphoriques.

Enfin, une étude intéressante montre que les réseaux de traitement de l'information sémantique ne sont pas figés dès la naissance. Ceux-ci évoluent au fil du temps : on observe un changement développemental des réseaux traitant la sémantique des gestes, en particulier chez les enfants. En réalisant une étude longitudinale en EEG chez les enfants, Sheehan, Namy et Mills [165] ont montré que les déflections typiquement liées au traitement sémantique (N400) changeaient entre 18, 26 mois et l'âge adulte.

Les analyses présentées ci-dessus, permettent de montrer à quel point le geste manuel et la parole sont

intimement liés. En particulier, les études récentes en imagerie cérébrale permettent d'avancer rapidement des hypothèses quant aux traitements que subit l'information liée au geste lors de sa perception.

1.5 Récapitulatif et problématique du manuscrit

1.5.1 Les enjeux d'une meilleure compréhension dans les interactions gestes manuels/parole

La rapide revue de littérature proposée ci-avant permet d'avoir un aperçu sur les raisons pour lesquelles la gestualité est une notion importante dans la communication parlée. En particulier, il a été montré un possible rôle (débattu) de la gestualité dans l'apparition du langage et de manière plus consensuelle un rôle des gestes (en particulier, du geste de pointage) dans l'acquisition du langage chez l'enfant. Hormis son rôle dans l'évolution, c'est depuis l'antiquité que la gestualité a toute sa place dans les interactions entre êtres humains, une attention particulière lui ayant été conférée dans le but d'améliorer l'efficacité communicative lors de discours par exemple. En effet, quel que soit le rôle précis qui est attribué aux gestes, toutes les théories s'accordent sur le fait que les gestes manuels ont un rôle méliorateur dans la communication humaine.

Il a ensuite été montré, par l'intermédiaire du modèle de McNeill, qu'il était probable que ces différents liens entre les gestes manuels et la parole soient l'expression de mécanismes plus "profonds" qui, faisant subir des traitements à une idée, mèneraient à extérioriser cette idée sous une forme multimodale. Plusieurs modèles ont par ailleurs été exposés afin d'expliquer les possibles étapes de développement de l'idée en un message multimodal. Ainsi, selon cette théorie, l'étude précise de la coordination geste manuel/parole pourrait aider à la caractérisation plus fine des mécanismes sous-jacents à cette coordination *i.e.* à décrire la "traduction" d'une idée en un flux manuel et un flux vocal. Les travaux précurseurs menés par Kendon et quelques autres travaux ultérieurs, ont mené à la présentation du rôle possible de la prosodie dans la coordination entre les deux modalités. Ce constat fait émerger **l'étude de la deixis** comme un cadre théorique intéressant pour mieux comprendre la coordination entre les modalités gestuelle et vocale, la fonction de deixis étant au carrefour entre la prosodie et les gestes manuels (plus précisément, la focalisation prosodique et le geste de pointage ayant un rôle commun de désignation dans les deux modalités d'intérêt).

Des arguments venant de plusieurs types de recherches(perception/production, études comportementales/neuroimagerie, études plus ou moins contrôlées…) et d'études s'intéressant à une variété de gestes importante ont montré que les gestes et la parole étaient bien des modalités en totale coopération/coordination. Globalement, une grande partie des recherches mentionnées ci-dessus (et présentes dans la littérature) étudient la relation entre gestes manuel et parole de façon qualitative (*i.e.* décrivent les liens sémantiques entre les deux modalités, essaient de préciser le rôle des gestes dans la communication,…). Afin de décrire pleinement les phénomènes, il est intéressant de compléter ces résultats d'**études quantitatives** permettant de *chiffrer* les phénomènes mis en jeu (et ainsi appuyer ou réfuter certains résultats mis en avant grâce à un point de vue différent sur les données, typiquement : la caractérisation temporelle permet de savoir s'il est *possible* qu'un traitement en suive un autre, sachant la durée de chacun des traitements à réaliser).

Dans un souci de prendre en compte la totalité de l'acte communicatif (et en particulier, son sens), une grande partie des études qualitatives se sont appuyées sur des données vidéo : la vidéo permet une capture complète d'une scène et des processus précis ont été mis au point au fil des années afin de pouvoir annoter celles-ci de façon standardisée. L'annotation vidéo comporte un bon nombre d'avantages (annotation directe du signal multimodal, et donc prise en compte des interactions entre les modalités étudiées) mais comporte aussi quelques inconvénients (annotation manuelle des évènements, annotation d'un signal en deux dimensions). L'étude quantitative des relations entre geste et parole a été abordée par quelques équipes de recherche. Ces études ont pour certaines eu recours à la **capture de mouvements** (utilisant les ultrasons pour de Ruiter [38], le theremin chez Rusiewicz [160], les infrarouges pour Rochet-Capellan et Leonard et Cummins [157, 113], …) afin d'obtenir des mesures tridimensionnelles annotables automatiquement. Ces deux approches sont complémentaires et ne sont pas incompatibles.

Toutes les études pour lesquelles un dispositif dedié à la capture de mouvement a été utilisé ont utilisé des paradigmes très contrôlés, en particulier en utilisant des productions vocales souvent réduites à des mots, groupes

de mots ou non-mots en isolation. Il est possible que ces productions vocales non naturelles influent fortement sur la coordination entre le geste manuel et la parole associée. Une prolongation intéressante de ces travaux consisterait donc à prendre en compte des **productions de parole constituées de phrases grammaticalement correctes** afin de se rapprocher des productions naturelles.

De même, les protocoles contrôlés ont le plus souvent mené les équipes de recherche à étudier la production/perception des gestes accompagnant la parole en s'appuyant sur la production/perception d'un participant seul. Or il a été mentionné en filigrane l'intérêt des gestes dans la communication, il semble alors évident qu'un protocole ne permettant pas de communication avec autrui peut induire une modification des comportements observés. La **mise au point d'expériences permettant une interaction** est un point important pour les études sur la coordination gestes manuels/parole, parfois sous-estimé dans les études portant sur cette coordination.

1.5.2 Programme et plan de cette thèse

Cette thèse a pour but principal de caractériser de façon quantitative les interactions entre les deux modalités manuelle et parole. Comme vu précédemment, les études présentées ci-après se placent dans le cadre de la désignation, ce cadre théorique représentant un environnement de choix pour étudier les interactions gestes manuels/parole. En s'appuyant sur la revue de bibliographie présentée ci-avant, au niveau expérimental quelques points importants seront respectés.

Les études utiliseront un système de suivi de mouvement précis (de type Optotrak – suivi infra-rouge), parfois avec appui de la vidéo, ceci permettra l'acquisition de données tridimensionnelles précises qui donnera la possibilité une annotation objective et semi-automatique des données de mouvements (manuels et articulatoires). Les études précédentes ont montré un rôle possible de la prosodie dans la coordination des productions des gestes manuels et de la parole et certaines études ont mis au jour des liens entre les mouvements articulatoires et les mouvements de la main. C'est ainsi que les variables d'intérêt considérées dans ces travaux sont les indices prosodiques et articulatoires de la parole et les mouvements des mains. Enfin, les études proposées se feront dans le souci d'avoir des études très contrôlées dans un premier temps (tout en veillant à conserver un corpus constitué d'éléments bien formés –phrases complètes) puis de relâcher progressivement les contraintes expérimentales pour rendre la situation expérimentale de plus en plus naturelle.

Selon ce fil conducteur, cette recherche se déroulera selon quatre phases expérimentales. Dans un premier temps (Chapitre 2), le manuscrit s'attachera à décrire la coordination entre gestes manuels et parole dans la focalisation prosodique dans une tâche assez contrôlée (tout en conservant des réponses assez "naturelles" car celles-ci s'intègrent dans un contexte communicatif et comportent des phrases complètes). Une deuxième partie (Chapitre 3) sera consacrée à l'étude de l'effet sur la coordination geste manuel/parole du lien (qu'il soit fonctionnel, ou de sens) qui unit le geste manuel et la parole que celui-ci accompagne. Une troisième partie s'intéressera à une situation communicative en interaction, en s'efforçant de proposer un protocole original permettant à la fois une interaction "naturelle" et un suivi de mouvements précis. Par ailleurs, l'étude de l'influence d'une perturbation de la communication sur l'interaction sera une condition d'intérêt de cette étude (comme décrit plus loin). Enfin, comme dans la troisième partie (Chapitre 4) quelques contraintes sont encore imposées sur les productions, une dernière étude proposera des pistes pour la réalisation d'une expérience totalement naturelle. Différents types de gestes seront considérés afin de tester l'influence des différents gestes sur la coordination dans les deux premières parties puis la coordination entre parole et geste de pointage sera explorée plus en détail dans la troisième partie. Une quatrième partie (Chapitre 5) se concentrera sur la mise au point de protocoles plus naturels et sur la description d'une méthode d'annotation possible pour les interactions multimodales. Ce manuscrit se clôturera par une discussion générale qui, après une rapide synthèse des résultats, tirera quelques conclusions en les mettant en perspective avec les résultats vus dans cette partie introductive avant de s'intéresser à des perspectives possibles pour ces travaux.

Chapitre 2

Les gestes manuels et la focalisation prosodique : des données sur la coordination parole/main

Cette première étude nous permettra de poser les bases méthodologiques et expérimentales de l'ensemble de notre travail, elle en fournit l'expérience noyau à partir de laquelle les autres expériences nous conduiront à complexifier progressivement notre approche.

Cette expérience princeps est ainsi volontairement assez contrainte. Elle consiste à analyser la coordination entre différents types de gestes manuels, et des phrases les plus simples possibles, de structure grammaticale sujet/verbe/complément, avec deux possibilités d'emplacements pour la focalisation, soit sur le sujet soit sur le complément. Pour ce faire, un protocole expérimental permettant la production naturelle d'une correction (et donc de focalisation contrastive prosodique), mis au point par Dohen [43] a été utilisé. Il vise à permettre aux locuteurs de produire alternativement, en fonction du lieu de correction (sujet ou complément), des focalisations sur l'un ou l'autre des éléments de la phrase. La question est de savoir comment les gestes associés à la parole voient leur exécution se modifier selon l'emplacement de la focalisation afin de connaître les évènements temporels de chaque modalité qui se coordonnent et d'estimer s'il existe des stratégies différentes de coordination.

Cette étude vise également à aborder une question expérimentale qui nous semble très importante, et qui est la question de la nature cognitive des mécanismes de coordination. Si ces mécanismes sont essentiellement liés à un couplage plus ou moins automatique entre deux systèmes moteurs partenaires (le système orofacial et le système bracchiomanuel) alors la coordination devrait s'observer quel que soit le type de geste produit. Si au contraire la coordination est linguistiquement et cognitivement contrôlée, alors elle doit dépendre crucialement du type de geste produit, et notamment différer selon qu'il s'agit d'un geste de communicatif ou non, ou non. Ainsi, cette étude vise à comparer la coordination pour plusieurs types de gestes.

2.1 Protocole expérimental

La description du protocole dans ce chapitre ainsi que dans les suivants commencera toujours par la description des corpora qui permettent de fixer le cadre linguistique dans lequel l'expérience est menée. Suite à la présentation des participants et à la description des conditions d'expérimentation, la description de la tâche expérimentale à proprement parler sera effectuée.

2.1.1 Corpus

Le corpus utilisé pour cette expérience se compose de 4 phrases déclaratives en français. Ces 4 phrases sont présentées dans la Table 2.1.

Sujet	verbe	Objet
Mumu	tient	le bébé
Baba	vend	le balai
Mumu	voit	le bonbon
Baba	sent	le melon

TABLE 2.1 – Corpus de l'expérience

Les 4 phrases utilisées ont une structure grammaticale identique simple : Sujet (un prénom inventé) - verbe (verbe au présent de l'indicatif, troisième personne du singulier) - Objet (un article défini et un nom commun). Par ailleurs, les phrases ont toutes la même structure syllabique (Sujet : CVCV, Verbe : CV, Objet : CV+CVCV). Les syllabes sont toutes du type Consonne-Voyelle (CV) pour les parties de la phrase qui nous intéressent (Sujet et Objet). Les différents mots utilisés pour le Sujet ou pour l'Objet sont constitués de consonnes bilabiales à l'attaque (/m,b,p/ et de voyelles "ouvertes" (/a, e, ɜ/) ou "protruses" (/y, õ, ø/). Ce choix sera expliqué plus tard.

2.1.2 Participants

Huit femmes et deux hommes ($\mu_{age} = 30,2$, $\sigma_{age} = 8,94$) ont pris part à cette étude. Tous les participants étaient droitiers (test d'Oldfield [139]), ne présentaient pas de troubles de l'audition et avaient une vision normale ou corrigée sans troubles lors de la passation de l'expérience.

2.1.3 Conditions expérimentales

Le design expérimental mis au point permet d'étudier la coordination entre production de gestes de différents types et focalisation de localisation variable. Les variables indépendantes étudiées sont ainsi l'emplacement de la focalisation (facteur : FOCALISATION) et le type de geste (facteur : GESTE).

Niveau du facteur	Exemple
Focalisation sur le début de l'énoncé (Foc1)	**Mumu** tient le bébé
Focalisation sur la fin de l'énoncé (Foc2)	Mumu tient le **bébé**

(a) Niveaux du facteur FOCALISATION

Niveau du facteur	Description
Parole seule	Aucun geste produit
Pointage	Production d'un geste déictique communicatif
Battement	Production d'un geste communicatif non déictique
Contrôle	Appui simple sur un bouton (non communicatif)

(b) Niveaux du facteur GESTE

TABLE 2.2 – Les facteurs de l'expérience

Il est nécessaire de faire varier l'emplacement de la focalisation (Foc) au sein des énoncés de parole si on veut pouvoir caractériser une quelconque coordination entre focalisation et production de gestes. En effet, si la focalisation était systématiquement placée en fin d'énoncé par exemple, on ne pourrait pas distinguer un effet dû à la focalisation d'un effet dû à la fin d'un énoncé. Ainsi, le facteur FOCALISATION est décomposé en deux niveaux : la focalisation a été étudié en début d'énoncé (focalisation sur le sujet de la phrase, nommé *Foc1* par la suite) ou en fin d'énoncé (focalisation sur l'objet de la phrase, nommé *Foc2* dans ce qui suit). Une illustration des deux niveaux du facteur est donnée en Table 2.2a [1].

1. Les caractères gras représentent la partie focalisée de la phrase

Comme nous l'avons vu dans la Section 1.2.1, les gestes peuvent être classés selon leur relation à la parole. Il est possible que la production d'un geste n'ayant qu'une lien faible avec la parole ait une interaction limitée avec la production de celle-ci. Nous nous intéressons ici à divers types de gestes afin de voir si cela se vérifie. Par ailleurs, l'étude de la bibliographie nous montre une influence de la production de gestes sur la production de parole (*cf.* Section 1.4.2.3), nous essayons ici de décrire cette influence. Le facteur GESTE est constitué de quatre niveaux : trois types de gestes distincts ont été étudiés ainsi qu'une condition sans geste. La condition sans geste (*Parole seule*) nous permet, par la comparaison avec les conditions avec geste, d'étudier l'éventuelle influence de la production de gestes sur la production de parole. Les productions de trois types de gestes ont par ailleurs été prises en compte, chaque type de geste étant étudié séparément (en tant que niveau différent du facteur GESTE). Le geste de pointage (condition *Pointage*) est un geste déictique et donc naturellement lié à la focalisation (*cf.* Section 1.3.4). Le geste de battement (condition *Battement*) est un geste fréquemment utilisé dans le discours qui permet en particulier de mettre en valeur une partie de celui-ci, il a donc un rôle d'emphase qui est similaire dans une certaine mesure à la focalisation prosodique (pour rappel, *cf.* Section 1.2.1). Ces deux types de gestes sont clairement communicatifs et liés à la parole par des liens pragmatiques (démonstration ou mise en relief) avec la parole focalisée. Un autre type de geste (condition *Contrôle*), n'ayant aucun rapport avec la situation communicative a été étudié afin d'évaluer l'incidence du caractère communicatif du geste sur la coordination parole/geste.

Le paradigme expérimental utilisé est un paradigme par blocs. Au sein de chaque bloc, une seule condition GESTE est utilisée et le facteur FOCALISATION varie. L'ordre de passage des blocs a été varié parmi les participants. Chaque bloc comporte 16 essais (4 phrases avec 2 niveaux distincts de focalisation, 2 répétitions). Au sein de chaque bloc, l'ordre des phrases ainsi que la position de la focalisation ont été randomisés. On désignera par la suite comme "essai" une réalisation du quintuple (Numéro de participant, Condition GESTE, Condition FOCALISATION, Numéro de phrase, Numéro de répétition).

2.1.4 Dispositif expérimental

L'expérience s'est déroulée sur le site Stendhal du Département Parole et Cognition du GIPSA-LAB. Une représentation schématique de la salle d'expérience est fournie en Figure 2.1. Les participants étaient assis sur une chaise dans une chambre sourde, sur leur côté droit une table sur laquelle était repérée une position de repos (disposée à une position confortable pour chaque participant) qui a servi de position de référence pour les mouvements de la main droite. Sur cette table, à une distance de $26,5$ cm en avant de la position de repos, était disposé le bouton sur lequel appuyer en condition *Contrôle*. Un écran semi-transparent disposé deux mètres devant le participant permettait la projection des stimuli visuels. L'écran avait une largeur de 50 cm et une hauteur de 12 cm. Les cibles présentées sur cet écran avaient une taille de $10\,\mathrm{cm} \times 10\,\mathrm{cm}$ et étaient écartées de 30 cm centre-à-centre.

Les expérimentateurs ainsi que les ordinateurs permettant la diffusion des stimuli et l'enregistrement des données étaient sur la gauche du participant. Les stimuli ont été générés et diffusés par le logiciel Presentation [135]. Un haut parleur permettait la diffusion des stimuli audio. Le niveau du haut-parleur était fixé pour une écoute confortable. L'enregistrement des données consistait en la récupération du signal Optotrak (Optotrak 3020) (qui se faisait sur un ordinateur dédié) et au stockage du signal audio (capté par un micro (AKG C1000S) et enregistré sur un enregistreur numérique Marantz-PMD 670).

Le système de suivi de mouvements Optotrak est constitué de trois caméras solidaires sensibles à l'infrarouge. Ces caméras permettent de suivre le déplacement de diodes émettrices infrarouges. Ces diodes sont des diodes actives (émettrices) et sont donc alimentées électriquement : chaque diode est reliée à un boitier d'alimentation par un fil. Un multiplexage des diodes a lieu et chaque diode émet du rayonnement infrarouge avec une fréquence propre, ce qui permet de différencier les diodes dans le signal final. Dans les expériences qui suivent, ces diodes sont collées sur différentes parties du corps des participants par l'intermédiaire de morceaux d'œillets adhésifs (pour l'EEG) et de sparadrap. Le signal récupéré par les caméras (position des diodes à une précision < 1 mm) est traité et l'utilisateur récupère une matrice de coordonnées tridimensionnelles représentant les positions 3D des différentes au cours du temps.

Enfin, le dispositif de suivi de mouvements tridimensionnel Optotrak est placé en face du participant à une distance de 5 mètres environ et à une hauteur de 1m80 environ. Cette position légèrement en hauteur permet d'éviter les problèmes liés au masquage des données Optotrak par l'écran.

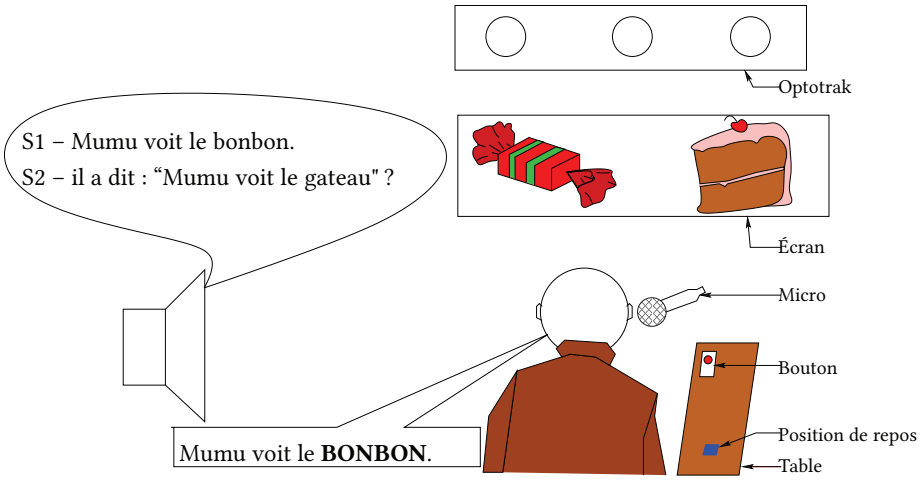


FIGURE 2.1 – Dispositif expérimental

2.1.5 Description de la tâche

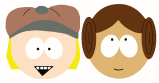
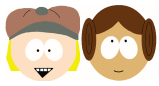
La tâche de parole utilisée consiste en la correction d'un dialogue entre deux acteurs qu'entend le participant, ce type de tâche a déjà été utilisé pour obtenir de la production de focalisation contrastive (*cf.* [43]). Le dialogue est diffusé grâce à un haut-parleur. Dans ce dialogue, le Locuteur 1 prononce une affirmation, cette affirmation est mal comprise par le Locuteur 2. Le Locuteur 2 demande alors de l'aide au participant afin de comprendre ce qui a été dit, comme illustré en Table 2.3. Le participant doit ensuite corriger la phrase du Locuteur 2 pour l'expérimentateur assis à côté de lui. Cette correction réalisée par le participant le fait naturellement produire de la focalisation contrastive prosodique sur la partie corrigée de l'énoncé. Afin de limiter le risque d'erreurs pour le participant, les mots erronés choisis étaient aussi des mots de structure syllabique CVCV mais ne partageant aucune syllabe commune avec le mot correct.

	Foc1	*Foc2*
Locuteur 1	– Mumu tient le bébé.	– Baba vend le balai.
Locuteur 2	– Elle a dit : « Baba tient le bébé » ?	– Elle a dit : « Baba vend le gâteau » ?
Affichage	 	 
Participant	– **Mumu** tient le bébé.	– Baba vend le **balai**.

TABLE 2.3 – Exemples de stimuli et de production de parole du participant

Afin d'éliciter également plus naturellement des gestes, des stimuli visuels s'affichaient sur un écran à la suite du dialogue entre les deux locuteurs. Les images représentaient deux mots possibles pour la correction. Parmi ces deux images, l'une représente le mot correct (prononcé initialement par le Locuteur 1), l'autre représente un mot incorrect (prononcé initialement par le Locuteur 2). Le stimulus associé au mot correct est la cible, l'autre stimulus est un distracteur. La Table 2.3 donne un exemple de stimuli visuels pour chacune des deux conditions *Focalisation*. La position (à droite ou à gauche) des images correcte/incorrecte a été randomisée afin de rendre la prévision de l'emplacement de l'image correcte impossible avant l'affichage de celle-ci. La liste des images utilisées comme stimuli visuels est représentée en Table 2.4. Afin de se familiariser à la fois avec les images des

deux personnages utilisés (“Baba” et “Mumu”) comme sujets et avec les images des compléments d’objet, une phase d’entraînement préliminaire est réalisée.

Type d’image	Sujet	Objet
Correct		
Incorrect		

De gauche à droite et de haut en bas : Première ligne : Baba, Mumu, Bonbon, Balai, Bébé, Melon
Deuxième ligne : Baba, Mumu, Panda, Vélo, Gâteau, Jambon

TABLE 2.4 – Stimuli visuels utilisés

Il était demandé aux participants de garder leur index droit sur le repère de position de repos, sauf lorsqu’ils voulaient produire un geste. Ainsi dans la condition *Parole seule*, leur main droite devait rester au repos avec l’index sur le repère position de repos. Dans les conditions GESTE nécessitant la production de gestes (*Pointage, Battement, Contrôle*), la main droite devait partir de la position de repos et y revenir après l’exécution du geste. Les participants devaient produire un geste en même temps que la production de parole, mais aucune indication plus précise n’est donnée quant à l’instant de production du geste : la production devait être la plus naturelle possible. Dans la condition *Pointage*, le participant devait désigner par un geste de pointage la cible correspondant à la correction qu’il effectuait avec la parole. Dans la condition *Battement*, le participant devait réaliser un geste de battement (rapide mouvement vertical de la main). Dans la condition *Contrôle*, le participant devait aller appuyer sur le bouton situé en avant de la position de repos.

Un exemple complet de consigne est donné dans la Figure A.1, p.157. Dans cette figure, on voit que la consigne pour les conditions contenant des gestes est légèrement différente. Dans Figure A.1b, la partie notée <TYPE DE GESTE> peut prendre les valeurs *de pointé ; de battement ; non communicatif* et la partie <DESCRIPTION RAPIDE DU GESTE> les valeurs (respectivement) *Le geste de pointage sera effectué vers l’image correspondante ; Le geste de battement est un geste de haut en bas avec la main ; Le geste non communicatif est l’appui sur le bouton situé devant vous.*

2.1.6 Acquisition des données

2.1.6.1 Suivi de mouvements

Les positions des diodes émettrices infra-rouge Optotrak sont représentées sur la Figure 2.2. Le dispositif permet le suivi de la trajectoire des diodes au cours du temps toutes les 5 ms (f_e= 200 Hz).

Trois diodes (D_{MD}, D_{MI}, D_{MO}) sont placées respectivement sur le dos de la main droite, la première phalange de l’index droit, l’ongle de l’index droit. Par défaut, c’est la trajectoire de D_{MO} qui est utilisée pour estimer la position de l’index, mais il peut exister des situations où cette diode n’est *a priori* pas visible et où les autres diodes pourraient être utilisées pour estimer la position de D_{MO}. La diode D_{MO} est collée sur une éclisse placée sur l’ongle de l’index et pointe donc dans la même direction que l’extrémité de l’index, D_{MD} pointe dans la même direction que le dos de la main et D_{MI} pointe dans la même direction que le dos de l’index (*cf.* Figure 2.3). L’intérêt de D_{MD} est de pouvoir suivre la trajectoire de la main surtout dans le cas du geste de battement. En effet, des essais préliminaires nous ont montré une grande variabilité inter- et intra-individuelle dans la réalisation de ce geste et ont par ailleurs suggéré que la majorité de ces gestes se faisaient soit avec la paume de main parallèle au tronc, soit avec la main fermée (dans les deux cas, D_{MO} et possiblement D_{MI} sont cachées). D_{MI} peut être utile dans les cas où un geste de pointage est réalisé avec l’index orienté vers le sol pendant la phase de *stroke* et où D_{MO} ne serait alors plus visible. Dans ce cas, D_{MI} donne une meilleure approximation de la position de D_{MO} que la diode D_{MD} puisque l’index peut bouger dans des directions différentes de la paume de la main.

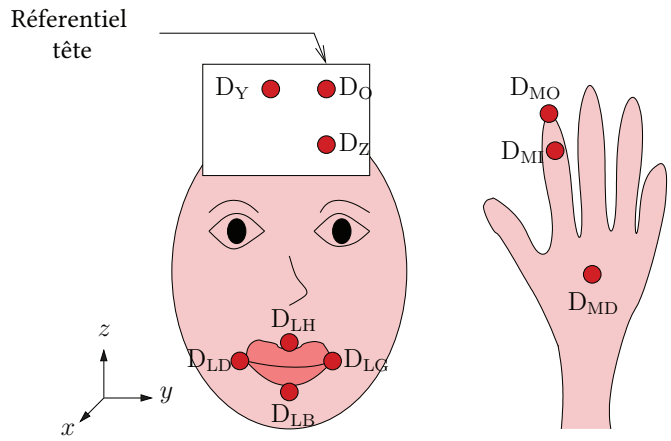


FIGURE 2.2 – Position des diodes Optotrak

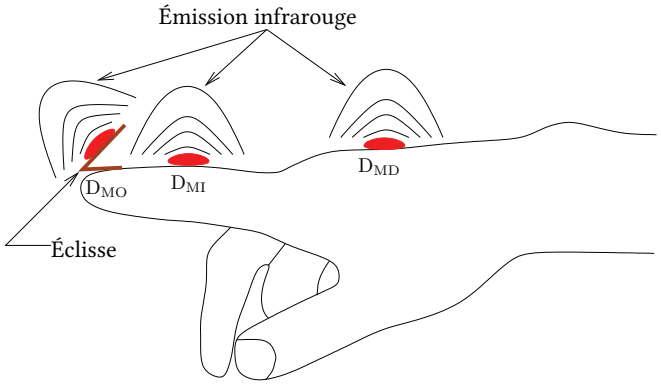


FIGURE 2.3 – Vue de côté des diodes Optotrak de la main droite

Afin de suivre la trajectoire des lèvres, quatre diodes (D_{LG}, D_{LD}, D_{LH}, D_{LB}) ont été placées sur les lèvres des participants respectivement sur la commissure gauche, la commissure droite, le milieu de la lèvre supérieure et le milieu de la lèvre inférieure. Par ailleurs, trois diodes ont été placées dans un plan rigide attaché à la tête du participant. Ces trois diodes permettent de soustraire les mouvements de la tête et du buste au mouvement des lèvres, afin d'obtenir le mouvement des lèvres dans un référentiel lié à la tête. En effet c'est le mouvement propre des lèvres qui est considéré dans cette étude.

2.1.6.2 Enregistrement du signal de parole

Le signal audio de parole a été enregistré à 22 kHz par un enregistreur Marantz-PMD. Un micro (AKG C1000S) placé devant la bouche du sujet (niveaux réglés dans une phase de calibration mais invariables au sein de chaque sujet) a permis l'acquisition du signal de parole.

2.2 Traitement des données recueillies

2.2.1 Prétraitements généraux

2.2.1.1 Validation préliminaire

La validation des données s'est réalisée en trois parties : validation gestuelle, validation du contenu et validation de la focalisation. Dans une grande majorité des cas, la tâche a bien été réalisée par les participants.

Cependant, certaines données ont du être supprimées des analyses. Dans $2,03\%$ des essais, le geste a soit été non produit, mal produit, produit avec un type de geste différent, ou contenait trop de données manquantes (masquage des diodes). Ces essais ont été écartés des analyses suivantes. Dans $3,44\%$ des cas, des hésitations, des erreurs sur la prononciation ont été relevées, de la même façon, ces essais ont été écartés pour les analyses. Enfin, tous les énoncés ont été validés acoustiquement par deux juges indépendants. La validation acoustique consistait en la détection de la présence d'au moins un corrélat acoustique de la focalisation contrastive prosodique dans chaque essai comme décrit dans [44]. Environ $2,30\%$ des données ont ainsi dû être rejetées par cette méthode.

2.2.1.2 Données articulatoires

Comme mentionné précédemment, les données articulatoires (trajectoires tridimensionnelles des diodes infrarouge placées sur les lèvres) ont tout d'abord été projetées dans un repère relatif à la tête. Cette projection permet de s'affranchir des mouvements de tête/du corps qui se superposent aux mouvements propres des lèvres dans les données. Le repère lié à la tête est généré à partir des trois diodes placées sur le "référentiel de tête" sur la Figure 2.2. Les trois diodes sont situées sur un même plan, et leurs trois coordonnées permettent de créer un repère orthonormal : si on note $\mathrm{D_O}$ la diode dans le coin du référentiel, $\mathrm{D_Y}$ la diode située sur la même ligne horizontale que $\mathrm{D_O}$ et $\mathrm{D_Z}$ la diode située sur la même ligne verticale que $\mathrm{D_O}$, un repère orthogonal lié à la tête est le repère $(\mathrm{D_O}, \overrightarrow{\mathrm{D_O D_Y}}, \overrightarrow{\mathrm{D_O D_Z}}, \overrightarrow{\mathrm{D_O D_X}})$ avec $\overrightarrow{\mathrm{D_O D_X}} = \overrightarrow{\mathrm{D_O D_Y}} \wedge \overrightarrow{\mathrm{D_O D_Z}}$. Tout point de l'espace ayant une coordonnée dans le repère initial (lié à l'Optotrak) peut alors être projeté dans le nouveau repère par la formule de changement de repère classique. Si on nomme $\mathcal{P}(t)$ la matrice de passage du repère (fixe) lié à l'Optotrak au repère (variable dans le temps) lié à la tête *i.e.* $\mathcal{P}(t) = \begin{pmatrix} x_{\overrightarrow{\mathrm{D_O D_Y}}}(t) & x_{\overrightarrow{\mathrm{D_O D_Z}}}(t) & x_{\overrightarrow{\mathrm{D_O D_X}}}(t) \\ y_{\overrightarrow{\mathrm{D_O D_Y}}}(t) & y_{\overrightarrow{\mathrm{D_O D_Z}}}(t) & y_{\overrightarrow{\mathrm{D_O D_X}}}(t) \\ z_{\overrightarrow{\mathrm{D_O D_Y}}}(t) & z_{\overrightarrow{\mathrm{D_O D_Z}}}(t) & z_{\overrightarrow{\mathrm{D_O D_X}}}(t) \end{pmatrix}$, et les coordonnées d'un point M dans le repère lié à l'Optotrak $M(t) = \begin{pmatrix} x(t) \\ y(t) \\ z(t) \end{pmatrix}$, alors les coordonnées de ce même point M dans le repère lié à la tête sont : $\begin{pmatrix} x\prime(t) \\ y\prime(t) \\ z\prime(t) \end{pmatrix} = \mathcal{P}(t)^{-1} \begin{pmatrix} x(t) \\ y(t) \\ z(t) \end{pmatrix}$ (ce calcul est toujours faisable puisqu'une matrice de passage est toujours inversible).

Les données extraites de la trajectoire des marqueurs des lèvres sont :

- L'ouverture des lèvres : distance euclidienne entre $\mathrm{D_{LH}}$ et $\mathrm{D_{LB}}$
- La protrusion de la lèvre supérieure : Coordonnée selon l'axe x de $\mathrm{D_{LH}}$

L'utilisation de ces grandeurs est justifiée par le fait que toutes les voyelles du corpus sont des voyelles ouvertes (/a, e, ɜ/) ou "protruses" (/y, õ, ø/). Les cibles articulatoires des voyelles ouvertes sont repérables dans nos données par des maxima d'ouverture des lèvres, les cibles articulatoires des voyelles protruses par des maxima de protrusion de la lèvre supérieure. Un des intérêts de l'Optotrak est que les coordonnées recueillies sont en millimètres. Ainsi, à la fois l'ouverture et la protrusion ont des valeurs interprétables directement (ce sont des distances euclidiennes, en millimètres).

2.2.1.3 Données manuelles

Suite à l'enregistrement du corpus, il s'est avéré, par chance, que la diode du bout de l'index ($\mathrm{D_{MO}}$) était visible 100% du temps. Ainsi, seule cette diode a servi à suivre les mouvements de la main et de l'index. Tous les gestes manuels étudiés dans cette étude sont principalement produits dans un seul et même plan et, en simplifiant à l'extrême, selon un seul axe : le geste de contrôle a clairement une trajectoire rectiligne, un geste de pointage a la plupart du temps une trajectoire rectiligne (*cf.* ci-dessous), et le geste de battement (hors préparation) a une trajectoire selon un seul axe (haut/bas). Ainsi, on peut espérer réduire la dimensionnalité du problème en projetant les trajectoires sur l'axe qui explique le mieux leur variance. Nous avons eu recours à une Analyse en Composantes Principales (ACP) pour réaliser cela. Pour chaque essai de chaque participant, une ACP a été

réalisée et il s'est avéré que la première composante (*i.e.* l'axe expliquant au mieux la variance) expliquait la variance à $89,6\%$ en moyenne sur tous les essais de tous les participants. La majeure partie de la variance étant expliquée par la première composante de l'ACP, il a été convenu de n'utiliser que cette composante pour étudier le mouvement du doigt.

Une fois l'ACP réalisée, afin de s'affranchir des petits mouvements parasites (tremblements,...) ou des erreurs de mesure (précision Optotrak), ces signaux ont été filtrés par un filtre passe-bas de Butterworth d'ordre 3 avec une fréquence de coupure à 15 Hz. L'utilisation d'un filtre de type Butterworth est justifiée par le fait que ce type de filtre offre un gain constant dans la bande passante (et surtout un retard de groupe constant pour les signaux ayant une fréquence inférieure à $\frac{f_c}{2}$ *i.e.* $7,5$ Hz).

2.2.2 Annotation des données

Afin de pouvoir étudier la coordination entre les différents signaux (et donc, la coordination geste manuel/parole), il faut avant tout repérer sur chaque signal des points d'intérêt. L'annotation des données s'est principalement réalisée avec deux outils : Praat [16] et Mathworks Matlab. Une interface graphique, développée sous Matlab (voir la capture d'écran Figure 2.4), permet de valider les annotations automatiques des points d'intérêt mentionnés ci-après et autorise la modification de celles-ci si nécessaire.

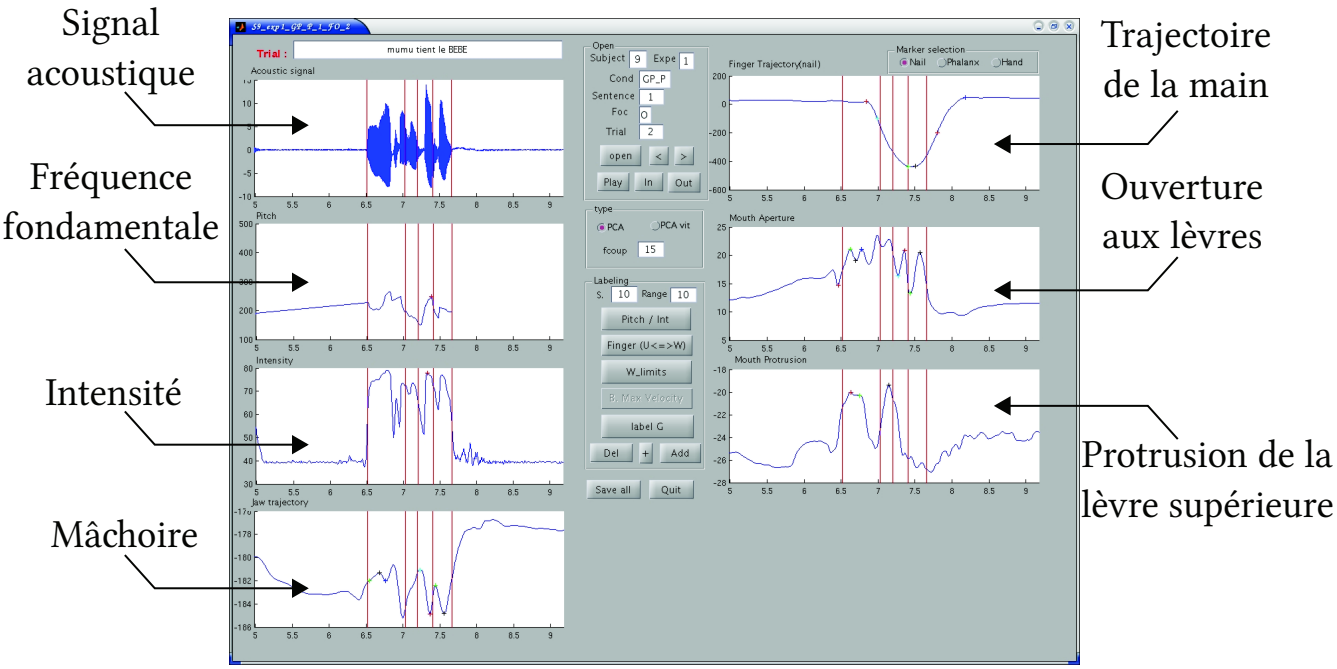


FIGURE 2.4 – Interface d'annotation Matlab

2.2.2.1 Segmentation & annotation acoustique

Praat a été utilisé afin d'annoter manuellement les limites acoustiques des syllabes. Ce logiciel a également été utilisé afin d'extraire la courbe de fréquence fondamentale (F_0) ainsi que la courbe d'intensité (Int) pour chaque essai. La fréquence fondamentale a été extraite par une méthode d'auto-corrélation avec une fenêtre d'intégration de 10ms, l'intensité a été extraite avec la même fenêtre d'intégration..

Ces deux grandeurs ont ensuite été annotées en repérant leur maximum *sur la partie focalisée* de chaque essai, ces maxima sont nommés F_0 et Int par la suite.

2.2.2.2 Données articulatoires

Pour chaque essai, l'articulatoire de la partie focalisée est annotée semi-automatiquement (annotation automatique vérifiée *a posteriori* manuellement). La partie automatique de l'annotation se fait de la façon suivante : L'utilisation de consonnes bilabiales avec fermeture initiale des lèvres en début de production permet de repérer facilement celles-ci articulatoirement (minimum d'ouverture), puis chaque mot suit un "patron" articulatoire typique (un "patron" en ouverture, un "patron" en protrusion), l'algorithme mis au point cherche à mettre en correspondance le "patron" typique et le signal enregistré. Chaque mot présent dans la partie focalisée de l'énoncé présente une combinaison de deux voyelles parmi /a, e, ɜ, y, õ, ø/. Pour chaque voyelle protruse, un maximum de protrusion de la lèvre supérieure a été annoté, pour chaque son réalisé avec les lèvres ouvertes, un maximum d'ouverture des lèvres a été repéré sur le signal.

Pour chaque essai, les deux maxima annotés (un maximum par syllabe) représentent les deux cibles articulatoires (CVs) du mot focalisé. Celles-ci seront nommées CV_1 et CV_2 par la suite.

2.2.2.3 Données manuelles

Pour chaque essai (sauf dans la condition *Parole seule*), la trajectoire de l'index a été annotée. Sur cette trajectoire, 5 points ont été annotés : l'instant de départ de la position de repos ou onset du geste ($\mathrm{P_{On}}$), l'apex du geste ($\mathrm{P_A}$), l'instant de départ de la position d'apex ou retour du geste ($\mathrm{P_R}$), l'instant de retour à la position de repos ou offset du geste ($\mathrm{P_{Off}}$). Chacun de ces points est annoté grâce à la détection d'un pic de vitesse : soit celui du lancer du geste ($\mathrm{P_{vit}}$), soit celui du retour du geste ($\mathrm{P_{vit2}}$). $\mathrm{P_{On}}$ (resp. $\mathrm{P_A}$) est le point où la vitesse atteint 10% de sa valeur en $\mathrm{P_{vit}}$ avant (resp. après) $\mathrm{P_{vit}}$. $\mathrm{P_R}$ (resp. $\mathrm{P_{Off}}$) est le point où la vitesse atteint 10% de sa valeur en $\mathrm{P_{vit2}}$ avant (resp. après) $\mathrm{P_{vit2}}$. La Figure 2.5 résume cette façon d'annoter pour les différents types de gestes.

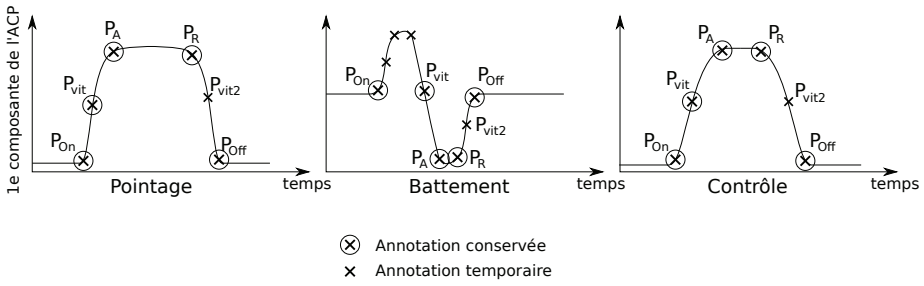


Figure 2.5 – Annotation typiques idéalisée des trajectoires des gestes manuels

Chaque geste est donc constitué de trois (ou quatre) principales phases (conformément à [96], puisque la phase de préparation n'est pas nécessaire dans le dispositif expérimental pour les gestes de *Pointage* et *Contrôle*). Pour tous les types de gestes, $\mathrm{P_{On}}$ et $\mathrm{P_{Off}}$ représentent l'instant où, respectivement, l'index part de et revient à la position de repos indiquée sur la table ; la signification de $\mathrm{P_A}$ dépend du type de geste (*cf.* ci-dessous) ; $\mathrm{P_R}$ est l'instant où le geste quitte sa position $\mathrm{P_A}$ pour retourner vers la position de repos ; enfin, les pics de vitesse $\mathrm{P_{vit}}$ et $\mathrm{P_{vit}}$ 2 sont respectivement les pics de vitesse pour atteindre $\mathrm{P_A}$ et $\mathrm{P_R}$.

Pour le pointage, l'apex représente l'instant où l'index et le bras sont étendus au maximum vers le stimulus visuel affichée sur l'écran. Souvent, la position d'apex est tenue (*i.e.* l'apex reste pointé vers la cible), c'est ce que Kendon appelle le *post-stroke hold*. L'index se déplace ensuite jusqu'à la position de repos sur la table : $\mathrm{P_R}$ est l'instant où ce mouvement de retour commence.

Le geste de contrôle a une trajectoire proche de celle du geste de pointage. Il est possible d'annoter les mêmes points sur sa trajectoire. La différence majeure entre les deux gestes est que le geste de contrôle ne désigne aucune cible, on observe par ailleurs que le *post-stroke hold* est plus court que pour le geste de pointage.

Le geste de battement contient une phase de préparation en sus des trois phases présentées ci-dessus. En effet, pour effectuer le geste de battement, la main part d'une position basse (sur la table à droite du participant), celle-ci

Signal annoté	Évènement annoté	Dénomination
Geste manuel	Départ du geste	$t_{\mathrm{P_{On}}}$
	Pic de vitesse du geste (du stroke)	$t_{\mathrm{P_{vit}}}$
	Apex du geste	$t_{\mathrm{P_A}}$
	Retour du geste	$t_{\mathrm{P_R}}$
	Fin du geste	$t_{\mathrm{P_{Off}}}$
Acoustique	Début/Fin de production vocale	$t_{\mathrm{Deb}}, t_{\mathrm{Fin}}$
	Pic de fréquence fondamentale du mot focalisé	$t_{\mathrm{F_0}}$
	Pic d'intensité du mot focalisé	t_{Int}
Articulatoire	Première et seconde cible articulatoire (vocalique) du mot focalisé	$t_{\mathrm{CV_1}}, t_{\mathrm{CV_2}}$

TABLE 2.5 – Évènements temporels annotés sur les différents signaux

doit donc s'élever avant de pouvoir effectuer le mouvement rapide de haut en bas qui constitue un battement. Une fois le point culminant atteint, le battement est effectué avant que la main retourne à sa position de repos. Comme dans l'article de Leonard et Cummins [113, page 14], les points annotés sont situés hors de la phase de préparation, le point considéré comme apex est le point le plus bas (en fin de mouvement vertical), le point $\mathrm{P_R}$ est *a priori* très proche de $\mathrm{P_A}$.

Par convention, on nommera *stroke* l'intervalle $[\mathrm{P_{On}}; \mathrm{P_A}]$, *plateau* ou *tenue* l'intervalle $[\mathrm{P_A}; \mathrm{P_R}]$ et *stroke de retour* $[\mathrm{P_R}; \mathrm{P_{Off}}]$. Par ailleurs, dans la suite, lorsqu'il sera fait référence aux instants de réalisation de ces différents évènements, ceux-ci seront nommés $t_{\mathrm{P_{On}}}, \ldots, t_{\mathrm{P_{Off}}}$, et les moyennes de ces instants (par type de geste) seront nommés $t^P_{\mathrm{P_{On}}}, \ldots, t^P_{\mathrm{P_{Off}}}$ pour le pointage, $t^B_{\mathrm{P_{On}}}, \ldots, t^B_{\mathrm{P_{Off}}}$ pour le battement et $t^C_{\mathrm{P_{On}}}, \ldots, t^C_{\mathrm{P_{Off}}}$ pour le geste de contrôle.

Toutes ces données ont été annotées automatiquement puis ces annotations ont été validées manuellement (et modifiées le cas échéant) grâce à l'interface graphique développée à cet effet. Dans le cas d'une annotation corrigée, le mode d'annotation reste le même : détection d'un pic de vitesse puis annotation des points à 10% de ce pic avant et après le pic, seul l'intervalle au sein duquel ces détections sont réalisées étant modifiable.

La Table 2.5 résume l'ensemble des évènements annotés sur les différents signaux.

2.3 Analyse des données

2.3.1 Prédictions

Comme expliqué en Section 2.1.3, les gestes de pointage et de battement entretiennent des relations particulières avec la parole et en particulier avec la focalisation contrastive prosodique.

Le pointage et la focalisation contrastive prosodique ont un rôle communicatif commun : la désignation ; ils ont par ailleurs la cible de la désignation en commun (un personnage dans le cas *Foc1*, un objet dans le cas *Foc2*). Ces points communs mènent à penser que les deux évènements devraient "être réalisés au même moment" ce qui entraînerait un effet du facteur FOCALISATION sur les instants de production du geste de pointage *i.e.* typiquement un pointage cooccurrent avec le sujet de l'énoncé en condition *Foc1*, et cooccurrent avec l'objet en condition *Foc2*. Aucune hypothèse plus précise n'est avancée quant aux lieux possibles de synchronisation. Au niveau du geste, l'apex est le premier instant où la désignation est stable, c'est donc un lieu de synchronisation possible important à étudier. Au niveau de la parole, la littérature propose des lieux de synchronisation avec le geste assez variés (et aucun "consensus" n'est avancé) : cibles articulatoires pour Rochet-Capellan [157], indices acoustiques pour de Ruiter [38] ou Rusiewicz [160].

Le geste de battement entretient selon la littérature un lien avec la focalisation car ils ont un rôle pragmatique commun : l'emphase. Une hypothèse possible est donc, comme ci-dessus, que le geste de battement soit réalisé "en même temps" que la focalisation contrastive prosodique, ce qui prédit un effet du facteur FOCALISATION sur les instants de réalisation du geste de battement. Ici encore, les lieux possibles de cooccurrence sont peu consensuels,

la littérature mentionne néanmoins la possible synchronie entre pic de fréquence fondamentale et "apex" (au sens ci-dessus).

Enfin le geste de contrôle est un geste non communicatif et non déictique : il n'entretient *a priori* aucune relation avec la parole. On s'attend ainsi à n'avoir aucun effet du facteur FOCALISATION sur les instants de production du geste de type contrôle.

Le paradigme nous permet enfin d'étudier l'influence de la production de gestes sur la production de parole, suivant Krahmer et Swerts [103] il est attendu un renforcement des corrélats articulatoires/acoustiques liés à la focalisation lorsque les gestes produits sont liés à la parole *i.e.* les gestes communicatifs et lors de la production de gestes par rapport à la condition parole seule.

2.3.2 Traitements *a priori* et analyses statistiques

2.3.2.1 Prétraitement des données temporelles et d'amplitude

Afin de pallier les variations segmentales et les réalisations interparticipants et intraparticipants des énoncées du corpus, et pour éliminer l'influence du temps de réaction, les données temporelles ont été normalisées sur la durée acoustique de chaque production. Soit x un évènement annoté, t_{Deb}, t_{Fin} respectivement le début et la fin acoustique de la production, si on note t^i_{x} l'instant de production non normalisé et t_{x} l'instant de production normalisé, on pose $t_{\text{x}} = \frac{t^i_{\text{x}} - t_{\text{Deb}}}{t_{\text{Fin}} - t_{\text{Deb}}}$, ainsi pour les temps normalisés, le temps 0 correspond au début acoustique de l'énoncé, le temps 1 à la fin acoustique de l'énoncé. Cette normalisation permet de pouvoir comparer temporellement les productions entre elles relativement à la durée acoustique de la parole.

Une fois ces normalisations effectuées, pour chaque mesure, les valeurs atypiques ont été corrigées. Pour tout triplet de GESTE×FOCALISATION×numéro de phrase, pour toutes les mesures effectuées, la moyenne ainsi que l'écart-type des essais ont été calculés. Toute valeur s'écartant de plus de deux écarts-types de la moyenne a été remplacée par la valeur moyenne. Le même processus a été appliqué pour les essais manquants (écartés dans la Section 2.2.1.1). Les valeurs atypiques remplacées représentent en moyenne $2,48\%$ des données totales et les essais manquants représentent $7,50\%$ du nombre total d'essais.

Le temps moyen de production d'un énoncé est de $1,076$ s ($\sigma = 0,13$ s), lorsque, dans la suite, on voudra donner des ordres de grandeur temporels, on passera d'une durée en temps normalisée t_N à une durée en secondes par la formule $t = t_N \times 1,076$.

Pour les données en amplitude, les valeurs des pics de fréquence fondamentale et d'intensité ne sont pas modifiés. Pour les mesures articulatoires, il est nécessaire d'opérer une transformation sur les données. En effet, on considère les voyelles (/a, e, ɜ/) et (/y, õ, ø/) comme des objets comparables en s'intéressant aux cibles articulatoires soit en ouverture (/a, e, ɜ/) soit en protrusion (/y, õ, ø/), or celles-ci sont extraites à partir de mesures différentes : un écartement entre les diodes situées sur les lèvres supérieure et inférieure pour les cibles d'ouverture et une mesure de l'avancement de la diode sur la lèvre supérieure pour les cibles de protrusion. Une façon de pallier ce problème est de normaliser ces données d'amplitude entre 0 et 1 pour chaque type de mesure : 0 correspond au minimum des réalisations des cibles articulatoires à la fois en ouverture et en protrusion, 1 correspond à leur maximum. Si on note $\mathcal{A}$ l'ensemble des mesures d'ouverture et a_j une de ces mesures, la mesure normalisée correspondante est : $a^N_j = \frac{a_j - \min \mathcal{A}}{\max \mathcal{A} - \min \mathcal{A}}$, de même pour les mesures de protrusion, si on note $\mathcal{P}$ l'ensemble des mesures de protrusion et p_j une de ces mesures, la mesure normalisée correspondante est $p^N_j = \frac{p_j - \min \mathcal{P}}{\max \mathcal{P} - \min \mathcal{P}}$. Ainsi toutes les mesures articulatoires sont restreintes dans un espace entre 0 et 1 et sont comparables. Bien entendu, pour des raisons à la fois morphologiques et de mode de production, cette normalisation n'a de sens que si on l'opère pour chaque individu (*i.e.* les ensembles $\mathcal{A}, \mathcal{P}$ sont considérés pour chaque participant et la normalisation entre 0 et 1 se fait pour chaque participant). Grossièrement, l'intérêt de cette opération est de donner des "poids égaux" à ces deux types de mesures lorsqu'on calculera des moyennes d'amplitudes des données (les cibles d'ouverture ayant des valeurs moyennes de $2,5$ cm et celles de protrusion de 3 cm, sans normalisation, les cibles de protrusion auraient un poids plus important).

2.3.2.2 Tests statistiques

Les données prétraitées sont la base du travail statistique réalisé. Toutes les analyses sont réalisées avec R [150] et les graphiques sont en majorité tracés avec le *package* ggplot2 [180]. Chaque triplet d'éléments GESTE×FOCALISATION×participant correspond plusieurs répétitions (4 phrases * 2 répétitions = 8 répétitions pour chaque triplet). Les données ont été agrégées selon les 3 dimensions GESTE, FOCALISATION, participant : les moyennes (et, le cas échéant, les écart-types) utilisées dans les analyses sont issues de cette agrégation. Les données ainsi agrégées sont directement utilisables pour des analyses à mesures répétées.

Les tests statistiques utilisés par la suite sont principalement des analyses de la variance (ANOVAs) à mesures répétées et des *t*-tests de Welch. Dans le cas des ANOVAs, la sphéricité des données a été testée par un test de Mauchly. Dans le cas d'une variance non sphérique (test de Mauchly significatif), les résultats de l'ANOVA sont corrigés par une correction de Huynh-Feldt. Cette correction est en général conseillée lorsque la sphéricité des données est assez mauvaise, on préfère en géneral utiliser la correction de Greenhouse-Geisser, plus conservative pour des valeurs estimées de ε en dessus de $0,75$. Dans les études présentées ci-après, les deux corrections donnent des résultats similaires au niveau de la significativité, seules les valeurs de la correction de Huynh-Feldt sont ici reportées.

Cette correction modifie le nombre de degrés de liberté des différents facteurs si les données sont non sphériques, par mesure de clarté, seuls les degrés de liberté non corrigés seront rapportés ici, par contre, les valeurs du F de Fischer correspondantes seront les valeurs en considérant les degrés de liberté corrigés.

Afin de tirer des conclusions plus précises sur les résultats des ANOVAs, des tests *post hoc* ont été réalisés lorsque cela était possible. Les tests *post hoc* réalisés sur le facteur FOCALISATION (2 niveaux) sont des tests t de Welch : similaire au test t de Student mais applicable aux populations à variances possiblement inégales, ce test corrige les degrés de libertés mais ici encore seuls les degrés de liberté non modifiés seront rapportés. Une correction de type Bonferroni est par ailleurs appliquée à ces résultats. Les tests *post hoc* sur le facteur GESTE (3 ou 4 niveaux) sont des tests de Tukey-Kramer modifiés par Dunnett (permettant d'utiliser des variances inégales).

Dans toute la suite, le niveau de significativité minimal retenu est $\alpha = 0,05$.

2.3.3 Résultats

2.3.3.1 Influence des facteurs d'étude sur les instants de réalisation des évènements acoustiques, articulatoires et manuels

Une ANOVA à mesures répétées à deux facteurs a été menée sur les différentes variables dépendantes normalisées et annotées tel que décrit en Section 2.2.2. Cette analyse de la variance générale permet de voir l'influence des facteurs FOCALISATION et GESTE sur les instants de production des différents points annotés. La Table 2.6 [1] montre les résultats de cette analyse de la variance. En gras apparaissent les résultats significatifs de l'analyse. On remarquera que le nombre de degrés de libertés varie dans la condition GESTE entre les variables gestuelles et les variables relatives à la parole. Ceci est dû au fait qu'il est impossible (ou plutôt, non pertinent) d'évaluer les variables gestuelles dans la condition *Parole seule*. Ainsi, pour le facteur GESTE seuls les niveaux *Pointage*, *Battement* et *Contrôle* de la variable dépendante gestuelle sont pris en compte, pour les variables dépendantes de la parole, on considère les quatre niveaux *Pointage*, *Battement*, *Contrôle* et *Parole seule*.

De manière similaire, une ANOVA sur les déviations standards des instants de production a été menée. Cette analyse a essentiellement pour but de voir si certains gestes sont moins "variables" relativement à la parole par rapport à d'autres. En effet, pour un évènement donné, plus l'écart-type de ses différentes production est faible, plus cet évènement est "stable" parmi ses réalisations. Par souci de complétude les résultats concernant les variables mesurées sur les instants de parole sont également rapportés. Ces données peuvent permettre de voir par exemple, l'influence de la production des gestes sur la stabilité des évènements de parole annotés. La Table 2.7 donne les résultats de cette ANOVA.

1. Dans cette table et dans la suite du manuscrit, les résultats tabulés significatifs sont repérés par une trame grise

		FOCALISATION	GESTE	FOCALISATION×GESTE
GESTE	Onset ($t_{P_{On}}$)	$\mathbf{F(1,9) = 83,4}$ – $\mathbf{p < 0,001}$	$\mathbf{F(2,18) = 5,7}$ – $\mathbf{p < 0,01}$	$F(2,18) = 1,2$ – $p = 0,32$
	Apex (t_{P_A})	$\mathbf{F(1,9) = 114,4}$ – $\mathbf{p < 0,001}$	$\mathbf{F(2,18) = 24,3}$ – $\mathbf{p < 0,001}$	$\mathbf{F(2,18) = 4,2}$ – $\mathbf{p < 0,05}$
	Retour (t_{P_R})	$\mathbf{F(1,9) = 99,5}$ – $\mathbf{p < 0,001}$	$F(2,18) = 0,6$ – $p = 0,48$	$\mathbf{F(2,18) = 9,9}$ – $\mathbf{p < 0,001}$
	Offset ($t_{P_{Off}}$)	$\mathbf{F(1,9) = 52,3}$ – $\mathbf{p < 0,001}$	$\mathbf{F(2,18) = 13}$ – $\mathbf{p < 0,001}$	$\mathbf{F(2,18) = 4,2}$ – $\mathbf{p < 0,05}$
	Pic de vitesse ($t_{P_{vit}}$)	$\mathbf{F(1,9) = 115,3}$ – $\mathbf{p < 0,001}$	$\mathbf{F(2,18) = 55,6}$ – $\mathbf{p < 0,001}$	$F(2,18) = 1,1$ – $p = 0,36$
PAROLE	Pic de F_0 (t_{F_0})	$\mathbf{F(1,9) = 1571,6}$ – $\mathbf{p < 0,001}$	$F(3,27) = 1,6$ – $p = 0,21$	$F(3,27) = 1,6$ – $p = 0,22$
	Pic d' intensité (t_{Int})	$\mathbf{F(1,9) = 2478,6}$ – $\mathbf{p < 0,001}$	$\mathbf{F(3,27) = 5,7}$ – $\mathbf{p < 0,05}$	$F(3,27) = 0,5$ – $p = 0,69$
	Cible articulatoire 1 (t_{CV_1})	$\mathbf{F(1,9) = 3746,1}$ – $\mathbf{p < 0,001}$	$F(3,27) = 1,1$ – $p = 0,36$	$F(3,27) = 1$ – $p = 0,39$
	Cible articulatoire 2 (t_{CV_2})	$\mathbf{F(1,9) = 2655,7}$ – $\mathbf{p < 0,001}$	$F(3,27) = 2,2$ – $p = 0,12$	$\mathbf{F(3,27) = 3}$ – $\mathbf{p < 0,05}$

TABLE 2.6 – ANOVA générale sur les données temporelles

		FOCALISATION	GESTE	FOCALISATION×GESTE
GESTE	Onset ($t_{P_{On}}$)	$F(1,9) = 1,6$ – $p = 0,23$	$\mathbf{F(2,18) = 6,9}$ – $\mathbf{p < 0,01}$	$F(2,18) = 0,08$ – $p = 0,93$
	Apex (t_{P_A})	$F(1,9) = 0,06$ – $p = 0,81$	$\mathbf{F(2,18) = 4,9}$ – $\mathbf{p < 0,05}$	$F(2,18) = 1,4$ – $p = 0,28$
	Retour (t_{P_R})	$F(1,9) = 0,5$ – $p = 0,48$	$\mathbf{F(2,18) = 5,3}$ – $\mathbf{p < 0,05}$	$F(2,18) = 1,2$ – $p = 0,31$
	Offset ($t_{P_{Off}}$)	$F(1,9) = 0,1$ – $p = 0,77$	$\mathbf{F(2,18) = 4,1}$ – $\mathbf{p < 0,05}$	$F(2,18) = 1,4$ – $p = 0,28$
	Pic de vitesse ($t_{P_{vit}}$)	$F(1,9) = 0,1$ – $p = 0,71$	$\mathbf{F(2,18) = 12,3}$ – $\mathbf{p < 0,001}$	$F(2,18) = 1,6$ – $p = 0,23$
PAROLE	Pic de F_0 (t_{F_0})	$F(1,9) = 0,44$ – $p = 0,52$	$F(3,27) = 0,95$ – $p = 0,43$	$F(3,27) = 1,31$ – $p = 0,28$
	Pic d' intensité (t_{Int})	$F(1,9) = 0,31$ – $p = 0,59$	$F(3,27) = 0,52$ – $p = 0,67$	$F(3,27) = 2,4$ – $p = 0,09$
	Cible articulatoire 1 (t_{CV_1})	$\mathbf{F(1,9) = 50,49}$ – $\mathbf{p < 0,001}$	$F(3,27) = 1,47$ – $p = 0,26$	$F(3,27) = 2,76$ – $p = 0,10$
	Cible articulatoire 2 (t_{CV_2})	$\mathbf{F(1,9) = 15,31}$ – $\mathbf{p < 0,01}$	$F(3,27) = 2,74$ – $p = 0,06$	$\mathbf{F(3,27) = 4,72}$ – $\mathbf{p < 0,01}$

TABLE 2.7 – ANOVA générale sur les écart-types des données temporelles

VARIABLES MANUELLES Le décours temporel moyen des différents gestes dans les différentes conditions de focalisation est représente en Figure 2.6. Sur cette figure, les points représentent les instants de production moyens calculés sur tous les participants et les barres d'erreurs autour de ces points ont des limites à ± 1 erreur-type des instants de production au sein des participants.

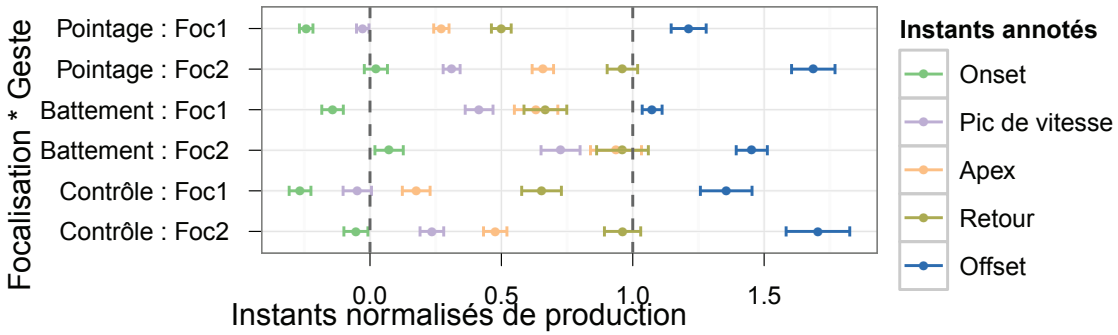


FIGURE 2.6 – Instants (moyenne et erreur-type) de production des différents moments annotés du geste

Instants de production des gestes manuels La Table 2.6 montre clairement un effet significatif du facteur FOCALISATION sur tous les instants annotés de production. La comparaison des moyennes (accessoirement, des tests *t post hoc*) montre que toutes les variables gestuelles sont produites plus tard dans la condition *Foc2* que dans la condition *Foc 1*. Ainsi, globalement, plus la focalisation est produite tard dans l'énoncé, plus le geste est également produit tard au cours de la tâche. Il est intéressant de noter qu'*a priori* ceci est vrai pour tous les gestes, qu'ils soient ou non en lien avec la focalisation (qu'il soit communicatif ou non). On peut dire que, dans cette tâche, la focalisation attire le geste.

Cette table montre également que le facteur GESTE a un effet significatif (sauf pour $t_{\mathrm{P_R}}$) sur les instants de production du geste. Ce résultat montre simplement que l'organisation temporelle des gestes est différente ce qui est facilement concevable puisque les gestes différents. Des tests *post hoc* ont été réalisés et montrent que $t^{C}_{\mathrm{P_{On}}} \lessapprox t^{P}_{\mathrm{P_{On}}} \lessapprox t^{B}_{\mathrm{P_{On}}}$ (mais ce résultat n'est pas significatif : $p > 0,05$: dans la suite, $\lessapprox$ représente des valeurs inférieures, s'approchant du seuil de significativité [1]), que $t^{P}_{\mathrm{P_{vit}}} \lessapprox t^{C}_{\mathrm{P_{vit}}} < t^{B}_{\mathrm{P_{vit}}}$ et $t^{P}_{\mathrm{P_A}} \lessapprox t^{C}_{\mathrm{P_A}} < t^{B}_{\mathrm{P_A}}$ (résultat non significatif entre *Contrôle* et *Pointage* mais significatif entre *Pointage* et *Battement*). Ceci est explicable par le fait que le geste de battement contient une phase de préparation, contrairement aux deux gestes de pointage et de contrôle. Enfin, pour l'instant de retour à la position de repos, on trouve que $t^{B}_{\mathrm{P_{Off}}} < t^{C}_{\mathrm{P_{Off}}} \lessapprox t^{P}_{\mathrm{P_{Off}}}$ (résultat significatif entre *Battement* et *Contrôle* mais non significatif entre *Pointage* et *Contrôle*).

Enfin, la colonne FOCALISATION×GESTE montre que l'interaction des deux facteurs est significative sur les instants $t_{\mathrm{P_A}}$, $t_{\mathrm{P_R}}$et $t_{\mathrm{P_{Off}}}$, ceci permet d'avancer que pour ces trois instants des gestes, la modification de l'emplacement de la focalisation entraîne des variations différentes pour les différents types de gestes *i.e.* que pour chaque geste, ces instants ne sont pas coordonnés de la même façon à la focalisation. Plus précisément, des tests post hoc sur les moyennes des décalages des différents instants de production du geste manuel montrent que le geste de pointage (au moins pour les instants $t_{\mathrm{P_A}}$, $t_{\mathrm{P_R}}$, $t_{\mathrm{P_{Off}}}$) est significativement plus décalé en passant de *Foc1* à *Foc2* que les autres types de gestes.

La Table 2.7 nous apporte par ailleurs des informations précieuses pour compléter ces premiers résultats. On n'observe aucun effet principal du facteur FOCALISATION sur les écarts-types des instants de réalisation des différents instants des gestes : les instants de production des gestes ne sont pas plus/moins variables lorsque la focalisation est située au début ou en fin d'énoncé. Cependant, on remarque un effet significatif du facteur GESTE pour tous les instants annotés. Les tests *post hoc* montrent que $\sigma_{t^{P}_{\mathrm{P_{On}}}} \lessapprox \sigma_{t^{B}_{\mathrm{P_{On}}}} < \sigma_{t^{C}_{\mathrm{P_{On}}}}$,$\sigma_{t^{P}_{\mathrm{P_{Off}}}} \lessapprox \sigma_{t^{B}_{\mathrm{P_{Off}}}} < \sigma_{t^{C}_{\mathrm{P_{Off}}}}$ *i.e.* les gestes de pointage ont des instants d'onset et d'offset moins variables que les instants d'onset et d'offset des gestes de battement (non significatif) et de contrôle (significatif). Par ailleurs, on trouve que $\sigma_{t^{P}_{\mathrm{P_A}}} < \sigma_{t^{B}_{\mathrm{P_A}}} \approx \sigma_{t^{C}_{\mathrm{P_A}}}$,$\sigma_{t^{P}_{\mathrm{P_{vit}}}} < \sigma_{t^{B}_{\mathrm{P_{vit}}}} \approx \sigma_{t^{C}_{\mathrm{P_{vit}}}}$,$\sigma_{t^{P}_{\mathrm{P_R}}} < \sigma_{t^{B}_{\mathrm{P_R}}} \approx \sigma_{t^{C}_{\mathrm{P_R}}}$ *i.e.* les gestes de pointage sont les gestes les moins variables d'un essai à l'autre, au moins pour ce qui est des instants $t_{\mathrm{P_A}}$, $t_{\mathrm{P_{vit}}}$et $t_{\mathrm{P_R}}$.

Enfin comme on peut le remarquer sur la Figure 2.6, certains instants sont moins variables que d'autres. La Table 2.8 donne les écarts-type des instants de production des gestes. Dans cette table, les écarts-types en condition *Foc1* et *Foc2* ont été calculés puis moyennés afin de ne pas prendre en compte l'effet principal de la focalisation sur les instants de production des gestes. Par ailleurs, cette table représente les ecarts-types inter-participants puisque les données ont été agrégées sur tous les participants (contrairement à l'ANOVA à mesures répétées –et donc intra-participants– en Table 2.7). Ce tableau montre que le geste de pointage est clairement moins variable relativement à la parole et d'un sujet à l'autre que les deux autres types de gestes et que les instants qui sont les moins variables sont l'onset, le pic de vitesse et l'apex pour le geste de pointage, l'onset et l'offset pour le geste de battement et l'initiation, le pic de vitesse et l'apex pour le geste de contrôle. La plus grande variabilité relativement à la parole pour les gestes de pointage (et dans une moindre mesure, de contrôle) des instants de retour et d'offset peut permettre d'argumenter sur le fait que la phase de retraction du geste n'est pas très importante dans sa relation à la parole, contrairement à la phase du stroke qui semble plus contrainte par la parole. La grande variabilité de ces instants pour le geste de battement sera abordée plus tard.

Geste	Initiation	Pic de vitesse	Apex	Retour	Offset
Pointage	0,11	0,089	0,11	0,15	0,24
Battement	0,15	0,20	0,28	0,28	0,15
Contrôle	0,14	0,15	0,15	0,23	0,35

TABLE 2.8 – Ecarts-type (entre les participants) des instants de production normalisés des gestes dans les trois conditions gestuelles

1. Le seuil de significativité est $p = 0,05$, les valeurs $p < 0,1$ sont considérées comme s'approchant du seuil de significativité

Durées des phases du geste Une autre question d'intérêt est de savoir si la production de gestes est simplement décalée dans le temps en fonction de l'emplacement de la focalisation ou bien si le geste est produit différemment. Comme on l'a vu précédemment (*cf.* Section 2.2.2.3), chaque geste est décomposable en trois phases : *stroke*, *plateau*, *stroke de retour*. Une ANOVA à mesures répétées a été menée sur les durées de ces phases ainsi que sur la durée totale de production des gestes ($t_{\mathrm{P_{Off}}} - t_{\mathrm{P_{On}}}$). Les facteurs de l'ANOVA sont toujours similaires : FOCALISATION×GESTE.

	FOCALISATION	GESTE	FOCALISATION×GESTE
Stroke	$\mathbf{F(1,9) = 45,53}$ – $\mathbf{p < 0,001}$	$\mathbf{F(2,18) = 18,71}$ – $\mathbf{p < 0,001}$	$F(2,18) = 0,45$ – $p = 0,64$
Plateau	$F(1,9) = 2,61$ – $p = 0,14$	$\mathbf{F(2,18) = 45,04}$ – $\mathbf{p < 0,001}$	$F(2,18) = 2,31$ – $p = 0,13$
Stroke de retour	$F(1,9) = 2,52$ – $p = 0,15$	$\mathbf{F(2,18) = 5,033}$ – $\mathbf{p < 0,05}$	$F(2,18) = 1,54$ – $p = 0,24$
Durée du geste	$\mathbf{F(1,9) = 13,50}$ – $\mathbf{p < 0,01}$	$\mathbf{F(2,18) = 16,92}$ – $\mathbf{p < 0,001}$	$F(2,18) = 0,94$ – $p = 0,41$

TABLE 2.9 – ANOVA sur les durées des phases du geste

La Table 2.9 montre les résultats de cette ANOVA. On observe un effet principal du facteur GESTE qui est évident : les trois types de gestes n'ont clairement pas les mêmes durées pour les différentes phases puisqu'ils sont différents. On voit que le facteur FOCALISATION a une influence sur la durée du geste et en particulier sur le *stroke* de celui-ci : le *stroke* du geste est plus long lorsque l'objet de la phrase est focalisé (condition *Foc2*). L'effet sur la durée du geste se limite à l'effet sur le *stroke a priori* puisque $t_{\mathrm{P_{Off}}} - t_{\mathrm{P_{On}}} = (t_{\mathrm{P_{Off}}} - t_{\mathrm{P_R}}) + (t_{\mathrm{P_R}} - t_{\mathrm{P_A}}) + (t_{\mathrm{P_A}} - t_{\mathrm{P_{On}}}) = d_{\mathrm{strokeRetour}} + d_{\mathrm{plateau}} + d_{\mathrm{stroke}}$ or on n'observe aucun effet sur la durée du *stroke* de retour, ni sur la durée du plateau.

En fait, en regardant les résultats par type de geste (tests t appariés sur les données par geste), on s'aperçoit que seuls les gestes de pointage et de battement ont leurs durées qui varient significativement avec l'emplacement de la focalisation et que l'allongement de la durée du *stroke* ne joue que pour partie dans cet allongement global de la durée du geste. Pour le geste de pointage, la tenue du geste est également plus courte en condition *Foc1* qu'en *Foc2* ($t(9) = -2,44$, $p < 0,05$), pour le geste de battement, c'est le *stroke* de retour qui est plus court en condition *Foc1* qu'en *Foc2* ($t(9) = -2,47$, $p < 0,05$).

VARIABLES DE LA PAROLE Les variables temporelles annotées sur le signal de parole sont une façon de représenter l'organisation temporelle interne de l'énoncé. La Table 2.6 montre un effet principal du facteur FOCALISATION sur tous les repères acoustiques ($t_{\mathrm{F_0}}$ et t_{Int}) et articulatoires ($t_{\mathrm{CV_1}}$ et $t_{\mathrm{CV_2}}$). Ceci était attendu puisque ces repères ne sont annotés que sur la partie focalisée de la parole, ainsi, ils se trouvent sur le sujet de la phrase en condition *Foc1* et sur l'objet en condition *Foc2 i.e.* ils interviennent significativement plus tard en condition *Foc2*. De manière plus intéressante, on n'observe aucun effet de la condition GESTE sur la survenue de ces indices (sauf pour l'emplacement du pic d'intensité, mais les comparaisons *post hoc* ne sont pas significatives). Ainsi, on peut dire que la production temporelle des indices acoustiques et articulatoires de la focalisation n'est pas modifiée par la production d'un geste (quel qu'il soit) *i.e.* la production d'un geste ne modifie pas l'organisation temporelle interne de l'énoncé.

La Table 2.7 n'apporte pas beaucoup de nouvelles informations, elle permet de renforcer la dernière affirmation en montrant que la production de gestes, en plus de ne pas modifier l'organisation temporelle des indices acoustiques au sein de l'énoncé, ne rend pas ces productions plus variables. D'après cette table, seul le facteur FOCALISATION joue un rôle dans la variabilité des instants de production des indices de la focalisation : les cibles articulatoires annotées sont plus variables en condition *Foc2* qu'en condition *Foc1* (p<0,001). Ce résultat est possiblement dû au fait que les cibles articulatoires liées à /e,ɜ/ sont légèrement plus variables dans leurs annotations que celles des autres phonèmes (ces cibles peuvent être considérées comme des cibles d'étirement, plutôt que d'ouverture).

2.3.3.2 Coordination des évènements manuels et de parole

Comme vu dans la Section 2.3.3.1, la focalisation "attire" la production du geste et ce, quel que soit le type de geste. Cette partie s'intéresse de façon plus précise à la manière dont se fait cette attraction. Il est intéressant de savoir quelles sont les parties du geste qui sont cooccurrentes avec la focalisation, en particulier, pour les gestes liés à la focalisation, est-ce que la partie porteuse de sens du geste (le stroke, la tenue) est cooccurrente avec la focalisation.

Une étude rapide des données montre que les instants $\mathrm{P_{On}}$ et $\mathrm{P_{Off}}$ sont le plus souvent hors de la partie focalisée (sur tous les essais, dans $94,80\%$ des cas, $\mathrm{P_{On}}$ intervient avant la focalisation et dans $98,13\%$ des cas, $\mathrm{P_{Off}}$ intervient après la partie focalisée). Ainsi, bien que ces instants soient déplacés avec le changement de position de la focalisation, ils ne sont pas cooccurrents avec la focalisation contrastive prosodique. Finalement, les trois évènements $\mathrm{P_{vit}}$, $\mathrm{P_A}$ et $\mathrm{P_R}$ sont les évènements qui peuvent *a priori* intervenir de façon concurrente à la production de focalisation. Ceci est intéressant pour les gestes de battement et de pointage puisque la partie susceptible d'intervenir en parallèle de la focalisation est alors la partie du geste qui a un rôle similaire à la focalisation (emphase, désignation) *i.e.* la partie du geste qui montre pour le pointage (tenue du geste), la partie de battement (mouvement de haut en bas) pour le geste de battement.

La Figure 2.7 donne des estimations de densité de probabilités des instants de production de ces trois évènement pour chaque type de geste et dans chaque niveau du facteur FOCALISATION. Par ailleurs, puisque la partie d'intérêt ici est la focalisation, les données ont été (sur la figure) normalisées sur la durée de la focalisation *i.e.* l'instant 0 correspond au début de la focalisation et l'instant 1 à la fin. La suite décrit la Figure 2.7 dans le cas de chaque type de geste et s'appuie sur la Table 2.10 qui détaille la proportion des instants des gestes annotés produits avant, pendant ou après la focalisation.

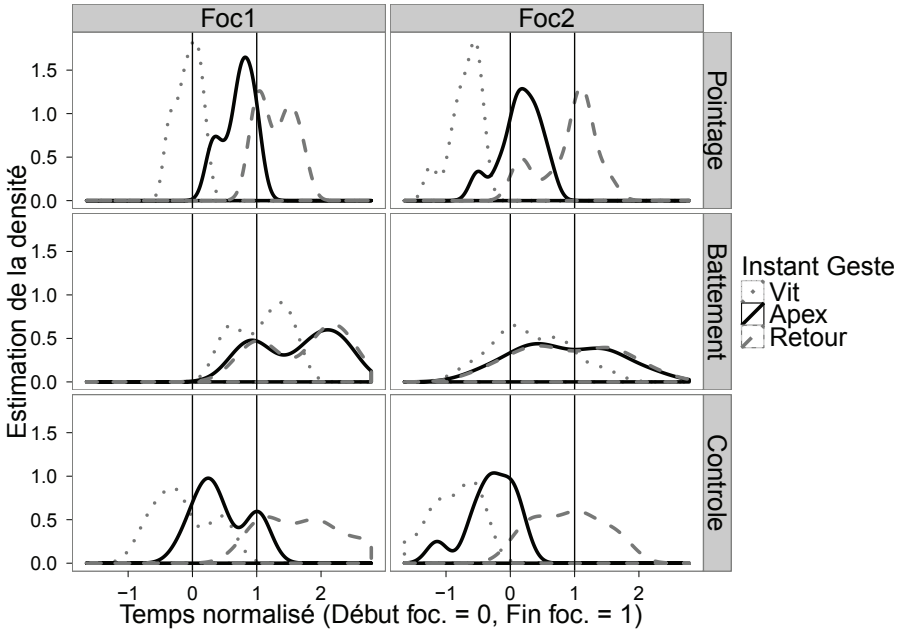


FIGURE 2.7 – Estimation de la densité de probabilité des instants de production du geste par rapport à la focalisation

Geste de pointage Le plus souvent, l'apex intervient au cours de la focalisation. Par ailleurs on voit clairement que l'apex intervient plutôt en milieu/fin de focalisation dans la condition *Foc1* et plutôt en début de focalisation dans la condition *Foc2*. En particulier, lorsque l'apex n'intervient pas dans la partie focalisée de la parole, il intervient après celle-ci dans le cas *Foc1* et avant celle-ci dans le cas *Foc2*. Le retour intervient généralement après la partie focalisée pour *Foc1* ; pour *Foc2*, le retour intervient soit dans la partie focalisée soit après celle-ci. Le pic

Geste/Foc \ Instant		Onset			Pic vitesse			Apex			Retour			Offset		
		$<$	$\in$	$>$	$<$	$\in$	$>$	$<$	$\in$	$>$	$<$	$\in$	$>$	$<$	$\in$	$>$
Pointage	*Foc1*	100	0	0	60	40	0	1,25	83,75	15	0	25	75	0	0	100
	Foc2	100	0	0	100	0	0	25	75	0	7,5	50	42,5	0	0	100
Battement	*Foc1*	80	20	0	2,5	38,75	58,75	0	26,25	73,75	0	23,75	76,25	0	0	100
	Foc2	100	0	0	28,75	53,75	17,50	12,5	50	37,50	12,5	42,5	45	0	5	95
Contrôle	*Foc1*	88,75	11,25	0	71,25	27,5	1,25	20	57,50	22,50	0	23,75	76,25	0	0	100
	Foc2	100	0	0	100	0	0	72,5	27,50	0	11,25	50	38,75	0	6,25	93,75

TABLE 2.10 – Proportions (en %) des instants du geste produits avant ($<$), pendant ($\in$), après ($>$) la focalisation

de vitesse du stroke intervient la plupart du temps juste avant la partie focalisée. Enfin, un résultat intéressant est que la tenue du geste de pointage a, la plupart du temps, une intersection non vide avec les instants de production de la focalisation (c'est le cas dans 85% des essais en condition *Foc1* et de $92,5\%$ des essais en condition *Foc2*), on observe même une tenue totalement produite dans la focalisation dans 23% du temps en *Foc1* et 40% du temps en *Foc2*. On peut donc dire que, globalement le geste de pointage est cooccurrent avec la production de la focalisation prosodique et plus précisément, il existe un recouvrement temporel non nul entre la phase du geste qui montre (tenue) et la production de focalisation.

Geste de battement Dans la Figure 2.7, on distingue assez mal les deux distributions correspondant à $\mathrm{P_A}$ et $\mathrm{P_R}$, ceci est expliqué par le fait que ces deux évènements sont très proches dans le cas du geste de battement (le geste n'est pas tenu, la tenue n'a pas de sens pour le geste de battement). La répartition des évènements gestuels est beaucoup plus variable parmi les participants que le sont ceux du geste de pointage. On observe en particulier, que l'apex du geste intervient principalement pendant la partie focalisée (en condition *Foc2*) ou après la partie focalisée (en condition *Foc1*). De la même façon, le pic de vitesse intervient surtout pendant et après la partie focalisée, mais très peu souvent avant. Globalement, on peut seulement dire que les instants de production du geste de battement sont (très) variables parmi les participants.

Geste de contrôle Ici on voit que l'apex du geste –*i.e.* le moment où les participants appuient sur le bouton– intervient souvent dans la partie focalisée de la parole dans la condition *Foc1*, mais très peu dans la condition *Foc2* où l'apex intervient le plus souvent avant la partie focalisée. Le retour du geste intervient majoritairement après la focalisation (et éventuellement dans celle-ci en condition *Foc2*). La tenue du geste (appui sur le bouton) intervient le plus souvent de manière concurrente (au moins en partie) avec la focalisation prosodique (dans $77,50\%$ des cas en condition *Foc1*, dans $88,25\%$ des cas en condition *Foc2*). On peut dire que globalement, le geste de contrôle est cooccurrent avec la production de focalisation prosodique dans le cas *Foc1* mais pas en condition *Foc2* bien qu'il soit quand même décalé vers l'élément focalisé.

2.3.3.3 Alignements possibles entre les instants des gestes manuels et ceux de la parole

Le but de cette section est de voir s'il existe des alignements entre les évènements des gestes manuels et les corrélats acoustiques/articulatoires de la focalisation. Par la suite, on dira que deux évènements sont alignés lorsque la différence temporelle de leurs instants d'occurrence est "proche" de zéro *i.e.* non significativement différente de zéro.

Afin de pouvoir traiter ce problème, les différences entre les instants du mouvement manuel ($t_{\mathrm{P_{vit}}}$, $t_{\mathrm{P_A}}$, $t_{\mathrm{P_R}}$) et les indices de la focalisation annotés sur les signaux relatifs à la parole ($t_{\mathrm{F_0}}$, t_{Int}et $t_{\mathrm{CV_1}}$, $t_{\mathrm{CV_2}}$) ont été calculées. Cette étude se limite à la recherche d'alignements avec $t_{\mathrm{P_{vit}}}$, $t_{\mathrm{P_A}}$, $t_{\mathrm{P_R}}$puisque la Section 2.3.3.2 a montré que seuls ces instants intervenaient de manière régulière en même temps que la focalisation contrastive prosodique. Pour chaque participant, les moyennes des alignements ont été calculées dans les trois conditions GESTE et les deux conditions FOCALISATION, ces alignements sont représentés en Figure 2.8. Les résultats présentés ci-dessous sont le fruit de deux types d'analyses statistiques : les alignements ont été comparés avec zéro par des tests t, les écarts-types de ces alignements ont également été comparés à zéro. Ceci permet de tester deux choses : est-ce

qu'il y a alignement (*i.e.* est-ce que deux évènements sont synchrones) et est-ce que les alignements sont réguliers (*i.e.* est-ce que deux évènements sont systématiquement séparés par un temps constant). Par abus de langage, on appelera parfois *alignement* la différence temporelle entre deux évènements, si celle-ci est *proche* de zéro.

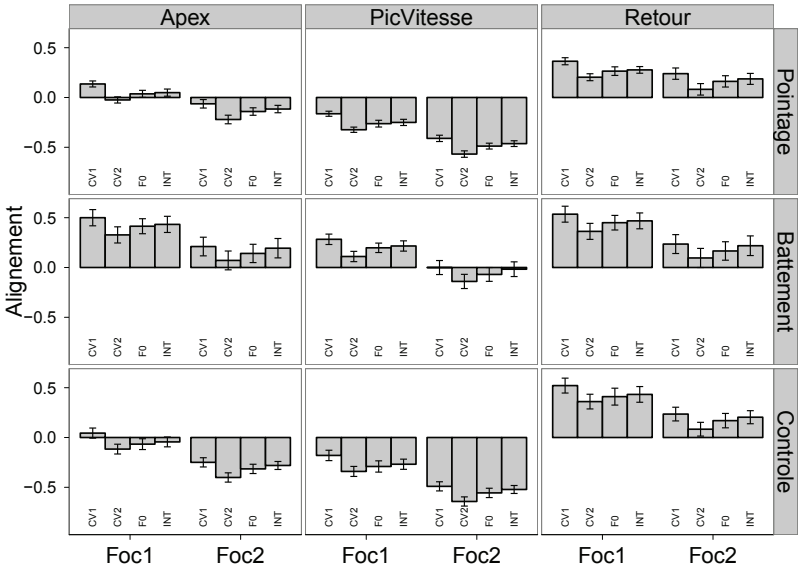


FIGURE 2.8 – Différences temporelles moyennes entre les indices de la focalisation et l'exécution des gestes

Sur la Figure 2.8 sont également représentées les erreurs-types pour chaque différence temporelle. Globalement, il est manifeste (et cela est confirmé par les tests statistiques) que le geste pour lequel les différences temporelles sont le moins variables est le geste de pointage. Ceci montre en premier lieu que c'est ce geste qui semble offrir le plus de régularité dans les alignements. L'étude des deux autres modalités gestuelles montre que le geste de battement semble être le moins régulier dans ses alignements avec la focalisation, le geste de contrôle ayant une régularité se situant entre ces deux gestes (toutes les différences sont significatives).

Geste de pointage Ce geste semble donc être celui ayant les différences temporelles les plus uniformes avec la parole. De manière plus précise, les instants les plus régulièrement proches de la parole sont $\mathrm{P_{vit}}$ et $\mathrm{P_A}$, ce qui place ces deux instants comme les candidats les plus sérieux à un éventuel alignement entre indices de la focalisation et gestes manuels. La valeur des différences temporelles est cependant plus petite (*i.e.* les deux évènements sont plus proches temporellement) dans le cas de l'apex du geste.

Statistiquement, on trouve que l'instant d'occurrence de $\mathrm{P_A}$ n'est pas différent de $t_{\mathrm{F_0}}$ ($t(9) = 1\,;\, p = 0,33$), ni de t_{Int}($t(9) = 1,4\,;\, p = 0,21$), ni de $t_{\mathrm{CV_2}}$ ($t(9) = -0,8\,;\, p = 0,45$), l'alignement le plus fin étant observé pour $\mathrm{CV_2}$ dans la condition *Foc1*. En condition *Foc2*, on trouve que $t_{\mathrm{P_A}}$ n'est pas statistiquement différent de $t_{\mathrm{CV_1}}$ ($t(9) = -1,5\,;\, p = 0,17$), on peut également noter que l'instant d'occurrence de $\mathrm{P_R}$ n'est pas différent de celui de $\mathrm{CV_2}$ ($t(9) = 1,4\,;\, p = 0,19$). Aucun alignement n'est observé pour $\mathrm{P_{vit}}$ et les évènements de la parole annotés ($p < 0,05$ pour toutes les différences temporelles).

L'ordre de grandeur de ces différences temporelles est assez faible… dans la condition *Foc1* : $t_{\mathrm{P_A}} - t_{\mathrm{CV_2}} \approx 26,48\,\mathrm{ms}$, $t_{\mathrm{P_A}} - t_{\mathrm{F_0}} \approx 39,32\,\mathrm{ms}$, $t_{\mathrm{P_A}} - t_{\mathrm{Int}} \approx 52,62\,\mathrm{ms}$; dans la condition *Foc2* : $t_{\mathrm{P_A}} - t_{\mathrm{CV_1}} \approx 68,16\,\mathrm{ms}$ et $t_{\mathrm{P_R}} - t_{\mathrm{CV_2}} \approx 88,12\,\mathrm{ms}$.

Hormis les valeurs des différences temporelles très faibles, on trouve également que les écarts types sont les plus faibles pour les différences temporelles avec les cibles articulatoires. Ainsi, même dans les cas où les deux évènements ne sont pas synchrones, l'écart de temps qui les sépare est plus systématique parmi les essais. En particulier, dans la condition *Foc1*, l'écart-type des différences temporelles est faible à la fois pour les différences temporelles entre l'apex et $\mathrm{CV_1}$ et pour l'apex et $\mathrm{CV_2}$. Ceci peut s'expliquer grâce à la Figure 2.9 sur laquelle sont représentées des estimations de densité de probabilité des instants de production de l'apex du geste de pointage dans les deux conditions *Foc1* et *Foc2*. Sur ce graphique, les instants de production de l'apex sont tracés

et les données sont mises à l'échelle de manière à ce que, pour chaque élément focalisé, le temps 0 corresponde à la première cible articulatoire de cet élément et 1 corresponde à la seconde cible articulatoire. On voit assez clairement qu'en condition *Foc1*, l'instant de production de l'apex est proche des deux instants CV_1 et CV_2 (on voit deux modes sur la courbe ce qui peut correspondre à deux stratégies différentes de réalisation). Dans la condition *Foc2*, la production de l'apex se rapproche de l'instant de réalisation de CV_1 seulement dans la plupart des cas.

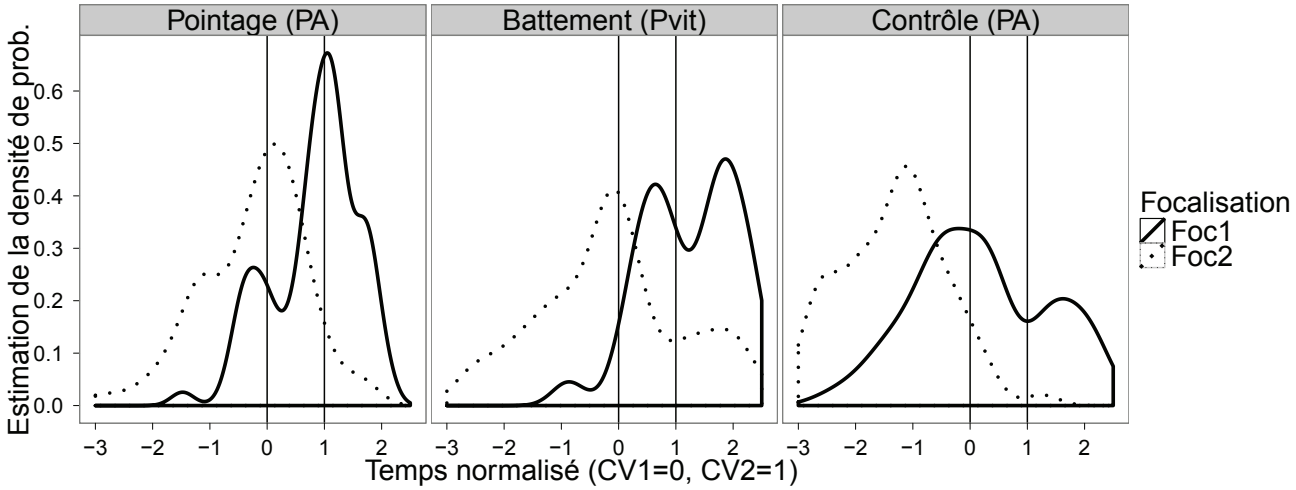


FIGURE 2.9 – Densité de probabilité estimée de l'instant de production de l'apex du pointage, pic de vitesse du battement et apex du geste de contrôle

Il y a donc une stratégie assez claire qui consiste à faire tendre la production conjointe geste de pointage + focalisation contrastive prosodique vers un alignement de l'apex avec l'une des deux cibles articulatoires de l'élément focalisé de l'énoncé.

Geste de battement Pour ce type de geste, l'évènement manuel annoté ayant les différences temporelles les plus régulières avec la parole est le pic de vitesse. Comme on le voit sur la Figure 2.8, en condition *Foc1* les tests statistiques ne donnent aucun alignement "possible" entre l'apex, le retour et les corrélats acoustiques et articulatoires de la focalisation annotés (toutes les différences temporelles sont différentes de 0, $p < 0,05$). On observe par contre un alignement entre les réalisations de $\mathrm{P}_{\mathrm{vit}}$ et de CV_2 ($t(9) = 2,1$; $p = 0,06$). En condition *Foc2*, les tests t donnent des alignements entre P_{A} et $\mathrm{P}_{\mathrm{vit}}$ et F_0 (pour P_{A} : $t(9) = 1,5$, $p = 0,160$, pour $\mathrm{P}_{\mathrm{vit}}$: $t(9) = -1$, $p = 0,340$) et Int(pour P_{A} : $t(9) = 2$, $p = 0,080$ et pour $\mathrm{P}_{\mathrm{vit}}$: $t(9) = -0,2$, $p = 0,820$) au niveau acoustique et entre P_{A} et $\mathrm{P}_{\mathrm{vit}}$ et CV_1 (pour $\mathrm{P}_{\mathrm{vit}}$: $t(9) = -0,006$, $p = 1,000$) et CV_2 (pour P_{A} : $t(9) = 0,8$, $p = 0,470$ et pour $\mathrm{P}_{\mathrm{vit}}$: $t(9) = -2$, $p = 0,080$). Ces alignements multiples en condition *Foc2* sont probablement le reflet des productions (au moins pour P_{A}) qui sont très variables à la fois entre les participants mais également au sein d'un même participant.

Ici encore on observe un alignement entre un élément du geste : $\mathrm{P}_{\mathrm{vit}}$ et les cibles articulatoires de la partie focalisée de la parole (CV_2 pour *Foc1* et CV_1 pour *Foc2*). Les différences temporelles les plus régulières observées sont celles entre $\mathrm{P}_{\mathrm{vit}}$ et F_0 en condition *Foc1* et $\mathrm{P}_{\mathrm{vit}}$ et $\mathrm{F}_0/\mathrm{CV}_1$ en *Foc2*.

Geste de contrôle Comme pour le geste de pointage, l'étude des écarts-types montre que les différences temporelles les moins variables sont celles obtenues avec l'apex du geste de contrôle (instant d'appui sur le bouton), ainsi c'est cet instant qui semble être coordonné le plus régulièrement aux instants annotés sur les signaux de parole.

En condition *Foc1*, l'apex est aligné avec F_0 ($t(9) = -1,2$, $p = 0,250$) et le pic d'intensité de la partie focalisée ($t(9) = -0,9$, $p = 0,390$) ainsi que CV_1 ($t(9) = 0,9$, $p = 0,410$). Cependant, aucun alignement

n'est trouvé en condition *Foc2* (comme suggéré par Figure 2.9 et Figure 2.8), tous les différences temporelles sont significativement différentes de 0 ($p < 0,05$). Aucun alignement n'est par ailleurs trouvé ni pour le pic de vitesse ni pour le retour (tous les différences temporelles sont différentes de 0, $p < 0,05$) sauf en condition *Foc2* où l'instant de retour est aligné avec CV_2 ($t(9) = 1,2$, $p = 0,260$).

2.3.3.4 Effets de la production conjointe de gestes manuels et de parole sur leurs productions respectives

	FOCALISATION	GESTE	FOCALISATION×GESTE
Durée de l'énoncé (non normalisée)	$\mathbf{F(1,9) = 12,1}$ – $\mathbf{p < 0,01}$	$F(3,27) = 2$ – $p = 0,14$	$F(3,27) = 0,35$ – $p = 0,79$
Durée de la partie focalisée	$\mathbf{F(1,9) = 7}$ – $\mathbf{p < 0,05}$	$F(3,27) = 0,65$ – $p = 0,59$	$F(3,27) = 1$ – $p = 0,4$
Instant d'onset vocal (non norm.)	$F(1,9) = 0,61$ – $p = 0,45$	$\mathbf{F(3,27) = 14,14}$ – $\mathbf{p < 0,001}$	$\mathbf{F(3,27) = 20,86}$ – $\mathbf{p < 0,001}$
Amplitude pic de F_0	$\mathbf{F(1,9) = 17}$ – $\mathbf{p < 0,01}$	$F(3,27) = 1,3$ – $p = 0,31$	$F(3,27) = 0,3$ – $p = 0,83$
Amplitude pic d'intensité	$\mathbf{F(1,9) = 76,2}$ – $\mathbf{p < 0,001}$	$F(3,27) = 0,8$ – $p = 0,5$	$F(3,27) = 2,2$ – $p = 0,1$
Amplitude CV_1	$\mathbf{F(1,9) = 39,7}$ – $\mathbf{p < 0,001}$	$F(3,27) = 2,5$ – $p = 0,08$	$\mathbf{F(3,27) = 3,8}$ – $\mathbf{p < 0,05}$
Amplitude CV_2	$\mathbf{F(1,9) = 128,6}$ – $\mathbf{p < 0,001}$	$\mathbf{F(3,27) = 6,6}$ – $\mathbf{p < 0,01}$	$F(3,27) = 0,9$ – $p = 0,46$

TABLE 2.11 – ANOVAs sur les durées/instants acoustiques de la parole

Durées des énoncés et de l'élément focalisé Les durées acoustiques de la parole (durée de l'enoncé et de l'élément focalisé) ont été calculées et une analyse de la variance a été menée pour étudier l'influence de la production de différents gestes et de la modification de l'emplacement de la focalisation sur ces durées. Il est évident que la durée de l'énoncé ici considérée est la durée *non normalisée* (puisque les données sont normalisées selon cette durée), on considère au contraire la durée normalisée de l'élément focalisé. Les résultats de l'ANOVA sont reportées dans la Table 2.11. Cette analyse de la variance montre un effet significatif de l'emplacement de la focalisation à la fois sur la durée de celle-ci par rapport à l'énoncé mais également sur la durée totale de la production vocale.

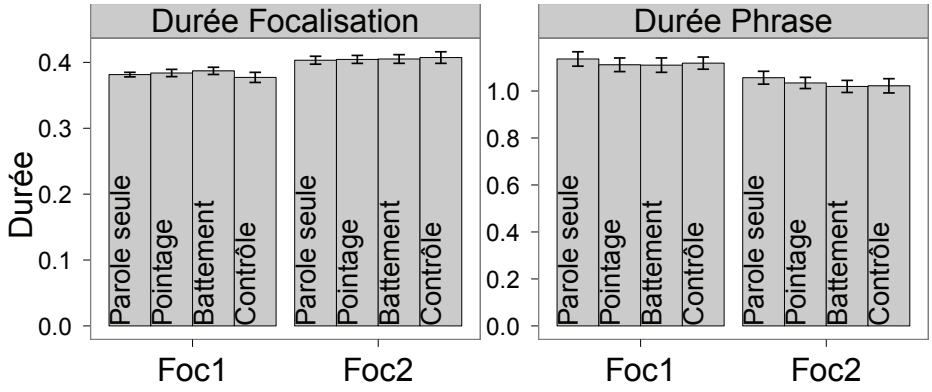


FIGURE 2.10 – Influence des facteurs sur les durées (de la focalisation, de l'énoncé) de la parole

La Figure 2.10 montre les moyennes et erreurs-types associées à l'ANOVA ci-dessus. On voit que l'effet de la focalisation sur la durée de l'énoncé ne va pas dans le même sens que l'effet de la focalisation sur la durée de l'élément focalisé. Ceci est en fait dû au fait que les données sont normalisées pour la durée de l'élément focalisé : en condition *Foc2*, l'énoncé dure moins longtemps mais la focalisation dure un temps égal à ce qu'on observe en condition *Foc1* ($t(9) = 0,43$, $p = 0,670$) ; ce qui entraîne qu'en proportion du temps total de production vocale, la focalisation est plus longue en condition *Foc2*. La moyenne de durée des énoncés en condition *Foc2* est $1,033$ s et $1,119$ s en condition *Foc1* : les énoncés produits en condition *Foc1* sont plus longs qu'en *Foc2* de 86 ms.

De manière intéressante, on note également que la production de gestes manuels ne modifie pas les durées de production (à la fois pour la durée totale de l'énoncé mais aussi pour la durée relative de la partie focalisée).

Onset de la parole Quelques études de la littérature se sont penchées sur la question de l'influence de la production manuels de gestes sur l'onset vocal de la parole associée (voir par exemple [52, 73, 114]), les résultats sont unanimes : la parole commence plus tard lorsqu'un geste est à effectuer en même temps que la parole. Cette expérience permet également de s'intéresser à cette question. Pour ce faire, l'"onset vocal" a été mesuré comme étant la différence temporelle (dans les données non normalisées) entre le début de la parole et la fin de la partie sonore du stimulus.

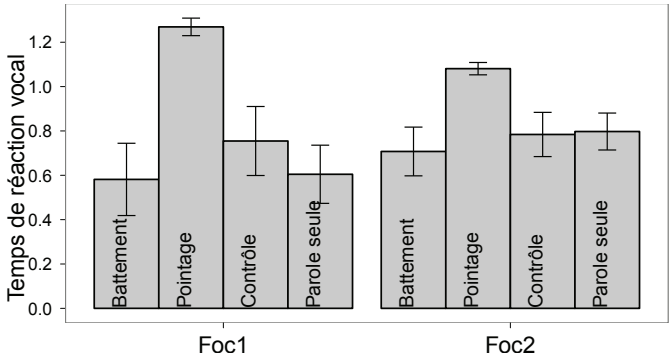


FIGURE 2.11 – Influence de la production de gestes sur l'"onset vocal"

La Figure 2.11 montre les valeurs des onsets vocaux mesurés. Une ANOVA (présentée dans la Table 2.11) montre qu'il n'y a aucun effet de l'emplacement de la focalisation sur le "temps de réaction" vocal mesuré, mais montre l'influence de la production de gestes sur ce temps. La Figure 2.11 ainsi que les tests *post hoc* menés montrent par ailleurs que le "temps de réaction" vocal est plus grand dans la condition *Pointage* que dans toutes les autres conditions ($p < 0,05$ pour les trois comparaisons) qui ont un temps de réaction similaires ($p > 0,05$ pour toutes les comparaisons). Enfin, l'ANOVA montre une interaction significative, celle-ci est due au fait que le "temps de réaction" vocal est plus court dans la condition *Foc2* que dans la condition *Foc1* pour le geste de pointage. Ce résultat peut mener à penser que la parole peut débuter alors que le geste est encore en cours de planification, mais que la planification doit être à un stade assez "avancé" pour permettre à la parole d'être exécutée *i.e.* si le geste de pointage intervient en même temps que la focalisation, en condition *Foc2*, la planification du geste peut se terminer pendant que le début de la phrase est articulée, ce qui n'est pas le cas en condition *Foc1* où le geste doit être "prêt" avant le début de la parole (la partie focalisée étant en début d'énoncé). Par ailleurs, il est intéressant de constater que l'écart-type des "onsets" mesurés est beaucoup plus faible dans la condition *Pointage* que dans toutes les autres conditions. Ceci permet d'affirmer que le début de production de la parole est fortement lié au temps supplémentaire mis pour la planification du geste de pointage (la parole commence "dès que possible" une fois la planification du geste dans un état assez avancé).

Enfin, les données des études précédentes trouvaient des décalages entre l'onset vocal de la tâche simple (parole seule) et la double tâche (pointage+parole) d'environ 100 ms pour Holender [73], de 99 ms pour Levelt et coll. [114], de 87 ms pour Feyereisen [52]. Ici le décalage entre les conditions *Parole seule* et *Pointage* est beaucoup plus important : 664 ms en moyenne en condition *Foc1* et 284 ms en moyenne en condition *Foc2*. Cette grande différence avec la littérature peut-être due au fait que pour l'expérience ici présentée, aucun signal visuel ne joue le rôle de signal go/no go : bien que la consigne demandait explicitement de répondre après l'apparition des stimuli visuels, l'apparition de celles-ci ne modifie *a priori* que peu le début de la parole en condition *Parole seule* (dans toutes les conditions, sauf le pointage) puisque la "bonne réponse" est totalement prévisible avec le stimulus auditif seul mais elle est primordiale en condition *Pointage* puisque le pointage doit désigner la bonne cible visuelle. Dans tous les cas, cette différence d'environ 400 ms (significative $t(9) = 8,08, p < 0,001$) entre le début de la parole dans la condition *Foc1* et *Foc2* pour le pointage est intéressante puisqu'elle permet de voir que la parole n'attend pas que le geste soit complètement "prêt" pour démarrer (sinon il n'y aurait pas d'effet de la focalisation).

Bien que la parole puisse être exécutée dès la fin de la présentation des stimuli audio, celle-ci "attend" que le geste soit "prêt" pour s'exécuter (c'est le cas dans le geste de pointage, puisque les trois autres conditions ont

des instants de début de production vocale similaires). Par ailleurs, le fait que la parole commence "plus tôt" dans la condition *Foc2* suggère qu'une partie de la planification du geste peut être faite pendant la réalisation de l'énoncé. La parole attend le geste s'il y a des contraintes communicatives à remplir (la cooccurrence de l'apex avec la partie focalisée dans le cas du geste de pointage), mais elle "ne perd pas de temps" à attendre que le geste soit totalement planifié afin de commencer et débute le plus tôt possible, tant que le temps restant à consacrer à la planification du geste ne met pas en péril les contraintes communicatives. La planification du geste de pointage prend *a priori* le même temps dans les deux conditions de focalisation mais les contraintes de coordination étant situées plus tard dans la condition *Foc2*, la planification du geste est terminée pendant le début de l'articulation (et le début de la parole intervient plus tôt).

Variation des amplitudes Une étude intéressante à mener est celle concernant les amplitudes des différents points annotés. La principale question posée est de savoir si la production de gestes (et leur relation avec la parole) influe sur l'amplitude des indices acoustiques et articulatoires liés à la focalisation. Les amplitudes étudiées sont ici celles des pic de fréquence fondamentale, pic d'intensité, cibles articulatoires sur la partie focalisée de la parole. Pour rappel, les données articulatoires sont normalisées de manière à pouvoir comparer les données en ouverture et en protrusion (*cf.* Section 2.3.2.1). Une analyse de la variance a été menée et est reportée dans la Table 2.11. Des histogrammes représentent les données sur la Figure 2.12.

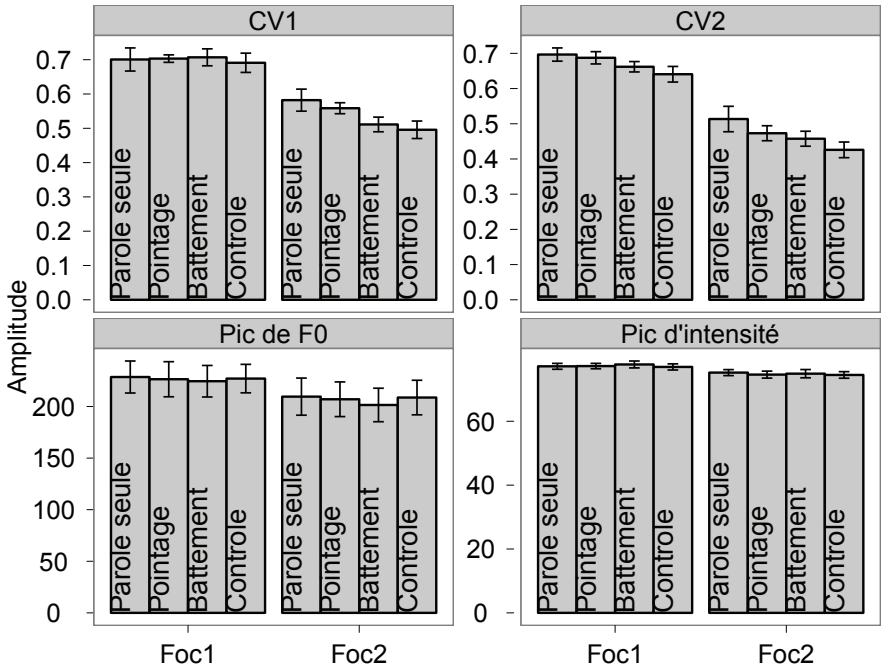


FIGURE 2.12 – Influence des facteurs sur les amplitudes des corrélats acoustiques (pic de F_0, d'intensité) et articulatoires (deux cibles articulatoires) de la focalisation

La table d'ANOVA ainsi que les histogrammes montrent que la condition de focalisation a un effet principal significatif sur les amplitudes des évènements annotés. Ceci peut être attribué au phénomène de déclinaison au sein d'un syntagme intonatif (décrit dans [44]), pour la fréquence fondamentale. L'ANOVA ne donne aucun effet de la production de gestes sur les amplitudes des corrélats acoustiques et articulatoires de la focalisation (sauf dans le cas de CV_2 mais les tests *post hoc* ne sont pas significatifs…) : il n'y a pas de différence (au niveau des amplitudes acoustiques/articulatoires) entre la condition parole seule et les conditions avec les gestes (et il n'y a pas non plus de différences entre les différents types de gestes).

2.4 Discussion

L'étude présentée dans cette partie a pour but d'étudier la co-production de gestes manuels et de parole dans le cadre de la désignation multimodale (par les gestes manuels et par la parole). Afin de décrire finement cette co-production, dix participants ont été enregistrés au cours d'une expérience permettant des mesures précises à la fois au niveau de la parole mais aussi des gestes manuels. Deux variables indépendantes ont été prises en compte : l'emplacement de la FOCALISATION qui pouvait être soit en début (*Foc1*) soit en fin (*Foc2*) de production vocale et le type de GESTE accompagnant la production vocale : Pas de geste (*Parole seule*), un geste communicatif et déictique (*Pointage*), un geste communicatif non déictique (*Battement*) et un geste non communicatif et non déictique (*Contrôle* : appui sur un bouton).

Les parties précédentes ont permis de montrer des résultats importants à la fois en ce qui concerne la coordination entre gestes manuels et parole mais aussi en ce qui concerne la modification de leurs productions respectives lors d'une production conjointe. En particulier, il a été montré que la focalisation "attirait" le geste manuel et ce, quel que soit le type de geste pris en compte : lorsque la focalisation intervient tard dans l'énoncé, alors le geste produit conjointement est exécuté tardivement. Les résultats mettent en avant une place particulière pour le geste de pointage : l'effet d'attirance, bien qu'observable pour tous les types de gestes est plus important sur le geste de pointage ; le geste de pointage est le geste ayant le plus de régularité dans ses instants de production au sein et à travers les participants ; enfin le geste de pointage est le seul geste pour lequel on observe une cooccurrence avec la focalisation dans toutes les productions. Au contraire, le geste de battement semble être cooccurrent avec la focalisation surtout lorsque la focalisation est en fin d'énoncé (sa cooccurrence dans la condition *Foc1* est plus limitée), et le geste de contrôle semble être cooccurrent avec la focalisation seulement lorsque la focalisation est en début d'énoncé (condition *Foc1*). L'étude des alignements potentiels entre la parole et les gestes manuels a montré que le geste de pointage semblait être le geste le mieux/le plus régulièrement coordonné avec les indices (principalement articulatoires) de la focalisation.

Finalement, quelques analyses ont également montré l'influence des facteurs sur la production conjointe de gestes manuels et de parole. Il a été montré l'influence de la position de la focalisation sur la durée des gestes manuels produits et ce, en particulier, à cause d'une phase de *stroke* plus courte en condition *Foc1*. L'impact de la production de gestes sur l'organisation interne de l'énoncé s'est révélé nul (tant au niveau temporel qu'au niveau des amplitudes mesurées) mais il a été noté que la production d'un geste de pointage (et seulement ce type de geste) retardait la production de parole (et la retardait encore plus en condition *Foc1* qu'en condition *Foc2*).

À la lumière de ces résultats, l'importance du lien communicatif qui existe entre gestes manuels et parole dans la coordination de ces deux modalités sera discutée avant de voir comment cette coordination se manifeste. Sera ensuite abordée le thème de l'adaptation mutuelle des productions manuelles et vocales en prenant en compte les résultats d'adaptation présentés précédemment.

2.4.1 L'importance du lien communicatif entre geste manuel et parole dans leur coordination

Comme synthétisé ci-dessus, les résultats ont en particulier montré que l'emplacement où se produit la focalisation attire la production du geste et ce, pour tous les types de gestes étudiés mais également le fait que le geste de pointage est le geste pour lequel cet effet est le plus fort. Par ailleurs, le geste de pointage est le geste le mieux coordonné à la production de focalisation et l'est de façon plus régulière que les autres gestes.

Comme avancé dans la Section 2.3.1, cette attraction de la production de gestes par la production de focalisation était attendue pour les deux gestes liés à la parole (et en particulier, à la focalisation). Cependant, aucun effet de la position de la focalisation n'était attendu sur les instants de production du geste non communicatif (*Contrôle*) : il n'y a a priori aucune raison pour que ce geste soit exécuté en même temps que la focalisation. Nonobstant, comme montré en Section 2.3.3.3, la production du geste de contrôle est légèrement décalée vers la fin de la production de parole dans la condition *Foc2*. Il a tout de même été montré que ce décalage était sans commune mesure avec le décalage subit pour le pointage (*cf.* p.50), ainsi on peut penser que le décalage du geste de contrôle est lié à un effet de couplage inter-effecteurs. Suivant l'approche dynamique de la coordination motrice comme

celle proposée par Kelso et coll. [91] pour la production de parole combinée à la production de petits tapotements sur une table, ici une hypothèse est que la production de focalisation requiert une énergie motrice plus forte qui attire tout acte moteur intervenant en parallèle (ici, tout geste manuel). Cette attraction ne fournit cependant pas une coordination fine ni régulière avec la production de focalisation : l'attraction existe mais est modérée. Ce n'est pas le cas du geste de pointage (et, dans une moindre mesure, du geste de battement) pour lequel la coordination est fine et régulière. Ceci suggère un effet supplémentaire du lien communicatif qui unit geste manuel et parole dans la coordination entre ces deux modalités : l'attraction motrice a lieu mais est renforcée par des buts communicatifs de plus haut niveau.

2.4.2 Comment s'effectue cette coordination ?

Pointage La Section 2.3.3.3 montre qu'on a dans plus de 85% des réalisations une intersection non nulle entre le l'intervalle de temps contenant le plateau du geste de pointage et l'intervalle de temps contenant la partie focalisée de la parole, plus précisément, on observe très souvent que l'apex du geste de pointage intervient au sein de la partie focalisée. Ces résultats sont en accord avec ce qui été trouvé dans les études de Rochet-Capellan et coll. [157] où une des conclusions est que la partie du geste de pointage qui montre (la tenue du geste) et la partie de la parole qui montre (l'accentuation lexicale) ont une intersection non vide.

Ces résultats sont concordants avec les hypothèses formulées précédemment puisque le pointage est cooccurrent avec la focalisation ce qui confirme dans un sens le fait que le pointage manuel et la focalisation prosodique désignent bien le même "objet" mais avec des modalités distinctes (dans l'espace réel pour le geste manuel et en parole pour la focalisation). Cela peut d'ailleurs constituer une extension des conclusions de de Ruiter [38] (et Rochet-Capellan [157]) qui ont tous deux trouvé un effet de l'emplacement de la focalisation contrastive (ou de l'accentuation lexicale) sur la coordination entre gestes manuels et parole. En effet, les résultats de ces études portaient sur des groupes de mots ou des non mots en isolation, ici le résultat est toujours valable pour des énoncés qui sont des phrases correctement construites grammaticalement.

Par ailleurs, il a été montré que l'apex était l'instant du geste de pointage le mieux (et le plus régulièrement) coordonné avec la focalisation contrastive prosodique et que les indices de la parole les mieux coordonnés à cet apex étaient des cibles articulatoires contenues dans la partie focalisée. Ainsi, la coordination gestes manuels/parole semble être réalisée par la coordination entre apex du geste de pointage et cibles articulatoires (CV_1 ou CV_2, selon l'emplacement de la focalisation), ce qui est dans la lignée des résultats de Rochet-Capellan [157] qui montre des alignements assez précis entre apex du pointage et position basse de la mâchoire pour la partie accentuée de la parole.

Finalement, la coordination ne se produit pas avec les mêmes cibles articulatoires dans les deux conditions de FOCALISATION. Lorsque la focalisation se trouve en début d'énoncé, l'apex intervient souvent en parallèle de la seconde cible articulatoire de la partie focalisée : cela permet au participant de réaliser une cooccurrence entre apex et cible articulatoire de la partie focalisée et d'optimiser le temps de réponse. Lorsque la focalisation se trouve en fin d'énoncé, l'apex intervient le plus régulièrement en synchronie avec la première cible articulatoire de la partie focalisée : la cooccurrence entre cible articulatoire et apex est réalisée et la contrainte temporelle étant moins forte, cette cible articulatoire est la première de la partie focalisée. Ainsi, on peut dire que le couplage inter-effecteurs est renforcé par des contraintes communicatives et motrices rendant la communication plus fine : l'apex du pointage est attiré par la production de focalisation (couplage moteur) et la tenue de celui-ci doit être cooccurrente avec la focalisation (contraintes communicatives) et plus précisément, l'apex doit être approximativement synchrone avec une cible articulatoire (contrainte motrice) de la partie focalisée (peu importe quelle cible au sein de la partie focalisée).

Battement Les résultats annoncés précédemment montrent que l'apex (resp. le pic de vitesse) du geste de battement intervient la plupart du temps en parallèle de la focalisation lorsque celle-ci se trouve en fin d'énoncé mais ce n'est pas le cas lorsqu'elle porte sur le début de l'énoncé… dans cette situation, l'apex (resp. le pic de vitesse) intervient en général après la partie focalisée. Ces résultats ne s'accordent pas directement avec les hypothèses formulées puisqu'une cooccurrence du battement et de la focalisation était attendue, le geste de

battement ayant un rôle pragmatique similaire à la focalisation ici (emphase).

Le pic de vitesse du *stroke* descendant semble être l'évènement du geste le mieux coordonné avec la parole (le pic de vitesse exhibe le moins de variabilité dans ses différences temporelles avec la parole). Ceci ne va pas dans le sens des résultats de Leonard et Cummins [113] pour qui l'apex semble être l'indice le plus régulièrement coordonné avec la parole (dans leur étude, l'apex utilisé est en fait le "retour" ici présenté mais les deux évènements sont très proches et ont une corrélation moyenne de $97,25\%$; $p < 0,001$), la tâche (lecture dans l'étude de Leonard et Cummins [113]) et la langue distincts peuvent expliquer ces différences, enfin les résultats ne sont pas très éloignés puisque les deux évènements $t_{\mathrm{P_A}}$ et $t_{\mathrm{P_{vit}}}$ ont une très forte corrélation ($92,95\%$; $p < 0,001$). Dans les résultats présentés précédemment, aucun alignement entre le pic de vitesse et les instants annotés de la focalisation ne sont visibles lorsque la focalisation porte sur le début de la production vocale (*Foc1*), au contraire, on observe des alignements entre $\mathrm{P_{vit}}$ et le pic de fréquence fondamentale, le pic d'intensité ainsi que la première cible articulatoire de la partie focalisée en *Foc2*. Un point important à rapprocher des résultats de Leonard et Cummins [113] (et en accord avec ces résultats) est que, dans les deux positions de focalisation, les deux évènements les plus régulièrement coordonnés (même si la différence temporelle des deux évènements n'est pas forcément proche de zéro, cette différence varie peu parmi l'ensemble des réalisations) sont le pic de vitesse du *stroke* descendant et le pic de fréquence fondamentale.

L'étude des productions montre cependant une variabilité très grande des productions à la fois intra et inter participants, ce qui mène forcément à prendre ces conclusions avec précautions. Cette grande variabilité était attendue puisqu'une grande partie des participants a fait remarquer la difficulté de produire des gestes de battement sur demande. Cependant, tous les participants étaient conscients de l'utilisation de tels gestes dans le discours à des fins d'insistance. Ainsi, ce type de gestes semble être exécuté de manière très automatisée et inconsciente lors de la production de parole et, de fait, difficile à éliciter dans des études contraintes.

Contrôle Les hypothèses concernant le geste de contrôle allaient dans le sens d'une exécution du geste peu coordonnée avec la focalisation prosodique. Les résultats pour le geste d'appui sur le bouton vont dans le sens de ces hypothèses : l'apex du geste de contrôle semble être coordonné avec $\mathrm{CV_1}$ et les pics de fréquence fondamentale et d'intensité dans la condition *Foc1* mais aucun alignement n'est trouvé dans la condition *Foc2* (seul un alignement entre $\mathrm{P_R}$ et $\mathrm{CV_2}$ est trouvé). Le geste de contrôle est coordonné de façon assez disparate avec la parole, ce qui va de pair avec le faible lien qui l'unit avec la parole. La coordination suit ici les règles d'un couplage inter-effecteur (l'effort moteur fourni pour produire la focalisation attire la production du geste manuel), mais cette coordination est irrégulière car aucune contrainte forte n'asservit la production du geste (contrairement aux autres types de gestes).

Conclusion Globalement, il a donc été argumenté le fait que la coordination entre gestes manuels et parole est influencée par les couplages inter-effecteurs et soumise à au moins deux mécanismes qui l'influencent : les contraintes motrices (co-occurrence d'évènements moteurs "importants") et les contraintes communicatives (regroupement des fonctions de désignation, d'emphase). Le geste de pointage est celui ayant une coordination la plus fine et la plus régulière avec la parole. Ceci était attendu puisque ce geste est celui qui est *a priori* le plus en lien avec la parole dans cette tâche de désignation. Les données montrent une cooccurrence presque parfaite entre le pointage et la focalisation contrastive prosodique (contrainte communicative). La contrainte motrice observée semble être une cooccurrence de l'apex du geste avec une cible articulatoire (quelconque) de la focalisation. Ainsi, afin de remplir ces deux contraintes et de réaliser au mieux la tâche demandée (réponse rapide au problème posé), le système de production s'auto-adapte. Lorsque la focalisation se trouve en début d'énoncé, la coordination se fait avec la dernière cible articulatoire possible (afin de ne pas trop retarder le début de la réponse), lorsque la focalisation se trouve en début d'énoncé, la contrainte motrice, facilement réalisable, se fait sur la première cible articulatoire. Ainsi, le système de production modifie les productions dans les deux modalités afin de répondre au mieux au problème posé en respectant les contraintes (motrice, communicative) à respecter.

2.4.3 Les systèmes de production des gestes manuels et de la parole : deux systèmes interactifs et interadaptatifs

L'expérience ici présentée permet d'étudier directement l'influence de la production de gestes sur la production de parole. Bien entendu, la meilleure façon d'étudier l'influence de la production de parole sur la production de geste aurait été d'inclure une condition "geste seul", mais une telle tâche aurait été totalement non naturelle avec le paradigme expérimental utilisé.

Si on ne peut pas étudier l'influence de la production *vs* non production de parole sur la production de gestes, on peut tout de même voir si la production de gestes s'adapte à des productions différentes de parole. En effet, le facteur FOCALISATION, ayant attrait à la parole est constitué de deux niveaux (deux emplacements différents pour la focalisation contrastive prosodique). En plus des résultats généraux qui montrent un effet de la focalisation sur l'instant de production des gestes, les résultats précédents ne montrent aucun effet de la focalisation sur la durée des phases du geste de contrôle mais des effets sur les durées (en particulier, du *stroke*) pour les gestes de battement et de pointage. Ces deux types de gestes sont – moins finement pour le battement – alignés avec la partie focalisée de la parole, ainsi en condition *Foc1*, lorsque la focalisation intervient tôt, le *stroke* se doit d'être plus court (plus rapide) afin de réaliser la coordination. Lorsque la focalisation se trouve en fin d'énoncé (Condition *Foc2*), la durée du *stroke* peut être plus longue pour réaliser la coordination. Ces résultats suggèrent une *adaptation de la production des gestes manuels à la production de la parole*.

Par ailleurs, grâce aux conditions *Parole seule* et les autres conditions GESTE, on peut quantifier l'influence de la production de gestes manuels sur la production de parole conjointe. Au niveau de l'organisation interne de l'énoncé, il a été montré que la position des différents indices mesurés (emplacement des cibles articulatoires, des pics d'intensité et de fréquence fondamentale sur la partie focalisée de l'énoncé) et celle des limites acoustiques internes n'étaient pas influencés par la production de geste (quel que soit le type de geste concerné). Il en va de même pour l'organisation globale de l'énoncé : aucun effet sur les durées à la fois de la partie focalisée et de la production de parole complète. Enfin, il a été montré que l'amplitude des différents indices de la focalisation ne subissait aucune modification avec la production de gestes manuels. Cela suggère que la production de gestes n'influence pas les amplitudes articulatoires et acoustiques des indices de la focalisation. Cette hypothèse peut sembler contradictoire avec les résultats publiés par Krahmer et Swerts [103] où la production d'un "battement visuel" (*visual beat*) – un geste manuel de battement, un haussement des sourcils ou un hochement de tête – influent à la fois sur la durée du mot produit en même temps et sur le formant F_2. Ainsi, selon les auteurs, la production d'un "battement visuel" modifie la production de la parole cooccurrente dans "la même direction" que la modification produite par la production d'une focalisation. Cependant, la tâche demandée aux participants de cette étude semble complexe (et au moins, très peu naturelle) : il était demandé aux participants de produire un "battement visuel" en même temps que la production d'un mot d'une phrase porteuse ; le mot en question était porteur (ou non, selon la condition) de focalisation. Ainsi, dans certains cas, le battement réalisé l'était sur un mot non focalisé alors qu'un autre mot de la phrase était porteur de focalisation. Comme présenté dans l'introduction, nombre d'études dans la littérature rapportent un lien entre geste de battement (manuel ou non) et focalisation et leur cooccurrence (même si les lieux de cooccurrence sont au cœur de débats). Il est donc possible que dans l'article de Krahmer et Swerts [103], les participants aient contourné la difficulté de production d'un geste de battement cooccurrent avec un mot non focalisé par la production d'un pic de fréquence fondamentale (plus faible que sur le mot focalisé, mais présent) sur le mot "non focalisé". Ainsi ce ne serait pas la production de geste de battement qui induirait directement une modification de la production de parole mais bien la gêne générée par la production de ce geste. Au contraire, l'étude présentée ici, qui ne montre pas de modifications acoustiques/articulatoires lors de la production de gestes de battement ne demande pas aux participants de produire des combinaisons incongruentes de battement/focalisation (qui peuvent être source de gêne et donc de production de pic de fréquence fondamentale pour Krahmer et Swerts [103]).

Dans tous les cas, l'étude de Krahmer et Swerts [103] montre que la production de battement –manuel, des sourcils ou de la tête– intervient naturellement en cooccurrence avec un pic de fréquence fondamentale. Ceci renforce encore une fois le lien entre geste de battement et fréquence fondamentale. Une interprétation intéressante des résultats est le fait que la production de geste de battement soit à la base d'une activité musculaire

accrue qui pourrait causer une augmentation de l'activité musculaire liée à l'articulation et donc des modifications au niveau des caractéristiques acoustiques de la parole. Cette interprétation est possiblement transposable aux résultats présentés dans ce manuscrit : il a été montré que la production de focalisation attirait la production de gestes, et ce, même pour les gestes non liés à la parole, en raisonnant en termes d'activité musculaire, on peut dire que l'activité musculaire accrue nécessaire pour la production de focalisation attire les gestes (*i.e.* l'attraction est purement motrice et indépendante du contenu du geste manuel/de la parole et du lien les unissant).

L'étude du temps d'"onset vocal" n'a montré aucune différence dans le temps mis par les participants à produire la parole parmi les conditions *Parole seule*, *Battement*, *Contrôle* mais a également montré que ces trois conditions menaient à une production de parole plus rapide que la condition *Pointage*. Cette production retardée de la parole est en accord avec la littérature (voir par exemple Feyereisen [52] ou Levelt et coll. [114]). Ces résultats ne sont *a priori* pas le fruit du hasard puisque l'étude a également montré des écarts-types plus faibles dans les "onset vocaux" pour le geste de pointage que pour les autres gestes : l'instant de début de production de parole est moins variable quand un geste de pointage l'accompagne. Par ailleurs, les résultats pour la condition *Pointage* montrent que la parole est plus retardée en condition *Foc1* (focalisation en début d'énoncé) qu'en condition *Foc2*. Il est probable que le geste de pointage nécessite plus de temps pour être planifié que les autres gestes. Il est possible que la programmation cognitive/motrice nécessaire pour la réalisation d'un pointage soit plus complexe, et donc plus longue que pour les autres gestes, puisqu'il faut choisir entre deux cibles ; il est également possible que ce soit l'apparition des stimuli visuels qui contraigne la production de geste, puisque seul le geste de pointage ne peut être préparé avant l'apparition effective des stimuli visuels. Quelle que soit l'origine du délai supplémentaire pour produire le geste de pointage, il semble que *la parole "attend" le geste de pointage* et le délai introduit dépend avec quelle partie de la parole le pointage doit être coordonné.

Ce qui précède montre bien que les systèmes de production de la parole et des gestes manuels interagissent et s'adaptent mutuellement afin de respecter au mieux des contraintes liées à des stratégies communicatives.

Des études précédentes expliquaient cette interaction par une compétition pour un *pool* de ressources communes, cette rivalité entraînant un onset vocal retardé pour la condition parole+geste par rapport à la condition parole seule (*cf.* [52, 73, 114] par exemple). Dans l'étude présentée ici, ce retard dans l'onset vocal n'est observé que pour le geste de pointage comme le remarquait Feyereisen [52] pour les gestes iconiques. Ainsi, il ne semble pas que l'onset vocal soit influencé par la présence *vs* l'absence de geste manuel et donc par la mise à disposition d'une quantité suffisante de ressources, mais bien par les délais dûs à la préparation du geste de pointage : la parole attend que le geste soit prêt afin de pouvoir réaliser la coordination nécessaire dans le cadre de la désignation. Par ailleurs, on observe un effet important de la position de la focalisation sur l'onset vocal en condition *Pointage* : le délai supplémentaire observé en condition *Pointage vs Parole seule* pour l'onset vocal est beaucoup moins important en condition *Foc2* qu'en condition *Foc1*. Ceci permet de renforcer l'hypothèse selon laquelle les deux systèmes interagissent : lorsque la focalisation se trouve en fin d'énoncé, la parole peut commencer plus tôt puisque le geste de pointage a plus de temps pour se préparer/réaliser (et être cooccurrent avec la focalisation) que lorsque la focalisation est en début d'énoncé.

Enfin, le lien qui unit geste et parole est important dans la modulation de la coordination des productions. Comme présenté dans la Section 1.3.2 il existe des études prônant une production modulaire pour les gestes manuels et la parole, bien qu'émergeant d'une seule et même origine (*cf.* [104, 38]). L'étude ici présentée montre d'une part que lorsque la parole est "lancée", la production ou non de gestes manuels ne modifie pas son organisation interne, et d'autre part : que la production de geste (ici, au moins du geste de pointage) semble s'adapter dynamiquement à la parole en fonction des ancres sémantiques/pragmatiques utilisées pour la parole (ici, en fonction de l'emplacement de la focalisation) résultant en des durées différentes pour les phases du geste et possiblement en des points d'ancrages légèrement modifiés (ici, pas la même cible articulatoire dans les deux conditions de focalisation). Ces résultats vont donc plutôt dans le sens d'une approche intégrative des systèmes de production des gestes manuels et de la parole (*cf.* [96, 142, 128, 130] par exemple).

2.5 Conclusions

Finalement, cette étude a permis de mettre au jour plusieurs processus rentrant en jeu dans la coordination gestes manuels/parole : en plus du couplage inter-effecteurs, la coordination entre les différentes modalités est régulée par des contraintes communicatives. En effet, une coordination plus fine a été observée pour des gestes communicatifs et une coordination d'autant plus fine lorsque la fonction communicative (ici, pragmatique) assurée par le geste est la même que celle assurée par la parole (ici, geste et parole ont par exemple le même rôle de désignation, par l'utilisation du pointage et de la focalisation). Il a par ailleurs été montré que l'ancrage des deux modalités semble se faire entre la position d'apex et les cibles articulatoires pour le geste de pointage, et avec le pic de vitesse du geste et les cibles articulatoires, pic de F_0 pour le geste de battement. Un retard de l'instant d'onset de la parole entre la tâche simple et double (*Parole seule vs Parole+geste*) a également été mis en avant et ce, seulement pour la production de pointage, ce retard étant d'autant plus important que la coordination entre geste manuel et parole doit être effectuée en début d'énoncé. On peut donc dire que la parole adapte son début de production à la production du geste pour que la coordination se produise au moment voulu. Enfin, il a été montré que la production de gestes s'adaptait dynamiquement à la production de parole (en particulier par la modification du temps dédié au *stroke*) afin de réaliser cette coordination voulue pour des raisons d'efficacité communicative. En résumé : les systèmes de production du geste manuel et de la parole sont en interaction et s'adaptent l'un à l'autre afin de remplir au mieux des contraintes liées à des besoins communicatifs.

Une des limites évidente de cette étude (qui est aussi un de ses points forts) est le fait que les productions soit assez contrôlées : bien que la tâche soit le plus naturel possible (scénario induisant une correction), la production concurrente des gestes peut paraître artificielle puisqu'elle n'est pas obligatoire pour effectuer la correction. Après passation de l'expérience ceci s'est surtout révélé vrai pour le geste de battement qui apparait comme difficilement réalisable "sur demande". Ainsi, une perspective assez claire de ce travail et de tester la généralisation des résultats dans un contexte plus naturel.

Chapitre 3

Étude de l'influence du lien communicatif entre gestes manuels et parole sur leur coordination

Cette seconde étude conserve le même protocole expérimental, et donc la même nature du contrôle global de la situation d'interaction, clairement une situation « de laboratoire ». Mais elle vise à élargir le contenu linguistique des stimuli, en posant une question centrale qui est celle de la nature du lien communicatif entre les gestes et la parole.

L'expérience précédente manipulait deux acteurs sémantiques (un sujet et un objet) associés dans une phrase simple au sein de laquelle le focus linguistique pouvait se déplacer de l'un à l'autre, et elle a permis de montrer comment les gestes s'adaptaient à ce déplacement. Dans l'expérience présente, les phrases ne contiendront qu'un seul acteur, décrit dans une structure linguistique de type sujet - verbe - attribut du sujet. Le focus linguistique sera déplacé du sujet à son attribut (toujours par le jeu de la prosodie), mais l'un et l'autre se rapportent au même objet physique. La question sera pour nous de savoir comment sera gérée par la gestualité cette tension entre deux éléments lexicaux (sujet et attribut) pour un seul objet physique. Les hypothèses précises seront déclinées dans une section ultérieure.

Dans cette seconde expérience, la comparaison entre les trois types de gestes étudiés précédemment (pointage, battement et geste contrôle d'appui sur un bouton) est conservée. Cette expérience est donc conçue comme un prolongement et une expansion de la précédente, dont nous tirerons à la fois des confirmations, et des précisions sur la nature intime du lien entre le geste manuel et la parole.

3.1 Protocole expérimental

3.1.1 Corpus

Le corpus utilisé dans cette expérience est composé de six phrases déclaratives en français. Les six phrases sont présentées dans la Table 3.1.

Sujet	verbe	Attribut
Le bonbon	est	rouge
Le bonbon	est	jaune
Le béret	est	rouge
Le béret	est	vert
Le ballon	est	jaune
Le ballon	est	vert

Table 3.1 – Corpus de l'expérience

Selon un modèle similaire à celui présenté dans le Chapitre 2, les phrases utilisent une structure grammaticale simple et identique : Sujet (un article défini et un nom commun) - verbe (verbe être au présent de l'indicatif) - Attribut (une couleur). Tous les mots du corpus sont composés de syllabes de type CV pour le sujet et d'une syllabe de type CVC pour l'attribut. Enfin, pour des raisons similaires à celles exposées dans le chapitre précédent, les voyelles utilisées dans ces syllabes sont soit ouvertes (/a, ɛ, e/) soit protruses (/o, u, ɔ̃/).

3.1.2 Participants, design et dispositif expérimental

Les mêmes participants que ceux ayant pris part à l'expérience présentée ci-avant ont passé cette expérience. L'ordre de passage des expériences a été contrebalancé parmi les participants. Les détails concernant ces participants sont présentés en Section 2.1.2.

De même, le design expérimental est strictement similaire à celui présenté en Section 2.1.3 : deux facteurs sont étudiés (GESTE et FOCALISATION) et ces deux facteurs ont respectivement quatre (*Parole seule, Pointage, Battement, Contrôle*) et deux (*Foc1, Foc2*) niveaux.

Enfin, le dispositif expérimental présenté en Section 2.1.4 est inchangé pour cette expérience.

3.1.3 Description de la tâche

Dans le cadre de cette expérience, un prompt audio construit sur le modèle Sujet-Verbe-Attribut est diffusé par l'intermédiaire du haut parleur. Deux stimuli visuels s'affichent ensuite sur l'écran. Ces deux dessins représentent deux objets différents avec deux couleurs différentes. L'un des deux stimuli visuels contient une propriété commune avec le prompt : soit le Sujet du prompt correspond au même élément que celui représenté par un des deux dessins, soit l'Attribut du prompt correspond à la couleur d'un des deux dessins. Le participant doit alors corriger le prompt en fonction du dessin qui lui est lié. De manière similaire à l'expérience précédente, la correction induit naturellement une focalisation contrastive prosodique. Un exemple pour chaque niveau du facteur FOCALISATION est proposé en Table 3.2

	Foc1	*Foc2*
Prompt	– Le ballon est rouge.	– Le bérêt est rouge.
Affichage		
Participant	– Le **bonbon** est rouge.	– Le béret est **jaune**.

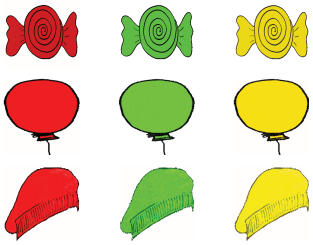
TABLE 3.2 – Exemples de stimuli et de production de parole du participant

Afin d'éviter toute confusion, le stimulus visuel distracteur représentait un objet différent à la fois de celui évoqué dans le prompt et de l'autre stimulus visuel et la couleur de celui-ci était également différente à la fois de la couleur évoquée dans le prompt et de la couleur de l'autre stimulus visuel.

Les deux conditions de FOCALISATION sont réalisées grâce à l'emplacement dans la phrase de la propriété partagée par le prompt et le stimulus visuel : lorsque la couleur évoquée dans le prompt est similaire à l'un des deux objets présentés à l'écran, la correction porte sur le type d'objet évoqué dans le prompt *i.e.* condition *Foc1* ; inversement, lorsque c'est l'objet du prompt qui est similaire à l'un des deux objets représentés à l'écran, la correction porte sur la couleur de celui-ci *i.e.* condition *Foc2*. L'ordre de présentation des prompts ainsi que l'emplacement des différents stimuli visuels (en particulier : du stimulus visuel non distracteur) ont été randomisés à la fois parmi les participants et au sein de chaque bloc. Une liste des images utilisées est fournie en Figure 3.1. L'ensemble des stimuli visuels (distracteurs ou non) est tiré dans un même jeu d'images.

Comme présenté en Section 2.1.5, une courte séquence d'entraînement a été réalisée afin de familiariser les participants avec les stimuli visuels et il était demandé aux participants de garder leur index droit sur la position de repos à tout moment hormis les instants où ils devaient produire un geste.

Un exemple complet de consignes présentées aux participants est donné en Figure A.2, p.158. Ici encore les consignes sont différentes pour chaque niveau du facteur GESTE. Lorsque la consigne requiert la production d'un



de haut en bas : un bonbon, un ballon, un béret. De gauche à droite : rouge, vert, jaune.

FIGURE 3.1 – Stimuli visuels de l'expérience

geste manuel (*cf.* Figure A.2b), la description sommaire des gestes qui est donnée est similaire à celle fournie en Section 2.1.5.

3.1.4 Acquisition des données

L'acquisition des données est totalement similaire à ce qui a été présenté en Section 2.1.6, à la fois pour les données gestuelles (capture de mouvements) et pour les données acoustiques (enregistrement audio).

3.2 Traitement des données et méthodes d'analyses

Le traitement des données est similaire à ce qui a été présenté en Section 2.2. Les analyses effectuées sur les données sont semblables à celles réalisées en Section 2.3.

3.2.1 Prédictions

Comme montré ci-dessus, cette expérience s'articule autour un protocole très similaire à l'expérience présentée précédemment. Néanmoins, les prédictions en sont différentes comme cela est exposé ci-dessous.

Dans le cas de la condition *Pointage*, dans l'expérience précédente (*cf.* Section 2.1.5), le mot focalisé dans l'énoncé correspond à l'objet désigné par le geste de pointage (ou du moins, à une représentation visuelle de cet objet). Plus précisément, la partie focalisée de la parole représente *exactement* ce qui est désigné par le geste de pointage (par exemple, *Foc1* : « **Baba** tient le bébé. » et le geste de pointage vers le personnage Baba ; *Foc2* : « Baba tient le **bébé** » et pointage vers un dessin représentant un bébé).

Dans l'expérience ici présentée, dans la condition *Pointage*, ce qui est désigné par le geste de pointage correspond au contenu complet de l'énoncé prononcé par le participant (par exemple, le participant dit « le **bonbon** est rouge » et pointe vers une image de bonbon rouge). Or, comme présenté en Section 3.1.2, la focalisation contrastive prosodique porte soit sur le sujet (*Foc1*) soit sur l'attribut (*Foc2*). Le pointage et la focalisation ne désignent donc pas strictement la même chose : ce qui est désigné par la focalisation (« bonbon ») n'est qu'une partie de ce qui est désigné par le pointage (un bonbon rouge). Un point important découlant de cette modification est que, pour une phrase donnée, le stimulus visuel désigné par le geste de pointage est le même dans les deux niveau du facteur *focalisation* (par exemple, *Foc1* : « Le **ballon** est rouge. » et pointage vers une image de ballon rouge ; *Foc2* : « Le ballon est **rouge**. » et pointage vers une image de ballon rouge).

Par la suite, on appellera *lien communicatif* entre geste manuel et la parole le lien qui existe entre les rôles respectifs du geste et de la focalisation tel que décrit ci-dessus : dans cette expérience, le seul changement important avec l'expérience précédente est donc la modification du lien communicatif entre geste manuel de pointage et parole.

Les prédictions faites sont en relation avec ce qui a été trouvé dans l'expérience précédente, en effet la tâche est globalement identique, mais les modifications dans le protocole mènent à prévoir de légères adaptations.

Pour le geste de contrôle (appui sur un bouton, non lié à la parole), le changement introduit dans la tâche ne devrait *a priori* pas induire de changement dans la coordination entre geste manuel et parole puisque le lien liant

les deux modalités est similaire à l'expérience précédente (aucun lien).

Pour le geste de battement, de façon similaire, le changement de tâche n'induira *a priori* aucun changement dans la coordination entre geste manuel et parole : ce qui unit geste de battement et focalisation est la mise en relief, or aucun changement dans ce sens n'est fait entre les deux expériences.

On devrait cependant observer une différence pour la condition *Pointage* : c'est le seul geste pour lequel le lien communicatif avec la parole a été modifié. Ce qui est désigné par le geste correspond au contenu complet de la phrase prononcée et ce qui est désigné par la focalisation n'est qu'une partie de la phrase prononcée. L'adaptation de la coordination entre geste manuel et parole peut se faire de deux façons différentes :

- Si c'est la fonction de désignation qui conditionne la coordination, alors aucun changement par rapport à l'expérience précédente n'est attendu : la focalisation devrait attirer le geste de pointage et donc le facteur FOCALISATION devrait avoir un effet significatif sur les instants de production du geste.
- Si c'est la correspondance entre l'élément désigné par le geste manuel et l'élément désigné par la parole, alors le comportement sera différent de celui de l'expérience précédente : l'élément désigné par le geste manuel étant le même dans les deux conditions *Foc1* et *Foc2*, la production de geste ne devrait pas être influencée par le niveau du facteur FOCALISATION *i.e.* le geste devrait être coordonné de la même façon que la focalisation porte sur le début de l'énoncé ou sur la fin de celui-ci.

Le paradigme ici présenté devrait de plus permettre de renforcer les résultats de l'expérience précédente quant aux lieux possibles de coordination entre les deux modalités.

3.3 Résultats

Les résultats de cette expérience sont intéressants selon deux points de vue. Cette partie décrira dans un premier temps les production d'une façon similaire à ce qui a été réalisé pour l'expérience précédente afin de valider/nuancer les résultats antérieurs, les productions des deux expériences seront ensuite comparées (ce qui est possible puisque les participants sont les mêmes).

La durée moyenne d'un énoncé est de $1,084\,\text{s}$ ($\sigma = 0,13\,\text{s}$). Ainsi, lorsque dans la suite on voudra donner des ordres de grandeur temporels, on passera d'une durée en temps normalisée t_N à une durée en secondes par la formule $t = t_N \times 1,084$.

3.3.1 Influence des facteurs sur les instants de production des événements manuels et de parole annotés

Deux ANOVAs à mesures répétées ont été menées à la fois sur les instants de production des gestes manuels et de la parole et sur les écarts-types de leurs productions. Ces deux analyses sont présentées dans les Table 3.3 et Table 3.4.

		FOCALISATION	GESTE	FOCALISATION×GESTE
GESTE	Onset ($t_{\text{P}_{\text{On}}}$)	$\mathbf{F(1,9) = 8,67}$ – $\mathbf{p < 0,05}$	$\mathbf{F(2,18) = 9,80}$ – $\mathbf{p < 0,01}$	$F(2,18) = 1,83$ – $p = 0,19$
	Apex ($t_{\text{P}_{\text{A}}}$)	$\mathbf{F(1,9) = 11,34}$ – $\mathbf{p < 0,01}$	$\mathbf{F(2,18) = 26,39}$ – $\mathbf{p < 0,001}$	$F(2,18) = 0,323$ – $p = 0,73$
	Retour ($t_{\text{P}_{\text{R}}}$)	$\mathbf{F(1,9) = 15,14}$ – $\mathbf{p < 0,01}$	$F(2,18) = 1,39$ – $p = 0,27$	$F(2,18) = 2,83$ – $p = 0,085$
	Offset ($t_{\text{P}_{\text{Off}}}$)	$\mathbf{F(1,9) = 12,76}$ – $\mathbf{p < 0,01}$	$\mathbf{F(2,18) = 13,59}$ – $\mathbf{p < 0,001}$	$F(2,18) = 0,32$ – $p = 0,73$
	Pic de vitesse ($t_{\text{P}_{\text{vit}}}$)	$\mathbf{F(1,9) = 11,14}$ – $\mathbf{p < 0,01}$	$\mathbf{F(2,18) = 68,87}$ – $\mathbf{p < 0,001}$	$F(2,18) = 1,47$ – $p = 0,25$
PAROLE	Pic de F_0 (t_{F_0})	$\mathbf{F(1,9) = 1918,19}$ – $\mathbf{p < 0,001}$	$F(3,27) = 2,28$ – $p = 0,10$	$F(3,27) = 2,29$ – $p = 0,44$
	Pic d' intensité (t_{Int})	$\mathbf{F(1,9) = 1481,33}$ – $\mathbf{p < 0,001}$	$F(3,27) = 1,17$ – $p = 0,33$	$F(3,27) = 1,17$ – $p = 0,33$
	Cible articulatoire 1 (t_{CV_1})	$\mathbf{F(1,9) = 2356,31}$ – $\mathbf{p < 0,001}$	$F(3,27) = 0,73$ – $p = 0,49$	$F(3,27) = 1,16$ – $p = 0,34$
	Cible articulatoire 2 (t_{CV_2})	$\mathbf{F(1,9) = 559,37}$ – $\mathbf{p < 0,001}$	$F(3,27) = 0,92$ – $p = 0,44$	$F(3,27) = 0,84$ – $p = 0,48$

TABLE 3.3 – ANOVAs générales sur les données temporelles (en parole et pour les gestes manuels)

On observe un effet principal du facteur *focalisation* sur toutes les données temporelles et un effet du facteur GESTE limité aux données temporelles manuelles (résultats similaires à l'expérience précédente). Une étude

		FOCALISATION	GESTE	FOCALISATION×GESTE
GESTE	Onset ($t_{P_{On}}$)	$F(1,9) = 0,11$ – $p = 0,74$	$F(2,18) = 0,88$ – $p = 0,43$	$F(2,18) = 0,48$ – $p = 0,62$
	Apex (t_{P_A})	$F(1,9) = 1,53$ – $p = 0,25$	$F(2,18) = 5,40$ – $p = 0,33$	$F(2,18) = 0,51$ – $p = 0,61$
	Retour (t_{P_R})	$F(1,9) = 0,24$ – $p = 0,64$	$\mathbf{F(2,18) = 6,31}$ – $\mathbf{p < 0,05}$	$F(2,18) = 0,11$ – $p = 0,89$
	Offset ($t_{P_{Off}}$)	$\mathbf{F(1,9) = 9,51}$ – $\mathbf{p < 0,01}$	$F(2,18) = 1,90$ – $p = 0,18$	$F(2,18) = 1,81$ – $p = 0,19$
	Pic de vitesse ($t_{P_{vit}}$)	$F(1,9) = 0,80$ – $p = 0,39$	$\mathbf{F(2,18) = 6,81}$ – $\mathbf{p < 0,01}$	$F(2,18) = 0,28$ – $p = 0,76$
PAROLE	Pic de F_0 (t_{F_0})	$F(1,9) = 0,033$ – $p = 0,86$	$F(3,27) = 1,02$ – $p = 0,40$	$\mathbf{F(3,27) = 3,57}$ – $\mathbf{p < 0,05}$
	Pic d' intensité (t_{Int})	$F(1,9) = 0,18$ – $p = 0,68$	$F(3,27) = 1,10$ – $p = 0,36$	$F(3,27) = 2,76$ – $p = 0,061$
	Cible articulatoire 1 (t_{CV_1})	$\mathbf{F(1,9) = 504,90}$ – $\mathbf{p < 0,001}$	$F(3,27) = 1,58$ – $p = 0,22$	$F(3,27) = 1,06$ – $p = 0,38$
	Cible articulatoire 2 (t_{CV_2})	$\mathbf{F(1,9) = 191,88}$ – $\mathbf{p < 0,001}$	$F(3,27) = 1,48$ – $p = 0,24$	$\mathbf{F(3,27) = 3,42}$ – $\mathbf{p < 0,05}$

TABLE 3.4 – ANOVAs générales sur les écart-types des données temporelles (en parole et pour les gestes manuels)

détaillée est proposée ci-dessous.

3.3.1.1 Variables manuelles

Instants de production des gestes manuels La Figure 3.2 donne un aperçu des instants de production des évènements manuels dans les différentes combinaisons de conditions.

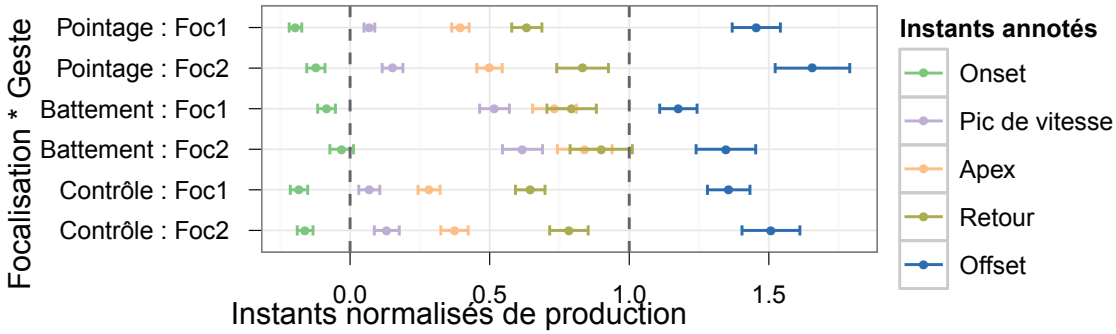


FIGURE 3.2 – Densité de probabilité estimée pour les instants de production du pic de vitesse, de l'apex et du retour annotés sur les gestes manuels

On observe un effet significatif de l'emplacement de la focalisation pour toutes les variables manuelles : en moyenne, tous les évènements manuels sont produits plus tard dans la condition *Foc2* que dans la condition *Foc1*. Cependant, les tests *post hoc* ne donnent une différence significative entre les deux conditions que pour les instants t_{P_R} et $t_{P_{Off}}$, pour les autres évènements manuels, les valeurs de t obtenues n'atteignent pas le seuil de significativité. Globalement, on peut donc toujours dire que la focalisation attire le geste mais dans une moindre mesure que dans l'expérience précédente (comme suggéré par les Figure 3.2 et Figure 2.6), en particulier, on observe peu d'influence de la focalisation sur l'instant d'onset du geste.

On observe un effet significatif principal du facteur GESTE sur les évènements manuels (sauf le retour), ce qui est normal puisque les gestes sont de type différent, il est à noter que la non-influence du facteur GESTE sur l'instant de retour était déjà observée dans l'expérience précédente. Les tests *post hoc* donnent des résultats similaires à ce qui a été évoqué dans l'expérience sus-citée : $t^C_{P_{On}} \lessapprox t^P_{P_{On}} < t^B_{P_{On}}$ (le geste de battement débute plus tard que les deux autres gestes), $t^C_{P_{vit}} \lessapprox t^P_{P_{vit}} < t^B_{P_{vit}}$, $t^C_{P_A} < t^P_{P_A} < t^B_{P_A}$ (les temps d'apex sont significativement différents pour tous les gestes) et $t^B_{P_{Off}} < t^C_{P_{Off}} \lessapprox t^P_{P_{Off}}$ (le geste de battement se termine avant les deux autres gestes).

Aucun effet de l'interaction des deux facteurs n'est observé sur les évènements manuels, pour rappel, dans l'expérience précédente, l'effet d'interaction résultait du fait que le geste de pointage était plus influencé par le

facteur FOCALISATION. Ainsi, dans cette expérience, tous les gestes semblent être affectés de la même façon par la variation de position de la focalisation.

Geste	Initiation	Pic de vitesse	Apex	Retour	Offset
Pointage	0,090	0,09	0,12	0,23	0,34
Battement	0,116	0,20	0,29	0,32	0,27
Contrôle	0,086	0,13	0,14	0,18	0,28

TABLE 3.5 – Ecarts-type inter-énoncé moyen des instants de production normalisés des gestes

L'ANOVA à mesures répétées sur les écarts-types des mesures manuelles (Table 3.4 et voir Table 3.5 pour les valeurs numériques de l'écart-type inter-énoncés) montre un effet de l'emplacement de la focalisation sur les variations des productions de l'instant d'offset (les instants d'offset sont moins variables en condition *Foc1*). Contrairement à l'expérience précédente, le facteur GESTE n'a pas d'effet sur les variations de production de tous les évènements manuels : seuls subsistent un effet sur les instants de production de retour ($\sigma_{t^P_{\mathrm{P_R}}} < \sigma_{t^C_{\mathrm{P_R}}} \approx \sigma_{t^B_{\mathrm{P_R}}}$) et du pic de vitesse ($\sigma_{t^P_{\mathrm{P_{vit}}}} < \sigma_{t^C_{\mathrm{P_{vit}}}} \approx \sigma_{t^B_{\mathrm{P_{vit}}}}$). On remarque alors que, bien que les écarts-types observés soient différents pour les instants du pic de vitesse et du début du *stroke* de retour, ceux de l'apex (qui est l'instant entre $t_{\mathrm{P_{vit}}}$ et $t_{\mathrm{P_R}}$) ne sont pas influencés par le type de geste. Globalement, les instants de production du geste de contrôle sont moins variables que dans l'expérience précédente, ce qui n'est pas le cas pour les deux autres types de gestes qui ont des instants de production aussi variables.

Durées des phases des gestes manuels Comme précédemment, une ANOVA à mesures répétées sur les durées des différentes phases des gestes a été menée. Le tableau de résultats obtenu est représenté dans la Table 3.6.

	FOCALISATION	GESTE	FOCALISATION×GESTE
Stroke	$\mathbf{F(1,9) = 7,92}$ - $\mathbf{p < 0,05}$	$\mathbf{F(2,18) = 18,61}$ - $\mathbf{p < 0,001}$	$F(2,18) = 0,60$ - $p = 0,56$
Plateau	$\mathbf{F(1,9) = 11,86}$ - $\mathbf{p < 0,01}$	$\mathbf{F(2,18) = 32,59}$ - $\mathbf{p < 0,001}$	$\mathbf{F(2,18) = 8,59}$ - $\mathbf{p < 0,05}$
Stroke de retour	$F(1,9) = 2,86$ - $p = 0,12$	$\mathbf{F(2,18) = 12,65}$ - $\mathbf{p < 0,001}$	$F(2,18) = 2,23$ - $p = 0,14$
Durée du geste	$\mathbf{F(1,9) = 10,10}$ - $\mathbf{p < 0,05}$	$\mathbf{F(2,18) = 24,77}$ - $\mathbf{p < 0,001}$	$F(2,18) = 0,068$ - $p = 0,93$

TABLE 3.6 – ANOVAs sur les durées des phases du geste

Cette table montre un effet du facteur GESTE sur les durées des gestes manuels, cela est dû au fait que les gestes ne sont pas exécutés de la même façon (résultat évident). Un effet principal intéressant est l'effet de la focalisation sur les durées globales des gestes. Les tests *post hoc* montrent que tous les gestes sont plus courts en condition *Foc1* qu'en condition *Foc2*. Plus précisément, pour le geste de pointage, c'est surtout la durée de la tenue du geste qui est plus courte en condition *Foc1* ($t(9) = -3,13$, $p < 0,05$), pour le geste de contrôle, le *stroke* ainsi que le plateau sont plus courts ($t(9) = -3,00$, $p < 0,05$ et $t(9) = -3,80$, $p < 0,01$), enfin pour le geste de battement, le raccourcissement est distribué dans les différentes phases du geste mais aucune phase n'est significativement plus courte en condition *Foc1*. Dans l'ANOVA, on voit également un effet de l'interaction FOCALISATION×GESTE sur la durée de la tenue : ceci est dû au fait que l'effet de la focalisation a une importance différente pour chaque type de geste et cet effet est le plus prononcé pour le geste de pointage ($+95$ ms en passant de *Foc1* à *Foc2*, $+42$ ms pour le contrôle et -6 ms pour le battement (non significatif pour le battement)). Finalement, pour le geste de pointage, contrairement à l'expérience précédente, il y a peu de différences pour l'instant d'*onset* du geste mais lorsque la focalisation porte sur la fin de la phrase, la main semble avoir tendance à "attendre" que la désignation en parole soit réalisée.

3.3.1.2 Variables de la parole

Pour les données relatives à la parole, on observe évidemment un effet de la focalisation sur les instants de production des pics de fréquence fondamentale et d'intensité et des cibles articulatoires : ces indices de la focalisation sont annotés sur la partie focalisée de l'énoncé et donc, ces indices arrivent plus tard en condition *Foc2* qu'en *Foc1*.

Tout comme dans l'expérience précédente, on n'observe aucun effet de la condition GESTE sur les instants de production de ces indices. On peut donc encore ici affirmer que la production d'un geste manuel ne modifie pas l'organisation temporelle interne de l'énoncé. Ceci est valable ici encore pour tous les types de gestes.

L'étude des écarts-types ne montre qu'un effet de la focalisation sur les instants des cibles articulatoires. Ceci pourrait être un biais de l'annotation (possible pour la seconde cible articulatoire, car celle-ci n'est pas évidente en condition *Foc2*), mais il est aussi possible qu'il soit le fruit des phonèmes qui précèdent ces cibles articulatoires : dans le cas de *Foc1* seules des plosives (courtes et donc ayant une durée moins variable) sont utilisées pour l'attaque des mots, dans le cas *Foc2*, les premiers phonèmes sont des spirantes et des fricatives (moins précises en durée).

3.3.1.3 Cooccurrence des gestes et de la focalisation contrastive prosodique

Comme présenté en Section 2.3.3.2, on peut se demander si le geste manuel est synchrone avec la production de focalisation. La Table 3.7 permet de voir la proportion des évènements manuels qui sont produit en même temps que la focalisation.

Dans cette expérience encore, mais dans une moindre mesure, les instants $\mathrm{P_{On}}$ et $\mathrm{P_{Off}}$ sont le plus souvent en dehors de la partie focalisée de la parole. Globalement ces indices sont donc non cooccurrents avec la focalisation dans l'immense majorité des cas, on notera de façon interessante que les gestes englobent le plus souvent la focalisation ($\mathrm{P_{On}}$ intervient majoritairement avant la focalisation, $\mathrm{P_{Off}}$ après celle-ci). De la même façon, les autres évènements manuels sont assez régulièrement cooccurrents avec la focalisation. Cependant en comparant cette table avec celle présentée pour l'expérience précédente (Table 2.10), on s'aperçoit que globalement, les gestes manuels sont produits plus tard que dans l'expérience précédente en condition *Foc1* et moins tard en condition *Foc2* : en condition *Foc1*, il y a plus de gestes dont $\mathrm{P_{vit}}$, $\mathrm{P_A}$, $\mathrm{P_R}$ interviennent après (ou pendant) la focalisation que dans l'expérience précédente, en condition *Foc2*, il y a plus de gestes dont ces trois instants interviennent avant (ou pendant) la focalisation.

Geste/Foc \ Instant		Onset			Pic vitesse			Apex			Retour			Offset		
		$<$	$\in$	$>$	$<$	$\in$	$>$	$<$	$\in$	$>$	$<$	$\in$	$>$	$<$	$\in$	$>$
Pointage	*Foc1*	100	0	0	74,17	25,83	0	0	81,67	18,33	0	30,83	69,17	0	0	100
	Foc2	100	0	0	100	0	0	52,50	47,50	0	22,50	53,33	24,17	0	10,83	89,17
Battement	*Foc1*	93,33	6,67	0	0	53,33	46,67	0	25	75	0	23,33	76,67	0	0	100
	Foc2	100	0	0	40,83	48,33	10,83	19,17	53,33	27,50	16,67	48,33	35	0,83	17,50	81,67
Contrôle	*Foc1*	98,33	1,67	0	64,17	35,83	0	13,33	77,50	9,17	0	24,17	75,83	0	0	100
	Foc2	100	0	0	100	0	0	78,33	21,67	0	20	64,17	15,83	0	10,83	89,17

TABLE 3.7 – Proportions (en %) des instants du geste produits avant ($<$), pendant ($\in$), après ($>$) la focalisation

On peut également tracer les densités de probabilité des instants de production des gestes comme en Figure 2.7. Cela donne des graphes similaires à ce qui était montré dans l'expérience précédente et qui sont présentés en Figure 3.3. Cependant on peut faire la même remarque que ci-dessus : toutes les densités de probabilités sont légèrement décalées vers la droite (*i.e.* production plus tardive) en condition *Foc1* et sensiblement décalées vers la gauche (*i.e.* production plus tôt) en condition *Foc2*.

Globalement :

– Pour le pointage, l'apex intervient souvent en même temps que la focalisation en condition *Foc1* ($81,67\%$) mais pas en *Foc2* ($47,50\%$) où il intervient majoritairement avant la focalisation. Le plateau du geste de pointage est exécuté le plus souvent sur un laps de temps correspondant au moins en partie à la partie focalisée de l'énoncé ($81,67\%$ du temps en *Foc1*, $77,50\%$ en condition *Foc2*), ce plateau est totalement

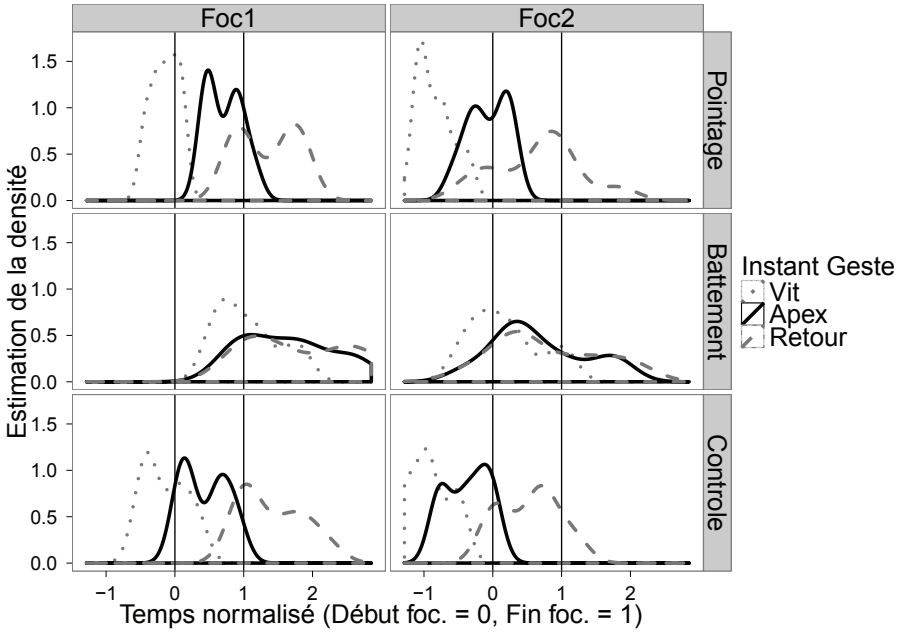


FIGURE 3.3 – Instants de production du geste par rapport à la focalisation

contenu dans la partie focalisée de la parole dans seulement $30,83\%$ des cas en condition *Foc1* et dans 30% des cas en condition *Foc2*.

- Pour le battement, les productions sont très variables et l'apex intervient le plus souvent dans la partie focalisée de la parole en condition *Foc2* et après celle-ci en condition *Foc1*. Ces résultats sont assez semblables à l'expérience précédente.
- Pour le geste de contrôle, l'apex intervient souvent dans la partie focalisée de la parole en condition *Foc1* ($77,50\%$ des cas) mais pas en condition *Foc2* où c'est plutôt le retour qui intervient dans la partie focalisée ($64,14\%$ des cas). De la même façon que pour le battement, les résultats sont proches de ce qui était observé dans l'expérience précédente.

Il est intéressant de noter que les instants de production semblent varier très peu entre les deux expériences dans la condition *Foc1*. Cependant, on observe bien un décalage entre les productions dans la condition *Foc2*. En particulier, tous les évènements du geste manuel sont produits plus tôt par rapport à la focalisation et ceci est particulièrement vrai pour le geste de pointage. L'analyse rapide des graphes de la Figure 3.3 permet de voir que sur toutes les distributions, deux modes ont l'air de se dessiner (de manière beaucoup plus nette que dans l'expérience précédente), en particulier pour le geste de pointage (et pour le geste de contrôle).

3.3.1.4 Variations des instants de production des gestes d'une expérience à l'autre

Globalement, les productions de gestes manuels interviennent légèrement plus tard en condition *Foc1* et plus tôt en condition *Foc2*. Cependant, la focalisation au sein de l'énoncé intervient plus tard que dans l'expérience précédente en condition *Foc1* et plus tôt en condition *Foc2* : il se pourrait alors que le décalage des gestes soit purement lié au décalage de la focalisation. Afin de répondre à cette question, une étude détaillée des décalages temporels de la focalisation est ici proposée.

Bien que les productions de geste en condition *Foc1* soient semblables relativement à la focalisation, les instants de production bruts (normalisés sur la durée de l'énoncé) subissent clairement un décalage en passant d'une expérience à l'autre. Ceci est explicable par le fait que l'article de la phrase prononcée n'est pas situé au même endroit dans les deux expériences : il est produit en milieu d'énoncé (avant l'objet) dans l'expérience précédente (par exemple : « Mumu tient *le* bébé ») et en début d'énoncé pour l'expérience courante (par exemple : « *Le* bonbon est vert »). Il faut noter que cette différence de structure syntaxique pourrait également mener à une

restructuration prosodique.

Cette différence dans les structures de phrase utilisées a bien entendu une influence sur les instants de production de la focalisation. D'une part, la partie focalisée en condition *Foc1* est produite plus tard dans l'expérience décrite que dans l'expérience précédente (où on avait l'instant de début de focalisation qui était égal à l'instant de début de production vocale *i.e.* temps normalisé 0). D'autre part, le début de production de la focalisation en condition *Foc2* intervient à un instant différent dans les deux expériences : elle commence au temps $0,54$ en moyenne sur tous les essais valides dans cette expérience alors qu'elle commençait au temps $0,59$ dans l'expérience précédente : la focalisation intervient légèrement plus tôt dans l'expérience décrite dans cette partie en condition *Foc2*.

Cette tendance à produire la focalisation plus tard en condition *Foc1* et plus tôt en condition *Foc2* se retrouve aussi sur les instants des production des évènements manuels : les gestes sont globalement produits plus tard en condition *Foc1* et plus tôt en condition *Foc2* si on compare à l'expérience précédente (*cf.* ci-dessus) ; ce décalage serait-il purement dû à celui de la focalisation ? La Table 3.8 donne dans les colonnes début foc, *stroke*, tenue, *stroke d'offset*, durée du geste, les différences temporelles entre l'expérience courante et l'expérience précédente pour les temps normalisés de la focalisation et les durées des phases des gestes manuels. Par ailleurs, cette table compare le décalage temporel entre les deux expérience pour les évènements manuels par rapport au décalage temporel entre les deux expériences pour le début de la focalisation : les colonnes $t_{\mathrm{P_{On}}}$, $t_{\mathrm{P_{vit}}}$, $t_{\mathrm{P_A}}$, $t_{\mathrm{P_R}}$, $t_{\mathrm{P_{Off}}}$ montrent cette différence (les décalages vont toujours dans le même sens pour la focalisation et les instants de production des gestes, sauf pour les différences en gras. Pour ces colonnes, si la différence est nulle, cela veut dire que les décalages subis par la focalisation et par l'exécution de l'instant manuel est exactement le même en passant d'une expérience à l'autre. Une différence positive signifie un décalage plus important du geste en passant d'une expérience à l'autre, une différence négative un décalage plus important de la focalisation (le sens du décalage est fourni par le signe de ce qui est indiqué dans la colonne "début foc).

GESTE	FOC.	début foc	$t_{\mathrm{P_{On}}}$	$t_{\mathrm{P_{vit}}}$	$t_{\mathrm{P_A}}$	$t_{\mathrm{P_R}}$	$t_{\mathrm{P_{Off}}}$	*stroke*	tenue	*stroke d'offset*	durée du geste
Pointage	*Foc1*	$+0,119$	$-0,072$	$-0,022$	$0,004$	$0,014$	$0,123$	$+0,077$	$+0,009$	$+0,109$	$+0,195$
	Foc2	$-0,059$	$0,086$	$0,099$	$0,099$	$0,068$	$-0,029$	$-0,013$	$+0,031$	$+0,097$	$+0,114$
Battement	*Foc1*	$+0,120$	$-0,062$	$-0,018$	$-0,020$	$0,005$	$-0,018$	$+0,042$	$+0,026$	$-0,024$	$+0,044$
	Foc2	$-0,055$	$0,048$	$0,052$	$0,040$	$0,006$	$0,052$	$+0,007$	$+0,034$	$-0,045$	$-0,004$
Contrôle	*Foc1*	$+0,116$	$-0,033$	$0,001$	$-0,009$	$\mathbf{-0,108}$	$-0,115$	$+0,024$	$-0,114$	$+0,008$	$-0,083$
	Foc2	$-0,045$	$0,062$	$0,058$	$0,057$	$0,132$	$0,152$	$+0,005$	$-0,075$	$-0,020$	$-0,090$

Les différences calculées pour les colonnes début foc, *stroke*, tenue, *stroke d'offset*, durée du geste sont : $\mathrm{t_{expe\ en\ cours}} - \mathrm{t_{expe\ precedente}}$

Pour toutes les autres colonnes, les différences calculées sont

$$|\mathrm{t_{x_{expe\ en\ cours}}} - \mathrm{t_{x_{expe\ precedente}}}| - |\mathrm{t_{debut\ foc_{expe\ en\ cours}}} - \mathrm{t_{debut\ foc_{expe\ precedente}}}|$$

TABLE 3.8 – Différences moyennes des instants normalisés de production entre les deux expériences

Cette table confirme bien que le décalage de la production de focalisation entre les deux expérience va dans le même sens que le décalage des productions des instants manuels annotés (une seule valeur en gras).

Ce tableau des décalage permet d'illustrer les résultats des ANOVAs (non présentées ici) menées sur l'influence des conditions sur les décalages entre les deux expériences. Une ANOVA étudiant le décalage acoustique de la focalisation donne un effet principal du facteur FOCALISATION ($F(1,9) = 396,88$, $p < 0,001$) mais aucun effet du type de geste produit ($F(3,27) = 0,64$, $p = 0,60$). Par ailleurs, des ANOVAs confirment globalement des comportements similaires pour les *décalages d'instants de production du geste* au sein de l'énoncé : la focalisation a toujours un effet sur le décalage des instants de production (le décalage "ne va pas dans le même sens" dans les deux conditions de focalisation) : respectivement pour l'initiation, le pic de vitesse, l'apex, le retour et l'offset : $F(1,9) = 51,38$, $p < 0,001$, $F(1,9) = 87,34$, $p < 0,001$, $F(1,9) = 63,61$, $p < 0,001$, $F(1,9) = 40,33$, $p < 0,001$, $F(1,9) = 21,83$, $p < 0,01$. A l'inverse, le type de geste n'a pas d'effet sur l'amplitude du décalage subi par les instants annotés sur l'exécution des gestes d'une expérience à l'autre (toutes les valeurs de F sont non significatives, sauf pour le retour où $F(2,18) = 3,63$, $p < 0,05$). Enfin, les ANOVAs donnent une interaction des facteurs significative seulement pour le décalage de l'instant d'apex ($F(2,18) = 4,34$, $p < 0,05$) : l'effet de la focalisation sur le décalage temporel subi par l'apex entre les deux expériences est différent selon les

gestes... Plus précisément, des tests t montrent que l'effet de la focalisation est plus fort sur le geste de pointage que sur les deux autres gestes. Ceci s'explique par le fait qu'en moyenne l'apex du geste de pointage, bien que suivant la tendance générale de tous les gestes à être produit plus tôt en condition *Foc2* dans cette expérience par rapport à la précédente, est produit significativement plus tôt (par rapport à l'expérience précédente) que les apex deux autres types de gestes. Cet effet d'interaction est représenté dans la Figure 3.4. Dans cette figure, les boîtes à moustache représentent le décalage temporel de l'apex des gestes et de la focalisation dans les différents niveaux des facteurs FOCALISATION et GESTE. Le décalage sur la figure est calculé comme étant la différence entre le temps de réalisation durant cette expérience et le temps de réalisation dans l'expérience précédente. Ainsi les boîtes sous la ligne en pointillés du zéro représentent des instants qui se sont produits plus tôt dans cette expérience. L'effet d'interaction est visible sur le volet relatif à *Foc2*, où on voit aisément que l'apex du geste de pointage intervient en moyenne beaucoup plus tôt que l'apex des autres gestes *i.e.* le décalage négatif est plus fort pour le geste de pointage.

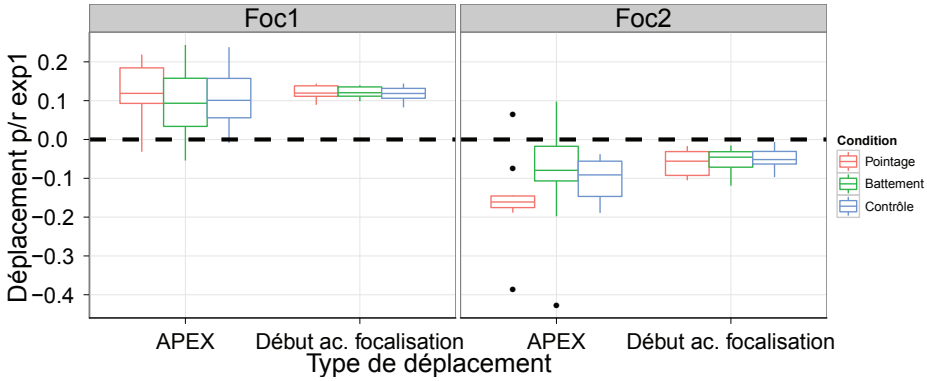


FIGURE 3.4 – Déplacement de l'apex et du début de la focalisation d'une expérience à l'autre dans les différentes conditions

Un autre point important à noter dans ce graphe (et confirmé par des tests t) est le fait que les décalages observés pour les réalisations des apex dans les conditions de battement et de contrôle ont des grandeurs similaires au décalage observé pour le début de la focalisation (*i.e.* les décalages dans l'exécution des gestes peuvent être expliqués par les décalages observés dans l'exécution de l'énoncé *i.e.* par les modifications segmentales et prosodiques du modèle de phrase) ; pour le battement, en condition *Foc1* : $t(9) = 0,86$, $p = 0,410$, *Foc2* : $t(9) = 0,84$, $p = 0,420$; pour le contrôle, en *Foc1* : $t(9) = 0,49$, $p = 0,630$, en *Foc2* : $t(9) = 2,18$, $p = 0,057$. Les résultats sont différents en condition *Foc2* pour le geste de pointage pour lequel il existe une différence significative entre le décalage de l'apex et le décalage de la focalisation entre les deux expériences, en condition *Foc1*, $t(9) = -0,20$, $p = 0,850$ mais en condition *Foc2*, $t(9) = 2,69$, $p < 0,05$. Ce résultat confirme encore une fois le comportement différent du geste de pointage qui ne subit pas les mêmes modifications que les deux autres types de gestes. Globalement, pour les gestes de contrôle et de battement, les décalages temporels observés entre les deux expériences sont similaires à ceux observés pour l'exécution de la focalisation : il y a toujours une attraction de ces deux types de gestes par la focalisation, comme dans l'expérience précédente. Ce n'est cependant pas le cas pour le geste de pointage pour lequel les valeurs des décalages temporels sont différentes de celles observées pour la focalisation en condition *Foc2*. Une étude par groupes est proposée ci-dessous.

3.3.2 Définition de deux groupes de participants et études par groupes

Dans le cas du geste de pointage, l'étude détaillée des productions permet de voir que les productions des participants sont en fait classifiables en deux groupes distincts.

Tous les participants produisent approximativement leurs gestes de pointage dans la condition *Foc1* à des instants similaires à ce qui était réalisé dans l'expérience précédente. Des tests t comparant les instants de production des apex par rapport au début de l'instant de production en condition *Foc2* montrent que les participants

se divisent en deux groupes : ceux ayant des productions similaires à l'expérience précédente et ceux pour lesquels l'attraction du geste par la focalisation est plus limitée. Ainsi les participants 8, 9, 10 ont un comportement similaire à l'expérience précédente (le facteur expérience n'a pas d'influence sur les instants de production d'apex du geste) et les participants 1, 2, 3, 4, 5, 6, 7 ont des productions différentes de ce qui était trouvé dans l'expérience précédente (effet significatif du facteur expérience). Dans la suite, le groupe **se comportant comme dans la première expérience** sera nommé **Groupe 1** et le groupe ayant un **comportement différent Groupe 2** . Les densités de probabilité estimées des instants de production d'apex pour l'ensemble des sujets (figure à gauche) et pour les deux groupes (figure à droite) dans les deux expériences sont représentées en Figure 3.5. Cette figure illustre globalement le fait que les productions du Groupe 2 sont assez différentes entre les deux expériences en condition *Foc2*, mais pas en *Foc1*.

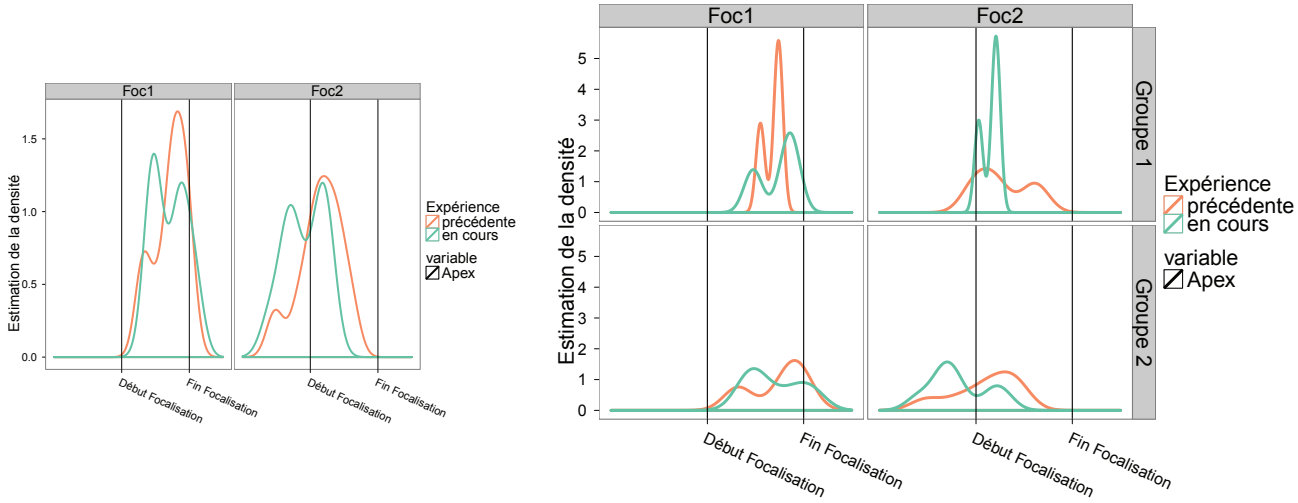


FIGURE 3.5 – Densités de probabilité des instants de production de l'apex du geste de pointage par rapport à la focalisation pour tous les participants (à gauche) et pour les participants répartis en groupe (à droite)

Ce découpage en groupe remet en question les analyses faites précédemment au moins pour ce qui est de l'effet de la focalisation : il est probable que les effets trouvés précédemment soient effectifs pour un groupe mais pas pour l'autre…

3.3.2.1 Étude par groupe : influence des facteurs de production

Afin de vérifier que la division en deux groupes des participants est bien justifiée, il est possible de refaire les analyses menées précédemment sur tous les participants pour chaque groupe de participants. Cependant, de telles analyses sur des groupes ne comportant que peu de participants n'ont pas de sens. Ainsi, seules quelques analyses qualitatives sont ici proposées afin de rendre compte du comportement des deux groupes constitués. La Figure 3.6 donne un aperçu des instants moyens de production des instants manuels pour chaque groupe dans les deux expériences. Les instants (moyens) de début et de fin de la focalisation sont représentés par des disques rouges, les instants manuels (moyens, ainsi que l'erreur-type associée) sont représentés par des barres verticales.

On voit sur ces graphes l'adaptation des deux groupes au changement d'expérience :

– Pointage – pour les deux groupes, le comportement observé dans la condition *Foc1* est similaire à celui observé dans l'expérience précédente : l'apex n'intervient que de manière marginale hors de la focalisation et intervient en général vers la fin de la focalisation, le retour se situe généralement après la focalisation. En condition *Foc2*, le comportement est différent selon les groupes : pour le Groupe 1, l'apex intervient (et intervenait) soit dans la partie focalisée soit légèrement avant celle-ci, le retour intervient (et intervenait) globalement dans la partie focalisée : les productions ne changent pas relativement à la focalisation par rapport à l'expérience précédente ; pour le Groupe 2 les choses sont légèrement différentes puisque l'a-

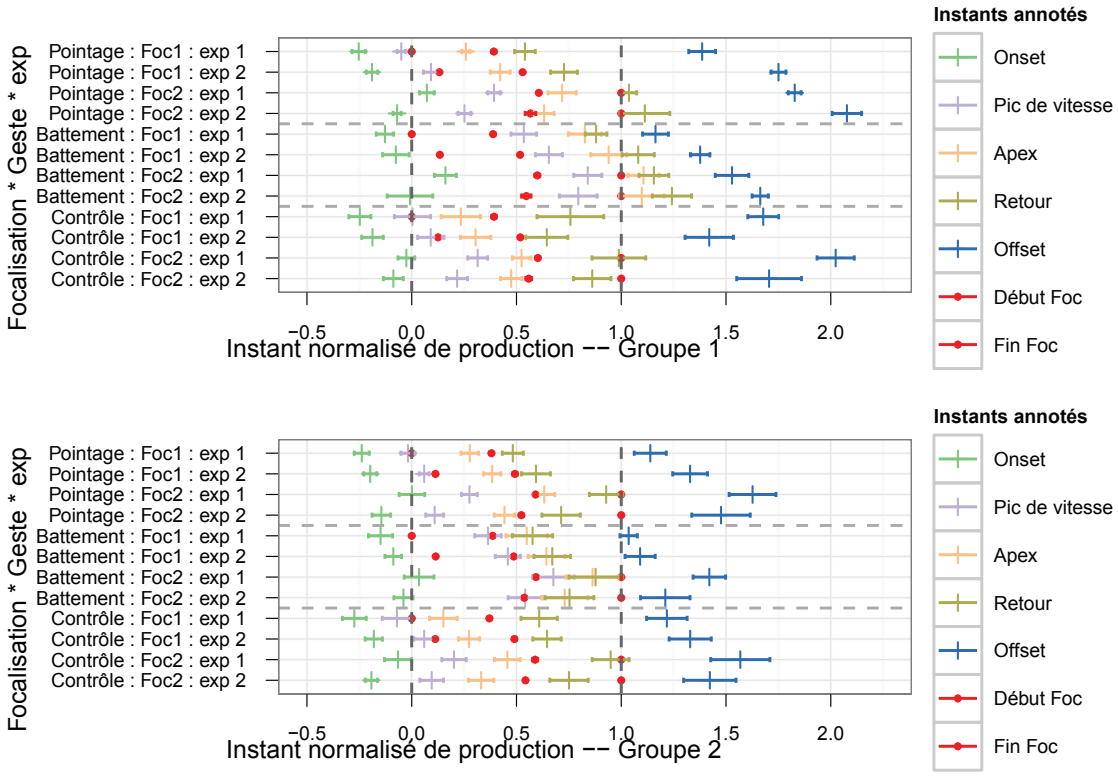


FIGURE 3.6 – Instants d'apex et retour moyens dans chaque groupe, chaque expérience et chaque condition GESTE et FOCALISATION

pex qui intervenait relativement tard dans la focalisation (et toujours au sein de celle-ci) intervient dans l'expérience présente le plus souvent avant la focalisation (et le retour toujours assez tôt au sein de la focalisation).

- Battement – pour les deux groupes, les productions de l'apex/retour ne semblent pas réellement changer d'une expérience à l'autre : on observe une production plus rapide du battement dans la condition *Foc2* pour le Groupe 2 (le geste arrive plus tôt dans l'énoncé) mais à part cela, aucune modification importante de l'instant de production du battement n'est notable. Comme dans l'expérience précédente, l'apex du geste de battement intervient dans la focalisation surtout en condition *Foc2*, ceci est (beaucoup) plus rare en condition *Foc1*. On observe, comme pour le geste de pointage, que la condition FOCALISATION n'a pas d'influence sur l'instant de production du battement pour le groupe Groupe 2.
- Contrôle – les modifications des productions du geste de contrôle ressemblent fortement à ce qu'on observe en condition *Pointage* : le geste, attiré par la focalisation dans l'expérience précédente est moins attirée dans cette expérience et plus spécifiquement, pour le groupe Groupe 2, les productions de gestes sont faites au même moment en condition *Foc1* et *Foc2*.

3.3.2.2 Variations des instants des productions *par groupe*

Les résultats présentés en Section 3.3.1.4 représentent une tendance sur l'ensemble des participants (Groupe 1 et Groupe 2 sont regroupés). Il est attendu, par construction, que les modifications soient en fait visibles au sein du Groupe 2 et que le Groupe 1 ne modifie pas l'instant de production d'apex par rapport à la focalisation.

L'étude de la Figure 3.6 et des valeurs numériques associées aux différences de décalage entre gestes manuels et focalisation présentées en Table 3.9 montre que, bien que la production de la focalisation ne soit pas vraiment impactée par le groupe étudié, la production des gestes, quant à elle, est considérablement impactée et ce, que ce

Condition	Foc	Groupe	Début foc	$t_{P_{On}}$	$t_{P_{vit}}$	t_{P_A}	t_{P_R}	$t_{P_{Off}}$	*stroke*	tenue	*stroke d'offset*	durée du geste
Pointage	Foc1	Groupe 1	0,132	−0,069	0,010	0,031	0,054	0,232	0,100	0,023	0,177	0,301
Pointage	Foc1	Groupe 2	0,113	−0,073	−0,036	−0,007	−0,003	0,076	0,066	0,003	0,079	0,149
Pointage	Foc2	Groupe 1	−0,039	−0,102	−0,101	−0,046	0,115	0,289	0,056	0,161	0,173	0,391
Pointage	Foc2	Groupe 2	−0,067	−0,078	−0,098	−0,121	−0,147	−0,083	−0,043	−0,025	0,064	−0,004
Battement	Foc1	Groupe 1	0,134	−0,082	−0,015	−0,020	0,067	0,079	0,061	0,087	0,012	0,161
Battement	Foc1	Groupe 2	0,114	−0,054	−0,020	−0,020	−0,020	−0,059	0,033	0,000	−0,039	−0,006
Battement	Foc2	Groupe 1	−0,052	−0,117	0,007	0,044	0,139	0,187	0,160	0,094	0,048	0,303
Battement	Foc2	Groupe 2	−0,055	−0,019	−0,078	−0,077	−0,068	−0,154	−0,058	0,008	−0,085	−0,135
Contrôle	Foc1	Groupe 1	0,125	−0,066	−0,038	−0,056	−0,238	−0,383	0,009	−0,182	−0,145	−0,317
Contrôle	Foc1	Groupe 2	0,111	−0,019	0,018	0,011	−0,074	−0,000	0,029	−0,085	0,073	0,018
Contrôle	Foc2	Groupe 1	−0,043	−0,018	−0,054	−0,006	−0,083	−0,274	0,012	−0,078	−0,190	−0,255
Contrôle	Foc2	Groupe 2	−0,045	−0,081	−0,061	−0,080	−0,153	−0,099	0,001	−0,073	0,053	−0,018

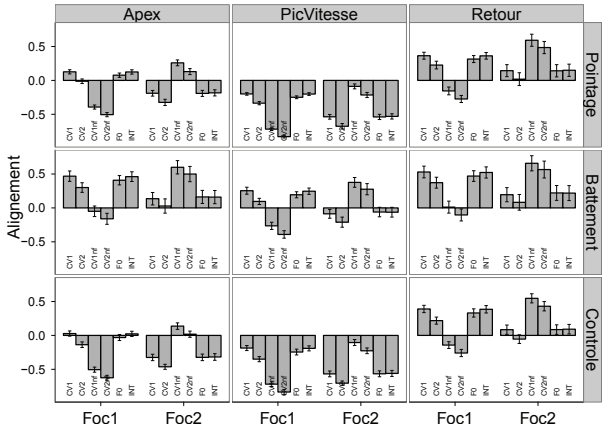
TABLE 3.9 – Différence des instants manuels par rapport à la différence des instants de production de la focalisation entre les deux expériences, pour chaque groupe et différences des durées des phases du geste entre les deux expériences

soit pour le geste de pointage (ce qui est normal, par construction) ou pour les autres types de gestes.

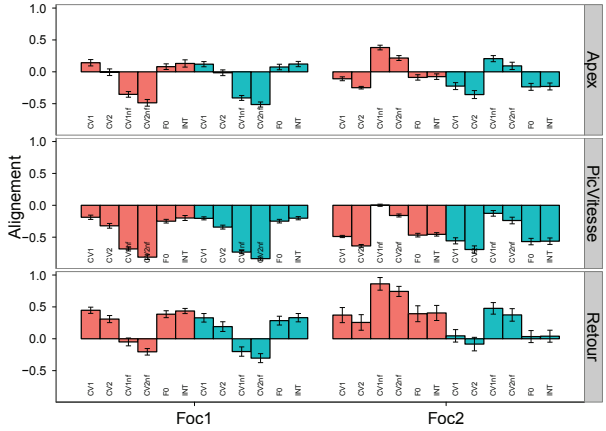
- Pour le geste de pointage, il apparaît que si les gestes démarrent approximativement au même moment dans les deux groupes et dans les deux conditions de focalisation, les instants de production des évènements suivants diffèrent. Dans la condition *Foc1*, dès le pic de vitesse, les décalages dans les productions avec l'expérience précédente sont visibles : le pic de vitesse ainsi que tous les instants suivants sont produits légèrement plus tard que dans l'expérience précédente pour le Groupe 1 (environ +50 ms, sauf pour l'instant d'*offset*). Dans la condition *Foc2*, les divergences dans l'adaptation à la tâche sont plus marquées : l'instant d'apex subit de grosses modifications par rapport à l'expérience précédente pour le Groupe 2 qui le produit beaucoup plus tôt, il en va de même pour tous les instants suivants, les décalages du Groupe 1 sont similaires au décalage subi par le du début de l'élément focalisé pour les évènements jusqu'à l'apex.
  Manifestement, les différences de durées des gestes entre les deux groupes se situent dans le *stroke* et le *stroke* de retour en condition *Foc1* et dans le *stroke*, la tenue et le *stroke* de retour en condition *Foc2*. Les différences entre les deux expériences se trouvent au niveau des durées du *stroke* et du *stroke* de retour.
- Pour le geste de battement, les deux groupes se comportent de façon similaire dans le cas *Foc1* et le comportement est très proche de ce qui est observé dans l'expérience précédente. En condition *Foc2*, les deux groupes initialisent leurs gestes légèrement plus tôt que dans l'expérience précédente, mais ce décalage est plus important pour le Groupe 1. Pour tout le reste du geste, le Groupe 1 effectue ses gestes moins vite que dans l'expérience précédente alors que le Groupe 2 est globalement plus rapide que dans l'expérience précédente. Le Groupe 1 est par ailleurs plus rapide dans ses exécutions du geste de battement que le Groupe 2.
  Les différences de durées se situent surtout au niveau de la tenue en condition *Foc1* et du *stroke*, de la tenue en condition *Foc2*. Les ajustements de durées entre les expériences se font ici encore surtout sur le *stroke* et le *stroke* de retour qui sont les deux seules phases qui ont un sens pour le geste de battement.
- Pour le geste de contrôle, en condition *Foc1*, les deux groupes suivent un pattern d'adaptation des productions à l'expérience semblable jusqu'à la position d'apex : les gestes sont produits plus tard (avec un petit retard supplémentaire pour le Groupe 2) tout comme la focalisation jusqu'à l'instant d'apex, puis la tenue est moins longue que dans l'expérience précédente dans les deux groupes et le stroke de retour est moins long que dans l'expérience précédente pour le Groupe 1. En condition *Foc2*, le deuxième groupe produit les gestes plus tôt que dans l'expérience précédente, globalement les écarts de production restent constants, sauf pour le *stroke* de retour qui est beaucoup plus rapide que dans l'expérience précédente pour le Groupe 1.
  Ici les différences (inter-groupes et inter-expériences) dans les durées se trouvent surtout dans la tenue et la durée du *stroke* de retour.

3.3.3 Alignements possibles entre instants de production de la parole et instants du geste

Afin de voir si les points de coordination sont stables par rapport à l'expérience précédente, il est intéressant de voir s'il existe des alignements possibles entre l'exécution des gestes et l'exécution de la parole. Tout comme précédemment, on cherche les alignements possibles en calculant les différences temporelles entre les instants manuels ($t_{\mathrm{P}_{\mathrm{vit}}}$, $t_{\mathrm{P}_{\mathrm{A}}}$, $t_{\mathrm{P}_{\mathrm{R}}}$) et les indices de la focalisation sur le signal de parole (t_{F_0}, t_{Int}et t_{CV_1}, t_{CV_2}). Il a été montré pour tous les gestes (et de façon marquée seulement pour le pointage) que les productions en condition *Foc2* étaient exécutées plus tôt par rapport à l'expérience précédente, et les résultats de l'expérience précédente ont montré une coordination entre les gestes et les cibles articulatoires, ainsi les instants de production des cibles articulatoires de la partie non focalisée de la parole ont ici également été annotées : $t_{\mathrm{CV}_1 nf}$, $t_{\mathrm{CV}_2 nf}$. Le traitement des données est le même qu'en Section 2.3.3.3 : les différences temporelles représentées sont des moyennes par participants agrégées sur tous les participants. Les analyses sont ici présentées de manière générale et par groupe lorsque cela est pertinent. Les différences temporelles sont représentées en Figure 3.7.



(a) Différences temporelles moyennes entre geste manuel et parole



(b) Délais moyens entre pointage et parole par groupe (en rouge, le Groupe 1, en bleu le Groupe 2)

FIGURE 3.7 – Délais moyens entre les indices de la focalisation et les instants annotés des gestes

Tout comme précédemment, cette figure permet de voir que le geste de battement est le geste qui donne les différences temporelles avec la parole les plus variables (*cf.* les barres d'erreur associées aux délais pour le battement), le geste de pointage ainsi que le geste de contrôle ont des différences temporelles plus régulières avec la parole.

Geste de pointage Les instants de ce geste qui semblent avoir des différences temporelles uniformes (*i.e.* écart-type faible) avec les instants annotés de la parole sont clairement l'apex et le pic de vitesse. Cependant, le retour a également un rôle intéressant qui sera détaillé par la suite. L'apex est ici encore l'instant du geste qui a les valeurs de différences temporelles les plus faibles avec les instants annotés de la parole.

Les analyses statistiques montrent qu'en condition *Foc1*, l'apex du geste de pointage n'intervient pas à un instant significativement différent de celui de CV_2 ($t(9) = -0,65$, $p = 0,530$), ces deux instants sont séparés en moyenne de 21 ms pour les deux groupes de participants. En condition *Foc2*, l'analyse de groupe est effectuée puisqu'il a été montré que les participants se divisent en deux groupes.

Les alignements observés sont légèrement différents selon le groupe pris en compte en condition *Foc2*. En effet, lorsque la focalisation porte sur l'attribut, le Groupe 2 continue à aligner les apex de ses gestes de pointage avec CV_2 du sujet de la phrase, mais ce n'est pas le cas pour le Groupe 1 pour lequel l'instant qui semble le mieux aligné avec l'apex est la première cible articulatoire de la focalisation. L'instant de retour est aligné avec les cibles articulatoires de la partie focalisée en condition *Foc2* pour le Groupe 2 mais on n'observe aucun alignement dans l'autre groupe.

Geste de battement Comme évoqué ci-dessus, c'est le geste qui a les différences temporelles les moins régulières avec la parole. L'instant le moins variable est le pic de vitesse du geste. La variabilité des instants de production des gestes (et donc des différences temporelles) mène à des alignements nombreux mais peu fins. En condition *Foc1*, l'apex du geste est aligné avec la cible articulatoire de l'élément non focalisé ($t(9) = -0,64$, $p = 0,540$), il en va de même pour le retour dont l'instant de production est similaire à l'instant de production de l'apex ($t(9) = 0,13$, $p = 0,900$), on n'observe cependant aucun alignement entre les instants de la parole et le pic de vitesse (le pic de vitesse est proche et intervient à un instant non significativement différent de la fin de la focalisation : $t(9) = 0,42$, $p = 0,680$, mais c'est le seul alignement observé). En condition *Foc2*, l'apex est encore aligné avec CV_1 ($t(9) = 1,48$, $p = 0,170$) et on observe également des alignements avec les pics de fréquence fondamentale ($t(9) = 1,68$, $p = 0,130$) et d'intensité ($t(9) = 1,64$, $p = 0,140$). Les mêmes alignements sont bien entendu observés pour le retour (respectivement : $t(9) = 1,87$, $p = 0,090$; $t(9) = 2,02$, $p = 0,070$; $t(9) = 1,98$, $p = 0,080$). Ces alignements sont également observés pour le pic de vitesse (respectivement : $t(9) = -1,33$, $p = 0,210$; $t(9) = -0,67$, $p = 0,410$; $t(9) = -0,88$, $p = 0,400$).

Les alignements se font globalement entre l'apex du geste et la première cible articulatoire de l'attribut de la phrase. Cependant, ces alignements souffrent d'un écart-type assez grand *i.e.* ils ne sont pas réguliers.

Geste de contrôle Pour ce geste encore, l'apex est l'instant qui offre les différences temporelles les plus régulières avec les instants annotés de la parole. En particulier, en condition *Foc1*, on trouve que l'apex est aligné avec la première cible articulatoire de l'élément focalisé ($t(9) = 0,72$, $p = 0,490$) et les pics de fréquence fondamentale ($t(9) = -0,71$, $p = 0,500$) et d'intensité ($t(9) = 0,60$, $p = 0,560$), aucun alignement n'est trouvé ni pour l'instant du pic de vitesse ni pour l'instant de retour. En condition *Foc2*, on observe un alignement entre l'apex et la seconde cible articulatoire de l'élément non focalisé de l'énoncé ($t(9) = 0,34$, $p = 0,740$) et des alignements entre le retour et la cible articulatoire de l'attribut ($t(9) = 1,24$, $p = 0,250$) et les pics de F_0 et d'intensité (resp. $t(9) = 1,23$, $p = 0,250$ et $t(9) = 1,30$, $p = 0,230$).

3.3.4 Effets de la production des gestes manuels sur les productions en parole

Il est intéressant de voir quelle influence ont la production de gestes et le déplacement de l'emplacement focalisé sur les durées et les amplitudes de la parole. Les résultats des ANOVAs sur ces grandeurs sont présentés en Table 3.10.

	FOCALISATION	GESTE	FOCALISATION×GESTE
Durée de l'énoncé (non normalisée)	$F(1,9) = 1,78$ – $p = 0,21$	$F(3,27) = 0,47$ – $p = 0,70$	$F(3,27) = 0,33$ – $p = 0,80$
Durée de la partie focalisée	$\mathbf{F(1,9) = 33,35}$ – $\mathbf{p < 0,001}$	$F(3,27) = 1,48$ – $p = 0,24$	$F(3,27) = 1,20$ – $p = 0,33$
Instant d'onset vocal (non norm.)	$F(1,9) = 0,80$ – $p = 0,39$	$F(3,27) = 2,19$ – $p = 0,11$	$F(3,27) = 0,36$ – $p = 0,78$
Amplitude pic de F_0	$F(1,9) = 0,90$ – $p = 0,34$	$F(3,27) = 0,066$ – $p = 0,98$	$F(3,27) = 0,0042$ – $p = 0,99$
Amplitude pic d'intensité	$\mathbf{F(1,9) = 4,26}$ – $\mathbf{p < 0,05}$	$F(3,27) = 0,0662$ – $p = 0,98$	$F(3,27) = 0,0042$ – $p = 0,99$
Amplitude CV_1	$\mathbf{F(1,9) = 13,49}$ – $\mathbf{p < 0,001}$	$F(3,27) = 1,79$ – $p = 0,16$	$F(3,27) = 1,04$ – $p = 0,38$

TABLE 3.10 – ANOVAs sur les durées/instants acoustiques de la parole

Durées des productions vocales La Table 3.10 permet de voir l'influence des facteurs sur les durées de production de l'énoncé et de la focalisation.

La durée brute de la production vocale n'est pas influencée ni par l'emplacement de la focalisation ni par la production de gestes (durées légèrement plus longues en condition *Foc1* mais ce n'est pas significatif).

Par ailleurs, seul l'emplacement de la focalisation a une influence sur la durée de l'élément focalisé , comme illustré en Figure 3.8 : l'élément focalisé en condition *Foc2* est plus long qu'en condition *Foc1*. Cet effet est probablement dû à l'allongement final qu'on observe en fin d'énoncé. De manière plus importante : la durée de l'élément focalisé n'est pas influencée par la production de geste et les différents types de gestes n'ont aucun effet sur la durée de l'élément focalisé.

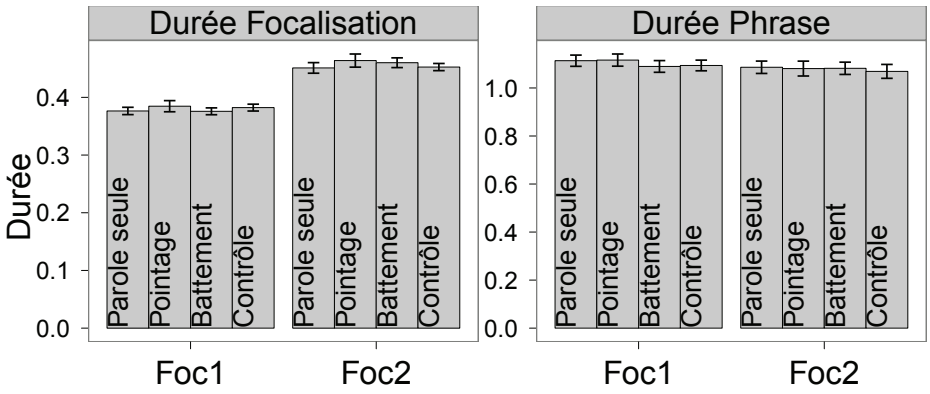


FIGURE 3.8 – Influence des facteurs sur les durées de la parole (durée de la focalisation, durée de l'énoncé)

"Temps de réponse" Le temps de réponse est, comme dans l'expérience présenté précédemment, une variable d'intérêt. L'ANOVA présentée dans la Table 3.10 ne montre aucune influence des facteurs FOCALISATION et GESTE sur la durée séparant la présentation des stimuli visuels du début de la parole. Dans cette tâche, contrairement à l'expérience présentée précédemment, l'apparition des images sur l'écran est nécessaire pour pouvoir réaliser la tâche. Bien que dans l'expérience précédente, aucune réponse n'ait été formulée avant l'apparition des images, la réponse en parole dans la condition gestuelle nécessitant l'apparition des dessins (*i.e.* condition de pointage) était plus longue.

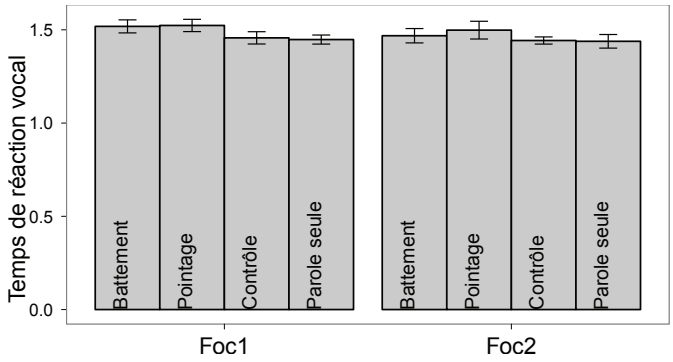


FIGURE 3.9 – Influence de la production de gestes sur l'"onset vocal"

La Figure 3.9 illustre les résultats de l'ANOVA et permet de voir que les instants de production suivant l'apparition des deux images est très peu variable (écarts-types faibles) au sein des productions. Dans toutes les conditions, le temps mis pour répondre vocalement est d'environ $1,45$ s. Ce temps est assez long et correspond sûrement au fait que la tâche ne soit pas automatique : une fois que les stimuli visuels apparaissent à l'écran il faut trouver quelle caractéristique (un nom ou une couleur) est en commun entre le stimulus et les images présentées, puis préparer l'énoncé à prononcer avant de "lancer" la parole.

Variation des amplitudes Les ANOVAs montrent que la production de gestes n'a pas d'influence sur l'amplitude des corrélats acoustiques (et articulatoires) de la focalisation.

Les ANOVAs (et ceci est illustré en Figure 3.10) montrent par contre un effet de la focalisation sur l'amplitude du pic d'intensité et sur la première cible articulatoire de la partie focalisée : le pic d'intensité est plus important dans la condition *Foc1* que dans la condition *Foc2* et l'amplitude de CV_1 est plus grande en condition *Foc2* qu'en *Foc1*. L'effet de la focalisation sur le pic d'intensité peut s'expliquer par le phénomène de déclinaison au sein du syntagme intonatif. L'effet observable sur la cible articulatoire va dans le sens contraire à ce phénomène mais peut être expliqué de deux façons. Premièrement, il est possible que les participants aient hyper-articulé afin de compenser la déclinaison. Deuxièmement, les cibles articulatoires utilisées dans les attributs (phonèmes /o,ɛ,u/)

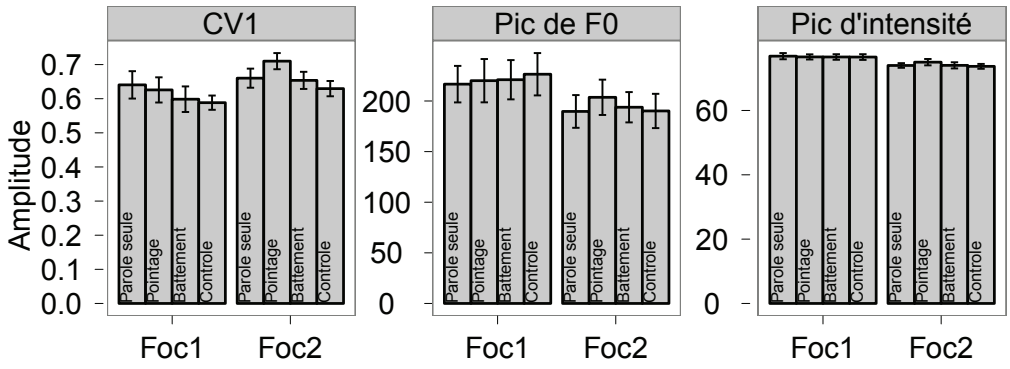


FIGURE 3.10 – Amplitudes des corrélats acoustiques et articulatoires de la focalisation

sont des cibles mieux définies par leur ouverture/protrusion que les cibles articulatoires utilisées en sujet de phrase (phonèmes /ɔ̃, a, e/).

3.4 Discussion

L'étude présentée dans cette partie a pour objectif d'affiner l'étude de la coproduction des gestes manuels et de la parole dans le cadre de la désignation multimodale exposée précédemment. Les conditions expérimentales, le dispositif expérimental, les participants sont tous identiques à l'expérience précédente, seule la tâche est différente entre les deux expériences. La tâche a été modifiée afin d'étudier une question plus fine que la question de la simple coordination avec la focalisation et ainsi, cette expérience est intéressante à commenter à la lumière des résultats de l'étude précédente. Plus précisément, le lien communicatif a été manipulé entre les deux expériences. Il est à noter qu'un effet de bord de cette modification est la modification des structures syntaxiques (l'article de la phrase n'est pas au même emplacement dans la phrase) qui a pu introduire une restructuration prosodique…

De manière globale, il a été montré que les résultats de l'expérience précédente ne se retrouvaient pas directement ici. Toutes choses étant similaires par ailleurs, seul le changement de tâche a pu mener à cette modification.

Pour ce qui est de la coordination geste/parole, les résultats montrent que les productions des participants peuvent se scinder en deux groupes dont les stratégies de coordination diffèrent. Ce découpage en groupes (qui n'était pas applicable dans l'expérience précédente, tous les participants suivant un même "motif" de production des gestes en rapport avec la parole) montre un effet de la tâche différent selon les participants. Globalement les gestes semblent toujours être "attirés" par la focalisation, comme dans l'expérience précédente, mais avec une attraction moins forte. L'étude par groupe montre une coordination proche de l'expérience précédente pour les gestes de contrôle et de battement et pour dans la condition *Foc1* pour le geste de pointage. Globalement, deux stratégies distinctes sont utilisées dans les conditions pointage et *Foc2* : un groupe (constitué de trois personnes sur dix) continue de suivre le motif de coordination qu'il utilisait dans l'expérience précédente, ce qui n'est pas le cas pour les sept autres participants qui utilisent une stratégie différente.

Pour ce qui est des productions respectives dans les deux modalités et leur influence mutuelle, les résultats de l'expérience précédente se retrouvent en partie. Au niveau temporel, il a été montré que ① la focalisation attirait toujours le geste globalement et ce, quel que soit le type de geste mais que cet effet était plus limité que précédemment (cet effet est en fait dû à une minorité des participants qui utilisent la même stratégie de coordination que dans l'expérience précédente) ② les gestes sont plus courts en condition *Foc1* et que les parties des gestes raccourcies sont principalement le *stroke* (et la tenue, dans une moindre mesure) ③ le *timing* interne de la parole n'est pas modifié par la production de gestes ④ que les durées de production de l'énoncé, de la focalisation ne sont pas non plus modifiées par la production de gestes. Au niveau de l'amplitude des corrélats acoustiques et articulatoires de la focalisation, il a été montré, comme dans l'expérience précédente, une absence d'effet de la production de geste. Quelques résultats sont cependant en porte-à-faux avec les résultats précédents, en particulier une quantité supplémentaire non négligeable des apex des gestes est produite hors des parties

focalisées (surtout dans la condition *Foc2*) ; aucune influence du type de geste n'est observée sur le temps séparant la présentation des stimuli visuels et le début de prononciation de la phrase de correction.

Il est rappelé ici que l'intérêt initial de cette étude était de la comparer à la précédente puisque le protocole expérimental, les participants, les conditions sont les mêmes mais la tâche diffère (en particulier, par la nature du *lien communicatif*). La discussion s'orientera donc principalement sur la comparaison des deux expériences, l'importance du lien communicatif sera discutée en premier lieu avant d'affiner les résultats concernant la coordination en exposant les deux groupes de participants mis au jour puis de discuter les résultats concernant les temps de réaction vocaux.

3.4.1 Coordination geste/parole : influence du lien communicatif

3.4.1.1 Stratégies et influence de la tâche

Comme annoncé en Section 3.2.1, le lien communicatif a été modifié en passant d'une expérience à l'autre dans la condition *Pointage*. En effet, dans l'expérience présentée plus haut, ce qui était désigné par la parole (la partie focalisée de l'énoncé) et ce qui était désigné par le geste de pointage était un référent similaire (par exemple, un bébé) ; dans cette expérience, les énoncés réalisés par les participants sont composées d'un objet et de la couleur de cet objet, la focalisation porte soit sur la dénomination de l'objet soit sur sa couleur, et au contraire, ce qui est designé par le pointage est l'ensemble {nom ; couleur} de l'objet dont il est question dans la phrase. Ainsi, la correspondance directe entre parole et cible du geste de pointage établie dans l'expérience précédente n'est plus de mise dans cette expérience.

Les prédictions faites sur l'influence de cette modification du lien communicatif présentées en Section 3.2.1 prévoyaient une adaptation différente de la coordination geste manuel/parole selon le type de geste considéré. Ainsi, il était attendu que les gestes de contrôle et de battement ne subissent pas de modification de leur coordination avec la parole puisqu'il n'y a pas de changement de la relation geste/parole par rapport à l'expérience précédente. Concernant le pointage, il était attendu que deux stratégies pourraient être mises en place pour s'adapter à la modification du lien communicatif entre geste de pointage et parole. Dans l'expérience précédente, la focalisation attirait le pointage (et plus précisément son apex, son plateau), ici les prédictions étaient : soit la focalisation attire le pointage (le regroupement des fonctions de désignation guide la coordination geste/parole), soit la focalisation n'a pas d'influence sur les instants de production du pointage (la coordination est guidée par l'élément désigné –qui est le même stimulus visuel quelle que soit la condition de focalisation).

Les résultats montrent globalement des productions similaires à l'expérience précédente en condition *Foc1* mais des comportements différents selon le type de geste en condition *Foc2*. Il a été montré une attraction moins forte par la focalisation en condition *Foc2* que dans l'expérience précédente : les gestes sont produits plus tôt que dans l'expérience précédente pour tous les types de gestes.

3.4.1.2 Alignements geste manuels/parole

Les résultats précédents vont dans le sens des prédictions faites en Section 3.2.1. Globalement la modification du lien communicatif entre le geste de pointage et la parole dans la tâche de désignation mise au point dans ces deux expériences influence la coordination entre la parole et le geste manuel.

- Battement : Le pic de vitesse est l'instant le moins variable, le pic de vitesse/l'apex sont souvent produits après la focalisation en condition *Foc1* et souvent dans la focalisation en condition *Foc2*. L'apex (et pic de vitesse en *Foc2*) est aligné avec la cible articulatoire de l'attribut de la phrase. L'étude par groupes montre que le battement est aligné avec la première cible articulatoire de la partie non focalisée en condition *Foc1* mais qu'en *Foc2*, le Groupe 1 a des productions alignées avec la fin de focalisation alors que le Groupe 2 aligne ses productions avec la cible articulatoire de la partie focalisée.
- Contrôle : L'apex est l'instant le moins variable, l'apex est souvent produit dans la focalisation en condition *Foc1* et avant la focalisation en condition *Foc2*. L'apex est aligné avec la première cible articulatoire de la partie focalisée en condition *Foc1* et avec la seconde cible articulatoire de la partie non focalisée de l'énoncé

en condition *Foc2*. L'étude par groupe ne montre aucune distinction dans les alignements pour les deux groupes de participants.

– Pointage : L'apex est l'instant le moins variable et est souvent produit au sein de la focalisation en condition *Foc1* et assez régulièrement avant celle-ci en condition *Foc2*. Les productions des participants sont alignées (l'apex est aligné) avec la seconde cible articulatoire de l'élément focalisé en condition *Foc1*. L'étude par groupe montre que les participants peuvent être séparés en deux groupes dont les productions diffèrent grandement en condition *Foc2* : le Groupe 1 dont les productions sont semblables à celles de l'expérience précédente (et qui alignent apex et la première cible articulatoire de la focalisation en condition *Foc2* et le Groupe 2 dont les productions manuelles sont pratiquement similaires entre la condition *Foc1* et *Foc2* (et qui alignent les apex des gestes de pointages sur la seconde cible articulatoire du sujet de la phrase également en condition *Foc2*).

Enfin, il a été montré que la production de geste manuel dans la condition *Foc1* n'est pas modifiée par rapport à l'expérience précédente et ce, pour tous les types de gestes : l'apex arrive approximativement au même moment au sein du mot focalisé dans les deux expériences, que l'article "le" soit présent ou non en tête de phrase. Ceci est vrai alors que l'instant d'initiation des gestes ne subit qu'un décalage temporel mineur comparé au décalage subit par la focalisation : le timing du geste est adapté afin d'obtenir des coordinations similaires à ce qui était produit dans l'expérience précédente.

Bien que ces résultats soient *a priori* concordants avec les hypothèses avancées, une critique évidente est que les propriétés segmentales des phrases du corpus (prononcées par les participants) ne sont pas les mêmes que celles de l'expérience précédente, et donc que, conclure directement à partir de ces résultats est trop hâtif.

Comme présenté dans la Section 3.3.1.4, la modification du matériel linguistique utilisé implique une modification des instants de productions des différentes parties de l'énoncé réalisé. Cette modification est en partie prise en compte par la normalisation des énoncés et on peut montrer, qu'une fois la normalisation effectuée, les "sujets" des phrases ont des durées normalisées similaires d'une expérience à l'autre et que seuls les "objets" ou "attributs" ont des longueurs normalisées qui diffèrent : plus long dans l'expérience courante. Un point important reste la longueur de ce qui est entre le sujet et l'objet/attribut (*i.e.* le verbe) qui est plus long dans l'expérience précédente. En plus de la modification du matériel linguistique utilisé (et impliquant une modification des durées des mots prononcés), la modification de la tâche apporte une modification de la structure grammaticale de la phrase un peu plus pernicieuse : la position de l'article est déplacée (depuis le "milieu" de la phrase –avant l'objet– dans l'expérience précédente au début de la phrase dans cette expérience). *A priori*, à cause de ces modifications structurelles/formelles de la phrase prononcée, si la production de gestes est liée à celle de la parole, il est possible que toutes les productions de gestes soient produites plus tard en condition *Foc1* et plus tôt en condition *Foc2* (car le verbe a une durée plus courte).

Ces modifications qu'on peut attendre *a priori* à cause des modifications sont réellement observées. Or il a été montré (*cf.* Figure 3.4) que ces modifications des instants de production de la parole avaient des grandeurs similaires aux modifications subies par les instants de production des gestes en condition *Foc1* pour tous les gestes et en condition *Foc2* pour les gestes de battement et de contrôle mais *pas* pour le geste de pointage pour lequel le décalage temporel de l'apex observé n'est pas explicable totalement par le décalage temporel des instants de la parole.

3.4.2 "Temps de réaction" et planification de la production

Un résultat intéressant à commenter est le "temps de réaction" (temps entre l'instant de présentation des images et le début de la parole) nécessité dans les différentes conditions GESTE dans les deux expériences. Il a été montré dans l'expérience précédente (Section 2.3.3.4) que le temps de réponse vocale était plus long dans la condition *Pointage* et ce, dans les deux conditions de focalisation, avec un démarrage plus tardif en condition *Foc1*. Cela pouvait être expliqué par le fait que, bien que l'apparition du stimulus visuel ne contraignait pas la planification de la parole à produire, celui-ci contraignait la planification du geste de pointage correspondant et, la parole attendant que le geste soit (au moins en partie) planifié, démarrait plus tard. Pour rappel, le temps séparant la présentation des images du début de production de parole est en moyenne de 650 ms en condition

Foc1 et 760 ms en condition *Foc2* pour les conditions Parole seule, Battement et Contrôle et de $1,27$ s en condition *Foc1* et $1,08$ s en condition *Foc2* pour le pointage. Le facteur FOCALISATION a un effet sur le laps de temps séparant la fin de présentation des images du début de la parole seulement en condition *Pointage*.

Dans l'expérience présentée dans ce chapitre, au contraire, les participants ont besoin de voir les images afin de planifier la parole à produire (et bien entendu, afin de planifier le geste de pointage à effectuer). Les instants de "début de parole" sont ici égaux à la fois parmi les gestes mais aussi parmi les conditions de focalisation. Le temps écoulé entre la fin de présentation des images et le début de la parole est en moyenne de $1,47$ s en condition *Foc1* et $1,45$ s en condition *Foc2* (différence non significative).

Les différences entre les deux expériences sont représentées sur la Figure 3.11.

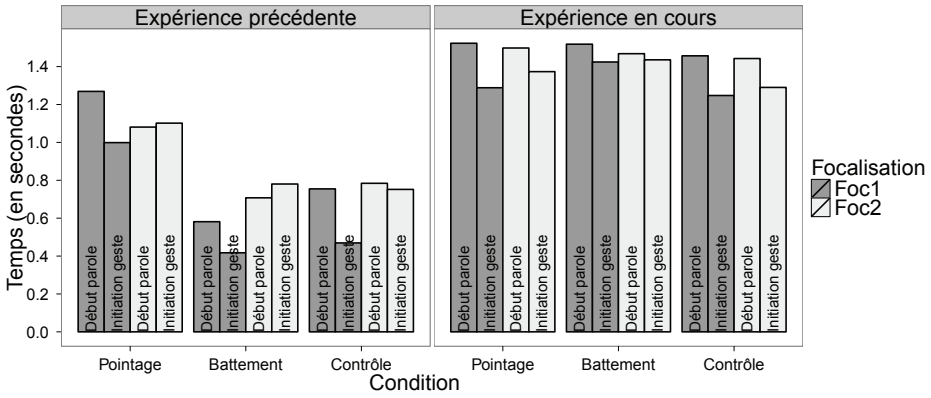


FIGURE 3.11 – Temps séparant l'apparition des stimuli visuels et le début de la parole/du geste manuel

Trois comportements sont notables en passant d'une expérience à l'autre :

- Les temps mis pour commencer l'exécution de la parole/du geste sont très largement différents
- Dans l'expérience de cette partie, pour le pointage, il n'y a plus de différence entre les deux conditions de focalisation
- Il n'y a plus de différences entre la condition *Pointage* et les autres

Le premier point est simple à expliquer, c'est ce qui a été évoqué ci-dessus, dans la seconde expérience, il est nécessaire d'avoir vu les images pour faire la correction (à la fois en parole mais aussi avec le pointage). Le second point est évident par ce qui a été montré au dessus : pour le Groupe 2 (le plus nombreux) ce qui importe dans la coordination, c'est l'élément désigné et les gestes de pointage subissent peu d'effet de la focalisation. Le dernier point est plus intéressant à commenter. L'instant de début de production pour la parole était différent entre la condition *Pointage* et les autres conditions de GESTE dans l'expérience précédente, ceci était expliqué par le fait que la "parole attend le geste pour des raisons communicatives" (*i.e.* la parole commence lorsque le geste est "assez prêt" pour avoir le temps de terminer sa planification et d'être exécuté à temps afin qu'apex et focalisation soient cooccurrents). Ici les raisons communicatives sont différentes (pour le Groupe 2 en condition pointage en tous cas) mais cette contrainte reste respectée. Le fait que les toutes les modalités exhibent un temps de réaction similaire permet de montrer que la planification du geste et de la parole se fait en parallèle : en effet, dans l'expérience précédente, la parole démarrait plus tard dans la condition geste de pointage et ceci était expliqué par le fait que cette condition était la seule dépendant de l'emplacement des stimuli visuels. Ici bien que la condition avec la production de geste de pointage soit encore la seule condition où le geste dépend de la stimulation visuelle, le temps nécessaire à la préparation du geste n'entache pas le délai mis pour produire la parole. Ce résultat peut sembler en contradiction avec les résultats de l'expérience précédente à première vue. Cependant, cela est explicable assez facilement : dans l'expérience précédente, la planification de la parole était réalisable alors que les stimuli visuels n'étaient pas visibles et, dès lors que les dessins apparaissaient, les participants produisaient la parole (et le geste associé le cas échéant) ; seul le geste de pointage nécessitait un délai supplémentaire de planification (infaisable sans l'apparition des stimuli visuels), ce délai supplémentaire influait sur le début de la parole, les deux modalités étant coordonnées. Ici, au contraire, l'apparition des images est nécessaire pour pouvoir planifier la parole (et le geste de pointage). Or, on n'observe aucune différence dans

les instants de début de parole pour les conditions parole seule et parole + geste manuel (en particulier, pour le geste de pointage !), ceci semble montrer que la préparation du geste (peu importe le type de geste) se fait en parallèle de la préparation de la parole et que cette préparation est moins longue. En effet, si la planification du geste de pointage était plus longue que celle de la parole, alors on observerait une différence entre les onsets vocaux dans la condition pointage par rapport aux autres conditions. Ces résultats sont en accord avec ce que trouvaient Levelt [114] et Feyereisen [52].

Il est cependant à noter que, contrairement à ce qui était présenté dans l'expérience précédente, le nombre réduit de types d'images et de couleurs rend la tâche un peu plus complexe puisque le choix du stimulus visuel utile pour la correction (et donc à désigner dans la condition pointage) n'est pas évident.

3.5 Synthèse du chapitre

Cette étude a permis de montrer que la coordination parole/geste manuels est régie par une multitude de facteurs. L'expérience précédente montrait (en particulier) l'influence de l'emplacement de la focalisation prosodique dans la coordination entre geste de pointage et parole, ici il a été montré que la focalisation n'était pas le seul mécanisme à entrer en jeu pour expliquer cette coordination et donc que la modification du *lien communicatif* avait une influence sur la coordination geste manuel/parole. En particulier, il a été montré que la modification du lien communicatif entre geste de pointage et parole dans le cadre de cette tâche de désignation amenait à une coordination avec la parole différente que dans l'expérience précédente : deux comportements distincts ont été identifiés. Dans la condition de pointage, pour certains participants (Groupe 1, 3 personnes sur 10), il importe (comme dans l'expérience précédente) de rendre cooccurrentes les fonctions de désignation, c'est à dire focalisation prosodique et pointage manuel tandis que pour d'autres participants (Groupe 2, 7 personnes sur 10), ce qui "pilote" la coordination est plutôt l'entité désignée dans son ensemble. Cette étude par groupe reste cependant à prendre avec précautions. Ceci permet d'aller plus avant dans l'hypothèse que les gestes (en particulier le geste de pointage) sont liés temporellement à la parole par des mécanismes de plus haut niveau que des simples entraînements moteurs.

Pour ce qui est de la coordination gestes manuels/parole, il a été montré que dans cette expérience, les gestes manuels n'étaient pas forcément cooccurrents à la focalisation (surtout pour le pointage, en condition *Foc2*) et que donc, contrairement à l'expérience précédente, l'attirance par la focalisation était moins forte. Globalement, si les productions sont similaires à ce qui était observé dans l'expérience précédente en condition *Foc1*, ce n'est pas le cas en condition *Foc2* où les gestes manuels sont produits plus tôt. Ce décalage peut-être expliqué par la production plus précoce de la focalisation pour les gestes de battement et de contrôle mais pas pour le geste de pointage. Finalement, les alignements qui ont pu être observés semblent intervenir entre les instants manuels et les cibles articulatoires annotées sur le signal de parole, comme dans l'expérience précédente.

L'étude s'est ici encore trouvée limitée par des productions du geste de battement qui souffrent de la même haute variabilité que précédemment et dont les productions ont été difficiles à exécuter pour certains. De la même façon, les productions du geste de contrôle, bien que différentes sur leur organisation temporelle, semblent partager beaucoup de points avec les productions de geste de pointage, ce qui mène à s'interroger sur le bien fondé de son appellation "contrôle"…

Chapitre 4

Étude de la coordination entre gestes manuels et parole en interaction

Les études présentées dans les parties précédentes ont cherché principalement à faire émerger des stratégies de coordination lors de la réalisation d'une tâche de désignation. Il a été montré un rôle particulier pour le geste de pointage qui entretient naturellement une relation intime avec la focalisation prosodique dans le cadre de la désignation. Ces études ont été menées dans un environnement très contrôlé, en évitant au maximum ce qui pourrait influencer l'exécution à la fois de la parole et des gestes manuels. L'avantage de cette pratique est le fait que les données ainsi acquises sont le plus simple possibles et fournissent *a priori* des résultats pouvant permettre de faire des hypothèses sur les motifs de coordination entre gestes manuels et parole. Ceci mène cependant à une limitation évidente : la désignation n'est utile que si on s'adresse à quelqu'un... Bien que dans les expériences présentées ci-avant, il ait été demandé aux participants d'effectuer la correction pour les expérimentateurs assis à leurs côtés, il est possible que cette contrainte ait été sous-estimée au fil des essais, les expérimentateurs se contentant de vérifier les productions, mais ne donnant aucun *feedback*. Aussi, il apparaît dès lors intéressant de plonger les participants dans une réelle situation d'interaction afin d'étudier la coordination dans un environnement moins contrôlé : c'est l'objectif central de l'expérience décrite dans cette partie.

Par ailleurs, une question d'intérêt dans la suite est l'influence de l'environnement communicatif sur la coordination gestes manuels/parole. En effet, il a été argumenté depuis longtemps que l'adaptation de la parole à l'environnement (typiquement, un environnement bruité) se traduit par des modifications diverses sur la parole, c'est l'effet Lombard. Cet effet est réalisé selon les théories soit dans un but d'augmentation de l'intelligibilité du message pour l'interlocuteur soit par réflexe (voir la thèse de Garnier [56] pour plus de détails). L'intelligibilité d'un message multimodal est constitué non seulement par l'intelligibilité propre de chaque mode mais également par l'intelligibilité de leurs interactions : chaque mode différent ne doit pas perturber les autres modes pour être "efficace". Ainsi, il est probable que, si la coordination entre gestes manuels et parole a un rôle dans l'intelligibilité du message (ce qui est le cas, d'après les études perceptives, voir par exemple [138]), celle-ci soit soumise à des contraintes plus fortes lors de communication dans un environnement adverse.

Ensuite, les deux expériences précédentes se sont intéressées à la coordination dans une tâche de désignation entre les gestes manuels et la focalisation prosodique (contrastive), qui assurait alors exclusivement la fonction de deixis. Nous testons dans cette nouvelle étude l'utilisation d'une autre solution pour « montrer avec la parole » impliquant l'usage des démonstratifs comme présenté dans la Section 1.3.4.

Enfin, seuls les gestes de pointage seront étudiés dans cette partie, c'est en effet pour ces gestes que l'étude de la coordination a le plus de sens dans le cadre d'une tâche de désignation "naturelle" puisque parole et geste manuel (de pointage) ont alors en commun un rôle de désignation.

Cette étude s'intéresse donc à la coordination entre production de gestes manuels et production de parole dans une condition d'interaction. Par rapport aux expériences précédentes, la situation d'interaction, plus naturelle, rend l'étude plus écologique, l'utilisation d'une autre modalité pour la désignation en parole (utilisation de démonstratifs) permet d'étudier une potentielle généralisation des résultats précédents et l'étude de l'influence

de la perturbation de la situation communicative est une troisième ouverture proposée par l'expérience ici présentée.

4.1 Questions de recherche

Cette expérience a pour but d'étudier la coordination entre gestes manuels et parole dans un environnement interactif peu contraint et d'estimer l'influence de l'environnement sur la coordination et sur les stratégies mises en place dans l'interaction. Pour ce faire, deux conditions sont comparées : *sans bruit* et *avec bruit*, l'influence de ces conditions sur la communication est étudiée.

La coordination est étudiée dans le cadre d'une tâche de désignation. Il a été montré dans les expériences précédentes une coordination entre la focalisation contrastive prosodique et plusieurs types de gestes (pointage, battement ou appui sur un bouton) : la focalisation attire le geste manuel. Dans ces expériences, la fonction de désignation était réalisée en parole par la production de focalisation prosodique. Un objectif de la présente étude est de déterminer si la coordination est liée à la fonction de désignation quelle que soit sa modalité (prosodie dans les études précédentes, démonstratifs dans l'étude présente) ou si le changement de modalité de désignation change la nature de la coordination. C'est donc la coordination entre la réalisation de gestes de pointage et la production de démonstratifs qui est étudiée ici, dans le but de tenter de généraliser les résultats précédents sur le lien entre focalisation et pointage.

Une autre question d'interêt est l'influence de la perturbation de la situation de communication sur la coordination entre gestes manuels et parole. La diffusion d'un bruit de conversations lors d'une interaction perturbe celle-ci et modifie les stratégies mises en place par les interlocuteurs, par exemple : la parole augmente en intensité, les gestes articulatoires sont plus amples [57], si le bruit ambiant devient trop intense, la parole est abandonnée au profit d'un langage signé comme cela est le cas dans les scieries par exemple [87, 110]. Ici, la même tâche est comparée dans une condition bruitée et une condition non bruitée. Ainsi, il est possible de comparer les mêmes productions multimodales dans les deux conditions et, par conséquent, de voir l'influence de la perturbation de la communication. Il est intéressant de voir si la perturbation de la situation communicative renforce les liens de coordination entre les deux modalités ce qui permet d'estimer l'importance de cette coordination sur "l'efficacité" communicative et de déterminer si la finesse de cette coordination a une influence sur la communication.

4.2 Protocole expérimental

4.2.1 Corpus

Dans cette expérience visant un mode d'interaction plus naturel, nous avons cependant maintenu une contrainte forte sur le corpus, afin de garantir un certain niveau de contrôle sur les productions vocales, et de régularité des mécanismes de pointage. Nous verrons dans le chapitre suivant comment explorer des situations totalement libres. Les productions vocales de cette expérience sont réduites à un modèle de phrase. Les participants sont tenus de respecter ce modèle de phrase lorsqu'ils veulent donner une consigne à leur partenaire dans l'interaction. Le corpus complet étudié est représenté en Table 4.1. Le modèle de phrase utilisé est *Le <OBJET> va là.* tous

Le bambou va là.
Le bonbon va là.
Le chameau va là.
Le chapeau va là.
Le lama va là.
Le pompon va là.

Table 4.1 – Corpus de l'expérience en interaction

les *<OBJET>* sont composés de deux syllabes de type Consonne-Voyelle (CV). Le démonstratif "*là*" est composé d'une syllabe CV. Toutes les voyelles utilisées sont des voyelles soit ouvertes (/a, ã/) soit protruses (/u, õ, o/).

Les différents "mots-cible" utilisés ont des "articulations" volontairement proches ce qui les rend difficilement discernables lorsque la communication est perturbée. En particulier, les articulations de *bambou* et *bonbon* ne sont discernables visuellement que par le phonème voisé de la première syllabe qui est protrus dans un cas seulement ; les articulations de *bonbon* et *pompon* et de *chameau* et *chapeau* ne sont pas discernables visuellement ; *lama* est le seul mot qui n'est pas ambigu visuellement.

4.2.2 Participants

Dix hommes et dix femmes ($\mu_{\text{age}}= 26,87$, $\sigma_{\text{age}}= 7,66$) ont pris part à l'étude. Tous les participants étaient droitiers d'après le "Edinburgh inventory" d'Oldfield [139], ne présentaient aucun trouble de l'audition et avaient une vision normale ou corrigée sans troubles lors de l'expérience.

4.2.3 Conditions expérimentales

Une seule variable indépendante est étudiée dans cette expérience. La variable étudiée est la variable BRUIT dont les deux niveaux sont *Sans bruit* et *Avec bruit*.

Un environnement bruyant est utilisé dans cette étude afin de dégrader la qualité du signal audio reçu par les deux interlocuteurs. L'effet du bruit peut être considéré au niveau de la parole comme un "allongement des distances" entre les interlocuteurs (*cf.* [176]). Une grande majorité des études de parole dans le bruit à été réalisée en utilisant un bruit à large bande spectrale (bruit blanc, parfois filtré par bandes, ou avec une enveloppe se rapprochant de la parole –*cf.* [47] par exemple). Ce bruit est intéressant car très perturbateur. Il n'est cependant pas très naturel, ce qui pose problème dans le cas de l'expérience en cours.

Des études plus récentes ont mis au point des bruits plus proches de la réalité, en particulier l'utilisation d'un bruit de type *cocktail-party* [188] est devenue plus fréquente dans les études souhaitant plonger des interlocuteurs dans un environnement bruité sans que celui-ci ne paraisse trop artificiel. C'est ce type de bruit qui est choisi par la suite. Le bruit de type *cocktail-party* couvre bien les fréquences de la parole (en dessous de 800 Hz, *cf.* Figure 4.1), et donne l'impression à l'auditeur d'être plongé dans une situation d'ambiance où beaucoup de personnes parlent autour de lui (dans un bar, un restaurant…). Le bruit utilisé est celui de la base de donnée BD_BRUIT [188] et est constitué de 8 voix mixtes, inintelligibles.

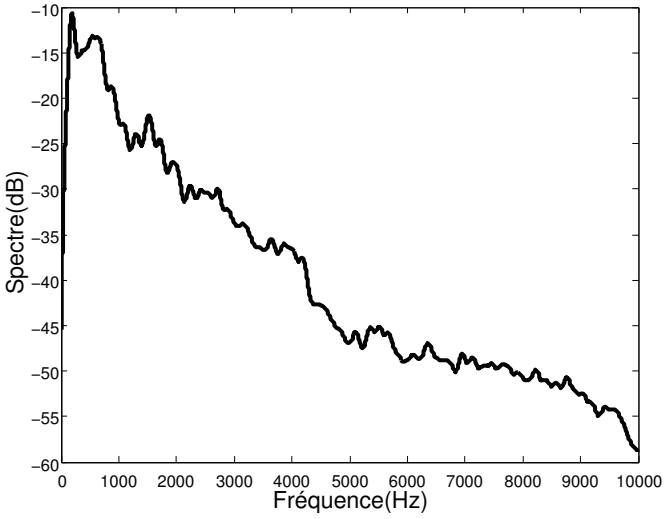


FIGURE 4.1 – Spectre du bruit *cocktail-party* utilisé

Le niveau de bruit est fixé de façon à ce que celui-ci perturbe la communication vocale mais ne la rende pas impossible : le message vocal est perceptible mais difficile à comprendre (si on se limite au canal acoustique).

4.2.4 Dispositif expérimental

La tâche se déroule en interaction avec un complice. Le complice a à sa disposition une réserve de cartes qu'il doit disposer sur un plateau de jeu. Sur ces cartes, sont dessinés des motifs relatifs aux mots du corpus (*bambou, bonbon, chameau, chapeau, lama, pompon*), comme présenté en Table 4.2. Le participant voit un modèle de disposition des images qui est invisible par le complice. Le but de la tâche est, pour le participant, de donner des instructions au complice afin de reproduire le modèle de manière coopérative. Chaque dessin apparaît à deux endroits différents dans le modèle, ceci afin d'augmenter (légèrement) la difficulté de la tâche.

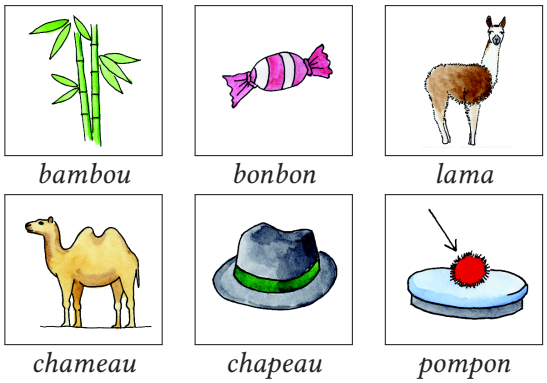


TABLE 4.2 – Cartes à la disposition du complice

4.2.4.1 Élements du dispositif

L'expérience s'est déroulée en chambre sourde sur le site Stendhal du GIPSA-LAB. Une représentation schématique de l'organisation de la salle est donnée en Figure 4.2.

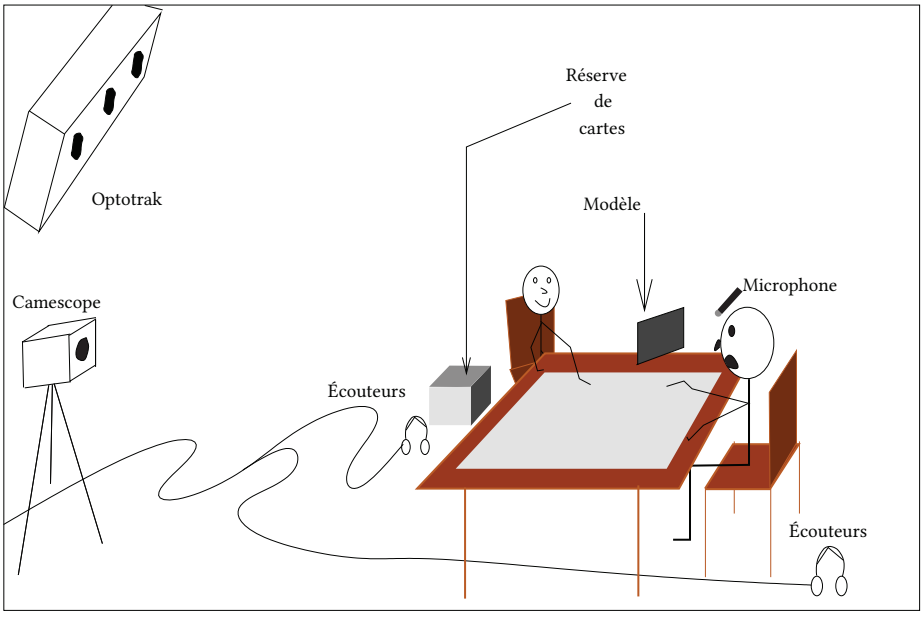


FIGURE 4.2 – Organisation de la chambre sourde pour l'expérience en interaction

La zone de jeu (70 cm de long, 53 cm de large) est disposée sur une table dans la salle avec un recul important par rapport à l'Optotrak (Optotrak 3020) afin que toute la surface de la zone de jeu ainsi que le participant soient couverts par le champ d'acquisition de l'Optotrak. Face à l'Optotrak, disposé au fond de la salle en hauteur, est assis le participant équipé avec les diodes Optotrak (comme présenté ci-dessous). Face au participant (à une distance de 2 m environ), mais légèrement décalé sur la droite, se trouve le complice qui dispose d'une réserve de cartes à côté de lui. Hors de l'axe participant-complice, sur la droite du participant, se trouve le modèle à reproduire diffusé sur un écran. Ce modèle est invisible pour le complice. L'écran sert également pour la diffusion écrite des consignes.

Une caméra vidéo est installée sous l'Optotrak et filme le plateau de jeu ainsi que le participant. Dans les conditions le nécessitant, le bruit est diffusé par l'intermédiaire de casques fermés (AKG-K77) identiques pour les deux interlocuteurs. Le bruit diffusé dans les deux casques est strictement identique : un doubleur jack$\rightarrow 2\times$jack est utilisé pour dédoubler le signal de bruit, les deux casques étant identiques, le même niveau de bruit est atteint dans les deux casques. Le niveau de bruit est fixé à 85 dB(C).

Sur le plateau de jeu (aussi appelé grille par la suite) se trouvent, en des lieux définis par avance, des points de faible taille représentant les emplacements-cibles correspondant aux emplacements des images sur le modèle.

La diffusion des modèles et du bruit *cocktail-party* (dans les conditions bruitées) se fait grâce au logiciel Presentation [135] : Le complice dispose d'un bouton qui lui permet d'avancer dans les étapes de l'expérience (présentation des consignes, début d'un essai –éventuellement, diffusion du bruit–, affichage du modèle, fin de l'essai –éventuellement, fin de diffusion du bruit–, passage à l'essai suivant ...).

L'enregistrement des données est réalisé par trois dispositifs distincts : un camescope numérique filme la scène, il est démarré manuellement en début d'expérience ; l'Optotrak filme les diodes émettrices infra-rouge (suivi de mouvement) et le micro (AKG C1000S) branché à un enregistreur multipistes (korg D1200) permet l'enregistrement du signal audio. Les dispositifs permettant l'enregistrement des données audio, de suivi de mouvement et permettant l'exécution du programme de diffusion des stimuli se trouvent à l'extérieur de la pièce afin de limiter les bruits parasites et d'avoir de l'espace dans la chambre sourde. Enfin, comme dans les expériences antérieures, les enregistrements audio et de suivi de mouvement sont synchronisés par un signal envoyé simultanément sur la piste audio enregistrée et sur une diode Optotrak.

4.2.5 Description de la tâche

Les tâches sont totalement similaires dans les conditions bruitée et non bruitée. Lorsqu'un essai débute, un modèle du même type que celui présenté en Figure 4.3 est diffusé sur l'écran et reste affiché jusqu'à la fin de l'essai (lorsque le complice signale que l'essai est terminé). Chaque grille est composée de douze images (chaque image du corpus apparaît deux fois) dont les positions changent à chaque essai. Le complice dispose d'une réserve de

FIGURE 4.3 – Exemple de modèle diffusé au participant

cartes composée des éléments présentés en Table 4.2. Le participant doit communiquer avec le complice afin que celui-ci dispose ses cartes sur la zone de jeu d'une façon similaire à ce qui est représenté dans le modèle.

Pour ce faire, le participant doit donner des indications au complice en utilisant les phrases du corpus (présentées en Table 4.1). Le participant est libre de dire ce que bon lui semble mais il doit utiliser les phrases du corpus lorsqu'il veut donner des indications de placement au complice : les autres phrases ne sont pas prises en compte pour le placement d'objets. Le complice n'interagit qu'en positionnant les cartes. Il n'est pas autorisé à poser des questions pour obtenir plus de précisions sur une position : si un énoncé a mal été compris, le participant doit répéter une phrase du corpus. Le participant est par ailleurs libre de ses mouvements, la seule contrainte étant qu'il doit rester assis sur sa chaise.

Les participants commencent par une phase d'apprentissage des cartes disponibles pour le complice afin de voir s'ils utilisent les bons mots pour désigner ces cartes (et ils apprennent le bon mot le cas échéant). Une phase d'entraînement est ensuite proposée au préalable de l'expérience. Cette phase d'entraînement sert à se familiariser avec la tâche et son but. L'entraînement est répété jusqu'à ce que le participant se sente à l'aise avec la tâche. Suivant cette phase d'entraînement, la condition sans bruit est réalisée (3 essais) puis la condition bruitée est réalisée (3 essais). Pour tous les participants, l'ordre de passage des facteurs de la condition BRUIT est fixe, ceci afin d'obtenir les productions les plus naturelles possibles lors de la condition non bruitée (le participant n'était pas prévenu par avance de la passation d'une condition bruitée). Ce choix sera discuté par la suite. Parmi les essais et entre les participants, les emplacements des différents objets sur le modèle sont randomisés.

Des exemples de consignes sont donnés dans la Figure A.3, p.159.

L'interdiction pour le complice de répliquer vient du fait qu'il est très compliqué pour le participant de respecter la consigne d'utilisation de phrases fixées si les deux interlocuteurs engageaient un dialogue. Cette décision permet de s'assurer que les données collectées sont utilisables pour un traitement statistique ultérieur. Bien que ceci rende *a priori* la situation moins naturelle, ce silence du complice ne rend pas la situation non interactive : il y a bien interaction entre les deux interlocuteurs.

4.2.6 Acquisition des données

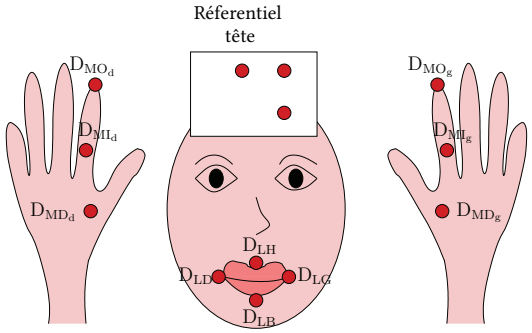
Vidéo L'enregistrement vidéo est lancé en début d'expérience. Le cadrage de la vidéo est réalisé afin de filmer à la fois le plateau de jeu et le participant et d'apercevoir le complice poser les cartes sur le plateau lorsque c'est le cas. L'utilisation de la vidéo est justifiée par le fait que des gestes *a priori* plus complexes que des gestes de pointage peuvent apparaître lors de la réalisation de la tâche (typiquement, des gestes iconiques, ou des gestes de battement, ...). L'occurrence de tels gestes est complexe à détecter à partir des données Optotrak, le visionnage de la vidéo permet quant à lui de classifier rapidement les gestes.

4.2.6.1 Audio

Le signal audio de parole a été enregistré au format WAV, 16 bits, avec une fréquence d'échantillonnage de 44.100 Hz.

4.2.6.2 Capture de mouvements

Le suivi de mouvement par Optotrak a été réalisé à $f_e = 150$ Hz. La résolution temporelle est ici légèrement inférieure à ce qui était fait dans les expériences précédentes pour des raisons techniques (un nombre plus important de diodes impose une fréquence d'échantillonnage plus basse). Le positionnement des diodes est similaire à celui des expériences précédentes mais des diodes sur la main gauche ont été rajoutées puisqu'il n'y a aucune raison de penser que les participants droitiers n'utilisent *jamais* leur main gauche pour réaliser des gestes manuels. Par ailleurs, la position des diodes sur les mains est légèrement modifiée pour tenir compte des différentes configurations de la main utilisées par les participants. En particulier, la diode du dos de la main a été déplacée vers le pouce afin de rester visible par les caméras lorsque le pointage est réalisé avec le pouce vers le plafond (paume verticale). La diode sur l'index a été reculée vers la première phalange afin de rester visible lorsque les pointages se font avec l'index à la verticale et le dos de la main orienté vers le plafond. L'emplacement des diodes émettrices est représenté en Figure 4.4.



Nom diode	Emplacement
$D_{MD_{d/g}}$	Dos main droite/gauche
$D_{MI_{d/g}}$	Première phalange index main droite/gauche
$D_{MO_{d/g}}$	Ongle index main droite/gauche
D_{LH}	Centre lèvre supérieure
D_{LB}	Centre lèvre inférieure
D_{LD}	Commissure lèvres droite
D_{LG}	Commissure lèvres gauche

FIGURE 4.4 – Emplacement des diodes Optotrak

4.3 Traitement des données

4.3.1 Traitements préliminaires

4.3.1.1 Validation des données

Les contraintes de validité pour les productions des participants étant faibles (le but est d'étudier leur comportement, en imposant le moins de contraintes possibles), peu de productions sont considérées comme "invalides" dans les données. Seules les productions contenant des énoncés n'appartenant pas au corpus ont été écartées des analyses. Ces données ont été repérées lors d'une phase d'annotation des vidéos (décrite au chapitre suivant) et elles ont été séparées en deux catégories : les phrases hors-corpus (contenant toute phrase sans rapport avec le placement des cartes, les productions dont le contenu ne contient pas les mots du corpus ...), les phrases du corpus erronées (phrase du corpus contenant une erreur, une hésitation, ...) ou les phrases interrompues (pause, éclat de rire,...). Les phrases hors-corpus représentent $7,23\%$ des données, les phrases du corpus erronées $5,30\%$ des données, les autres énoncés sont considérés comme valides. Le nombre total de productions prises en compte dans les analyses suivantes (données valides) est de 701 énoncés en condition non bruitée et 749 productions en condition bruitée.

4.3.1.2 Algorithme d'interpolation

L'algorithme mis au point pour estimer la position d'une diode quand celle-ci est masquée à un instant donné t repose sur la connaissance de la position relative des diodes voisines à des instants proches de l'instant t. Les coordonnées tridimensionnelles des diodes voisines (diodes de la main, diodes autour des lèvres) sont regroupées dans une matrice comme représenté en Figure 4.5 (pour les diodes de la main). Le numéro de ligne représente le numéro d'échantillon (*i.e.* le temps). Lorsqu'une diode est masquée, ses coordonnées sont toutes inconnues et sont notés Not A Number (NAN). Le but de l'algorithme d'interpolation est d'obtenir une matrice ne contenant aucune valeur inconnue.

$$\begin{pmatrix} x_{\mathrm{MO},1} & y_{\mathrm{MO},1} & z_{\mathrm{MO},1} & x_{\mathrm{MI},1} & y_{\mathrm{MI},1} & z_{\mathrm{MI},1} & x_{\mathrm{MD},1} & y_{\mathrm{MD},1} & z_{\mathrm{MD},1} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{\mathrm{MO},i} & y_{\mathrm{MO},i} & z_{\mathrm{MO},i} & x_{\mathrm{MI},i} & y_{\mathrm{MI},i} & z_{\mathrm{MI},i} & x_{\mathrm{MD},i} & y_{\mathrm{MD},i} & z_{\mathrm{MD},i} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{\mathrm{MO},j} & y_{\mathrm{MO},j} & z_{\mathrm{MO},j} & x_{\mathrm{MI},j} & y_{\mathrm{MI},j} & z_{\mathrm{MI},j} & NAN & NAN & NAN \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ NAN & NAN & NAN & NAN & NAN & NAN & NAN & NAN & NAN \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{\mathrm{MO},n} & y_{\mathrm{MO},n} & z_{\mathrm{MO},n} & x_{\mathrm{MI},n} & y_{\mathrm{MI},n} & z_{\mathrm{MI},n} & x_{\mathrm{MD},n} & y_{\mathrm{MD},n} & z_{\mathrm{MD},n} \end{pmatrix} \begin{matrix} \text{t=1} \\ \vdots \\ \text{t=}i\text{, les trois diodes sont visibles} \\ \vdots \\ \text{t=}j\text{, la diode } \mathrm{D_{MD}} \text{ est masquée} \\ \vdots \\ \text{t=}k\text{, les trois diodes sont masquées} \\ \vdots \\ \text{t=}n \end{matrix}$$

FIGURE 4.5 – Exemple de matrice représentant les données spatiales manuelles

Pour un instant t donné, deux cas différents où une interpolation est nécessaire se présentent (représentés dans la Figure 4.5) :

- Au moins une diode est visible (sur la figure, $t = j$)
- Les trois diodes sont cachées (sur la figure, $t = k$)

Dans le cas où au moins une diode est visible, on estime la position des autres diodes en faisant l'hypothèse selon laquelle sur un temps (très) court, le polygone formé par les diodes voisines est indéformable. Pour chaque ligne de la matrice, on calcule les vecteurs entre chaque couple de points (dans l'exemple, pour chaque ligne, on calcule $\overrightarrow{\mathrm{D_{MO}D_{MI}}}$, $\overrightarrow{\mathrm{D_{MO}D_{MD}}}$, $\overrightarrow{\mathrm{D_{MI}D_{MD}}}$). Lorsque les coordonnées d'une diode sont manquantes à un instant t, on interpole sa position à partir des vecteurs-position calculés aux instants antérieurs. Dans l'exemple, on trouve l'emplacement de $\mathrm{D_{MD}}$ à l'instant j de la façon suivante :

- Soit $(g, g\prime) \in \mathbb{N}^2$ le couple des plus petits entiers tels que $\begin{matrix}\mathrm{D_{MI},}_{j-g} \text{ et } \mathrm{D_{MD},}_{j-g} \\ \mathrm{D_{MO},}_{j-g\prime} \text{ et } \mathrm{D_{MD},}_{j-g\prime}\end{matrix}$ soient conjointement visibles
- On estime alors la position de $\mathrm{D_{MD},}_{j}$ par la formule :

$$\begin{cases} \overrightarrow{O\mathrm{D_{MD},}_{j}} = \overrightarrow{O\mathrm{D_{MI},}_{j}} + \overrightarrow{\mathrm{D_{MI},}_{j-g}\mathrm{D_{MD},}_{j-g}} & \text{, si } \mathrm{D_{MO}}_{j} \text{ n'existe pas} \\ & \text{, si } \mathrm{D_{MO}}_{j} \text{ et } \mathrm{D_{MI}}_{j} \text{ existent et } g < g\prime \\ \overrightarrow{O\mathrm{D_{MD},}_{j}} = \overrightarrow{O\mathrm{D_{MO},}_{j}} + \overrightarrow{\mathrm{D_{MO},}_{j-g\prime}\mathrm{D_{MD},}_{j-g\prime}} & \text{, si } \mathrm{D_{MI}}_{j} \text{ n'existe pas} \\ & \text{, si } \mathrm{D_{MO}}_{j} \text{ et } \mathrm{D_{MI}}_{j} \text{ existent et } g\prime < g \end{cases}$$

Cet algorithme est appliqué sur la matrice contenant les coordonnées des diodes organisées par temps croissant, la matrice résultante (M_c) ne contient que des lignes soit sans aucun NAN soit avec seulement des NANs et les données manquantes ont été interpolées par les données antérieures. Dans un deuxième temps, le même algorithme est appliqué sur la matrice des coordonnées organisées par temps décroissant, la matrice résultante (M_d) ne contient que des lignes totalement constituées de NAN ou sans donnée manquante. On note $x^{M_c}_{\mathrm{MO},t}$ (resp. $x^{M_d}_{\mathrm{MO},t}$) la première coordonnée de $\mathrm{D_{MO}}$ au temps t dans la matrice interpolée M_c (resp. M_d). Ces deux matrices sont alors "fusionnées" pour trouver la matrice interpolée finale. La fusion est expliquée par un exemple : si on suppose MO invisible entre les instants t_1 et t_2, $\forall t \in \mathbb{N}, t \in [t_1; t_2]$, on pose $r = \frac{t-t_1}{t_2-t_1}$ et la fusion s'écrit de la façon suivante (pour $x_{\mathrm{MO},t}$, les mêmes raisonnements sont applicables pour les trois coordonnées de $\mathrm{D_{MO}}$) :

$$x_{\mathrm{MO},t} = \begin{cases} x^{M_c}_{\mathrm{MO},t} & \text{si } x^{M_d}_{\mathrm{MO},t} \text{ n'existe pas} \\ x^{M_d}_{\mathrm{MO},t} & \text{si } x^{M_c}_{\mathrm{MO},t} \text{ n'existe pas} \\ x^{M_c}_{\mathrm{MO},t} * (1-r) + x^{M_d}_{\mathrm{MO},t} * r & \text{sinon} \end{cases}$$

. Cette façon de procéder permet de faire une interpolation donnant plus de poids à l'interpolation obtenue grâce aux données antérieures lorsqu'on estime une position en début d'intervalle contenant des NANs et donnant plus de poids à l'interpolation obtenue par les données postérieures lorsqu'on estime une position en fin d'intervalle contenant des NANs. L'interpolation obtenue est différente d'une interpolation linéaire puisque pour chaque instant t on prend en compte les interpolations *au temps* t trouvées avec les valeurs antérieures et postérieures (on ne prend pas en compte que les interpolations aux temps t_1 et t_2 pour trouver la valeur interpolée au temps t).

L'hypothèse de travail (les diodes sont placées sur un polygone indéformable pendant un court laps de temps) se vérifie aisément sur des laps de temps de l'ordre du dixième de seconde voire du centième de seconde pour les mouvements de la main et de la bouche. L'interpolation des positions est moins correcte sur des intervalles de temps longs. En particulier, bien que la position de la main lors d'un pointage puisse généralement être assimilée à une forme indéformable, ce n'est pas du tout le cas des lèvres lors de l'articulation d'une phrase. Cependant, les données liées aux lèvres ne sont pas cachées (le participant est en face des caméras Optotrak et le complice ne bloque pas le champ de vision des cameras pour les lèvres), donc la méthode est utilisable sur les données.

Une fois cette étape réalisée, les seules lignes de la matrice des coordonnées contenant des données manquantes sont les lignes composées uniquement de coordonnées de diodes invisibles ($t = k$ sur l'exemple). Dans le cas où toutes les diodes sont invisibles, une interpolation linéaire est utilisée pour combler les données manquantes. Par exemple, si toutes les diodes sont invisibles sur un intervalle $t \in]b, c[$, on estime une coordonnée de diode ($x_{\mathrm{MO},t}$ par exemple) à un instant t par : $x_{\mathrm{MO},t} = \frac{c-t}{c-b} \times x_{\mathrm{MO},b} + \frac{t-b}{c-b} \times x_{\mathrm{MO},c}$

Une fois ces deux interpolations réalisées sur un signal, aucune donnée n'est manquante. Les détails des interpolations réalisées sont présentés ci-dessous.

4.3.1.3 Données articulatoires

Les diodes placées autour des lèvres servent à la récupération des données articulatoires. Un traitement similaire à celui présenté en Section 2.2.1.2 a été appliqué à ces données (projection dans un repère lié à la tête, extraction de l'ouverture/protrusion des lèvres, filtrage des signaux). L'étirement des lèvres a également été extrait des données comme étant la distance euclidienne entre les deux diodes placées sur les commissures droite et gauche des lèvres ($\mathrm{D_{LD}}$ et $\mathrm{D_{LG}}$).

Les données articulatoires utilisées par la suite (ouverture/protrusion des lèvres, l'étirement étant simplement utilisé afin de lever certaines ambiguïtés lors des annotations) ne contiennent pas de données manquantes (diodes cachées lors de l'enregistrement), aucune interpolation des données n'est donc nécessaire.

4.3.1.4 Données manuelles

Bien que toutes les précautions aient été prises, les données de trajectoires manuelles recueillies par l'enregistrement Optotrak contiennent des données manquantes. En particulier, les diodes $\mathrm{D_{MO_g}}$ et $\mathrm{D_{MO_d}}$ ne sont pas toujours visibles (le masquage des diodes est parfois dû au fait que l'index est orienté dans une direction non visible pour les caméras, et souvent dû au fait que le complice doit passer dans le champ de vision de l'Optotrak pour déposer les cartes sur le plateau de jeu). Pour chaque main et pour chaque diode, les données issues des trois diodes ont été utilisées afin d'estimer la position de la diode lorsque celle-ci manquait.

L'annotation des données est réalisée en utilisant le signal de vitesse tridimensionnelle, pour un point M ayant une trajectoire tridimensionnelle représentée dans l'espace cartésien par des coordonnées $\overrightarrow{OM} = \begin{pmatrix} x \\ y \\ z \end{pmatrix}$, le vecteur vitesse s'écrit $\overrightarrow{v} = \frac{d\overrightarrow{OM}}{dt} = \begin{pmatrix} \frac{dx}{dt} \\ \frac{dy}{dt} \\ \frac{dz}{dt} \end{pmatrix}$ et sa norme est la vitesse tridimensionnelle : $v_M = \sqrt{\frac{dx}{dt}^2 + \frac{dy}{dt}^2 + \frac{dz}{dt}^2}$.
Ce signal de vitesse est le signal sur lequel se basent les annotations.

Afin de permettre une représentation graphique des signaux de trajectoire spatiales, il a également été décidé d'effectuer une analyse en composantes principales comme dans les expériences précédentes. Les valeurs issues de cette analyse ne sont utilisées qu'à des fins de représentation graphique, les valeurs ne sont en aucun cas utilisées dans les analyses qui suivent.

Ces signaux sont filtrés par un filtre de Butterworth d'ordre 3 ayant une fréquence de coupure à 15 Hz.

4.3.2 Annotation des données

Les données recueillies sont issues de trois appareils d'enregistrement différents, trois signaux sont donc disponibles pour une annotation. Ici encore, le but est d'étudier la coordination de la parole et des gestes manuels. Les annotations du signal audio et de capture de mouvement doivent être réalisées. La tâche n'étant contrôlée pratiquement qu'au niveau de ce qui est prononcé par les participants, ceci implique que les gestes manuels effectués peuvent être relativement complexes et difficiles à annoter automatiquement comme cela était fait précédemment : les gestes effectués ne sont pas nécessairement des gestes de pointage, il n'y a pas de position de repos (la main peut donc avoir des "positions de repos" en l'air, se déplacer entre deux gestes, etc…). Par ailleurs, il est impossible de savoir quel geste est associé à quel énoncé à partir des signaux audio et de trajectoire seuls (en particulier, si les mouvements sont enchaînés rapidement ou si les participants ne synchronisent pas leurs gestes avec leur parole, etc…)

Ainsi, il a été décidé d'annoter les trois signaux : vidéo, audio et capture de mouvements. L'annotation de la vidéo permet une annotation plus aisée du signal de suivi de trajectoire et fournit un aperçu plus détaillé (que l'audio et le suivi de mouvements seuls) du déroulement de l'expérience, permettant ainsi d'annoter rapidement

quelques informations relatives au déroulement expérimental. Les annotations sont réalisées sous ELAN [184] pour les données vidéo, sous Praat [16] pour la segmentation acoustique et sous Mathworks Matlab [120] pour l'annotation des gestes (manuels et articulatoires) et du suivi de fréquence fondamentale/intensité (réalisés sous Praat).

4.3.2.1 Segmentation et annotation acoustique

Segmentation automatique de la parole Le signal de parole pour chaque grille a été segmenté automatiquement grâce à un logiciel permettant de réaliser un alignement automatique forcé d'une transcription phonétique typique avec un signal audio. Ce logiciel s'appuie sur la théorie des modèles de Markov cachés (HMM) (voir par exemple le tutoriel de Rabiner [151]) et utilise le Hidden Markov Model Toolkit (HTK, [187]). L'identificateur de phonèmes utilisé est basé sur un algorithme de Viterbi. L'outil utilisé a été entraîné sur un corpus constitué d'une voix d'homme et d'une voix de femme. Ces deux modèles sont utilisés selon le sexe du participant afin de segmenter les fichiers audio.

Le niveau de granularité utilisé dans la segmentation est la syllabe : chaque production des participants est "découpée" acoustiquement en syllabes.

Ajustement de la segmentation acoustique, extraction et annotation des paramètres acoustiques Le logiciel Praat [16] a été utilisé pour affiner la segmentation automatique, extraire la fréquence fondamentale et l'intensité. Pour ces deux grandeurs, la fenêtre d'intégration utilisée mesure 10 ms. La fréquence fondamentale est extraite grâce à une méthode d'autocorrélation.

Pour chaque énoncé prononcé par les participants, ont été repérés sur les signaux correspondant au suivi des paramètres acoustiques le maximum et le minimum de ces paramètres pour chaque syllabe (extrema locaux) et le maximum et le minimum au sein de l'énoncé prononcé (extrema globaux).

4.3.2.2 Annotation vidéo

Une annotation vidéo est réalisée afin de guider l'annotation des données de suivi de mouvement et de décrire le déroulement de l'interaction de façon simple et surtout afin de permettre de lier correctement les gestes manuels aux énoncés. Comme dans une large majorité d'études utilisant le support vidéo, une grille d'annotation a été mise au point afin de permettre de réaliser ces deux points. La grille d'annotation mise au point a été élaborée dans le but de pouvoir annoter des interactions non contraintes utilisant un protocole similaire à celui présenté dans cette partie. Cette grille sera présentée dans le Chapitre 5 car sa mise au point constitue une contribution en tant que telle. Dans la présente étude, l'annotation vidéo sert principalement de contrôle global pour l'extraction des gestes manuels correspondant aux phrases du corpus. La direction du regard a également été annotée à la main (globalement, les participants regardent soit le modèle, soit leur interlocuteur, soit le plateau de jeu) en procédant à une annotation précise image par image. Un exemple d'annotation réalisé sous ELAN est donné en Figure 4.6.

4.3.2.3 Annotation des données articulatoires

Les données articulatoires sont annotées selon une méthode similaire à ce qui a été présenté en Section 2.2.2 : l'annotation est semi-automatique et est réalisée selon un patron défini par avance. Une vérification manuelle *a posteriori* de l'annotation automatique est réalisée (et des corrections effectuées si nécessaire). Les points annotés sur les trajectoires d'ouverture et de protrusion sont les maxima rencontrés sur les deux syllabes du mot-cible et sur le démonstratif. Les voyelles ouvertes (/a, ã/) sont annotées sur le signal d'ouverture des lèvres, les voyelles protruses (/u, õ, o/) sur le signal de protrusion. Le signal d'étirement est affiché sur l'interface d'annotation seulement pour indication.

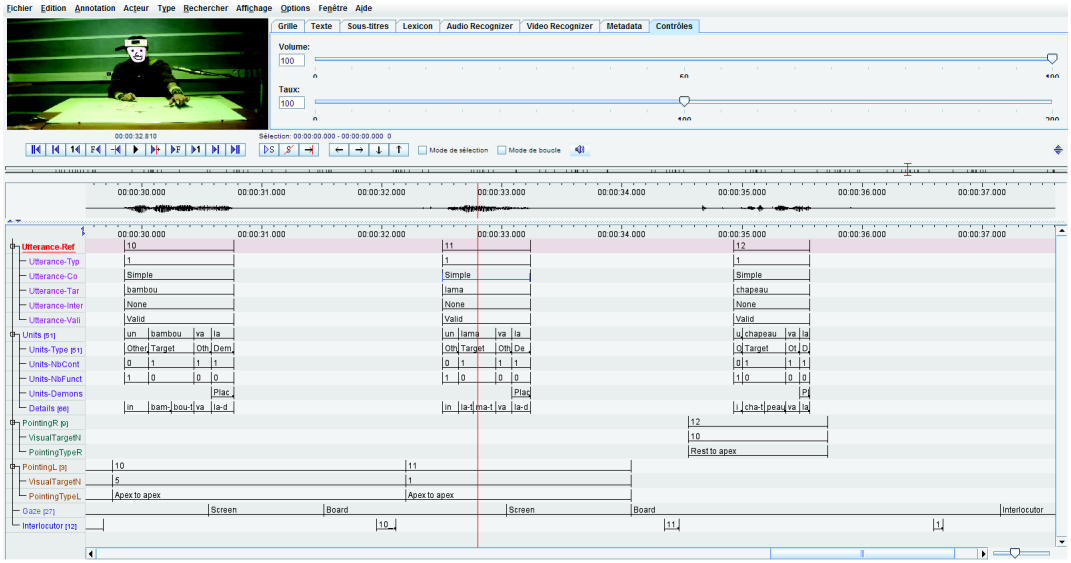


FIGURE 4.6 – Annotation d'une interaction sous ELAN

4.3.2.4 Annotation des données manuelles

Les données manuelles ont également été annotées de manière similaire à ce qui est décrit en Section 2.2.2. Seuls les gestes de pointage repérés lors de l'annotation vidéo sont analysés. L'annotation d'un geste de pointage se fait par la détection de deux pics de vitesse (celui du *stroke* et celui du *stroke de retour* – ou du *stroke* suivant si les pointages sont enchaînés). Une fois ces deux pics de vitesse détectés, on annote les points autour du pic dont les vitesses sont égales à 10% de la vitesse maximale. La détection des pics de vitesse se fait en recherchant ces pics dans un voisinage des instants annotés sur la vidéo : le premier pic de vitesse (du *stroke*) est cherché après l'instant de début du geste annoté en vidéo, le second pic (du *stroke* de retour) est recherché dans un voisinage qui suit l'instant de fin d'apex annoté en vidéo.

La trajectoire de chaque geste est représentée par six points, similaires à ce qui était présenté précédemment : l'onset (P_{On}), le pic de vitesse du *stroke* (P_{vit}), l'apex du geste (P_A), la fin de l'apex (P_R), le pic de vitesse du *stroke* de retour (P_{vit2}) et l'offset du geste (P_{Off}). L'onset est simplement le point précédant le premier pic de vitesse et dont la valeur de la vitesse atteint 10% de la valeur du premier pic de vitesse, l'offset le point suivant le deuxième pic de vitesse et dont la valeur atteint 10% de la valeur du deuxième pic de vitesse. Lorsque les gestes manuels sont enchaînés, les instants P_R, P_{vit2} et P_{Off} du premier geste d'une part et P_{On}, P_{vit}, P_A du second geste d'autre part, sont confondus.

Toutes les données sont annotées automatiquement puis vérifiées *a posteriori* (et modifiées le cas échéant). La vérification *a posteriori* est nécessaire afin de s'affranchir des gestes dont les profils de vitesse s'écartent de la "normale" (gestes hésitants, gestes extravagants, ...). Lorsque cela est nécessaire, la vidéo est mise à contribution pour annoter les gestes dont le profil de vitesse est trop complexe pour être interprété facilement.

4.3.2.5 Prétraitement des données temporelles

Comme présenté dans les expériences précédentes, les données ont été normalisées afin de pallier aux variations segmentales des différents énoncés du corpus : les données temporelles sont normalisées sur la durée de production vocale de l'énoncé correspondant. Le temps 0 correspond au début de production de parole, 1 correspond à la fin de l'énonciation.

Une fois cette normalisation réalisée, les valeurs atypiques sont prises en compte et corrigées. Une valeur est considérée comme atypique lorsqu'elle est hors de l'intervalle $[Q_1 - 1,5 * \text{IQR}; Q_3 + 1,5 * \text{IQR}]$ où Q_1 et Q_3 représentent les valeurs du premier et du troisième quartile et IQR représente l'écart inter-quartile : $Q_3 - Q_1$. La correction des valeurs atypiques est faite en remplaçant ces valeurs par la moyenne des données. La quantité

de données remplacée est donnée en Table 4.3. Dans cette table, chaque cellule représente la proportion de données temporelles remplacées par rapport au nombre total de données temporelles pour chaque "catégorie de donnée" annotée : Syllabes contient toutes les limites acoustiques des syllabes, F_0 contient les instants de pics de fréquence fondamentale, Intensité les instants de pics d'intensité, Articulatoire les cibles articulatoires annotées (2 sur le mot cible et une sur le démonstratif), Geste les cinq instants annotés sur les gestes. Cette table montre qu'une "faible" partie des données sont considérées comme atypique dans les deux conditions *Bruit* et *Sans Bruit* (la quantité de données atypiques est légèrement supérieure en condition bruitée).

	Syllabes	F_0	Intensité	Articulatoire	Geste
Sans Bruit	$5,67\%$	$5,54\%$	$2,23\%$	$6,62\%$	$5,56\%$
Bruit	$7,01\%$	$7,07\%$	$2,15\%$	$7,35\%$	$6,26\%$

La catégorie Syllabes rassemble l'ensemble des limites acoustiques des syllabes (trois pour l'objet et une pour le démonstratif), F_0 contient tous les pics de F_0, Intensité tous les pics d'intensité, Articulatoire toutes les cibles articulatoires (deux pour l'objet, une pour le démonstratif) et Geste les six instants annotés sur les gestes.

TABLE 4.3 – Proportion de données atypiques remplacées dans les données

4.4 Analyses

4.4.1 Description des analyses

Les analyses statistiques ici effectuées sont, comme dans les expériences précédentes sont dans la plupart des cas des tests t de Welch, et lorsque cela est nécessaire, des ANOVAs à mesures répétées. Dans toutes les analyses, sauf dans Section 4.4.2 ou lorsque celà est spécifié, seules les données "valides" (contenant une phrase issue du corpus et un geste) et "sans interruptions" (hésitations, pauses etc…) sont utilisées.

4.4.2 Description globale des interactions

Lors de la passation de l'expérience, il est assez clair que l'interaction dans un environnement bruité est plus difficile que dans un environnement sans bruit et ce, même pour une tâche simple comme celle ici présentée. Cette difficulté accrue se caractérise par plusieurs traits qui sont décrits ci-dessous.

4.4.2.1 Durées de productions vocales et temps de remplissage d'une grille

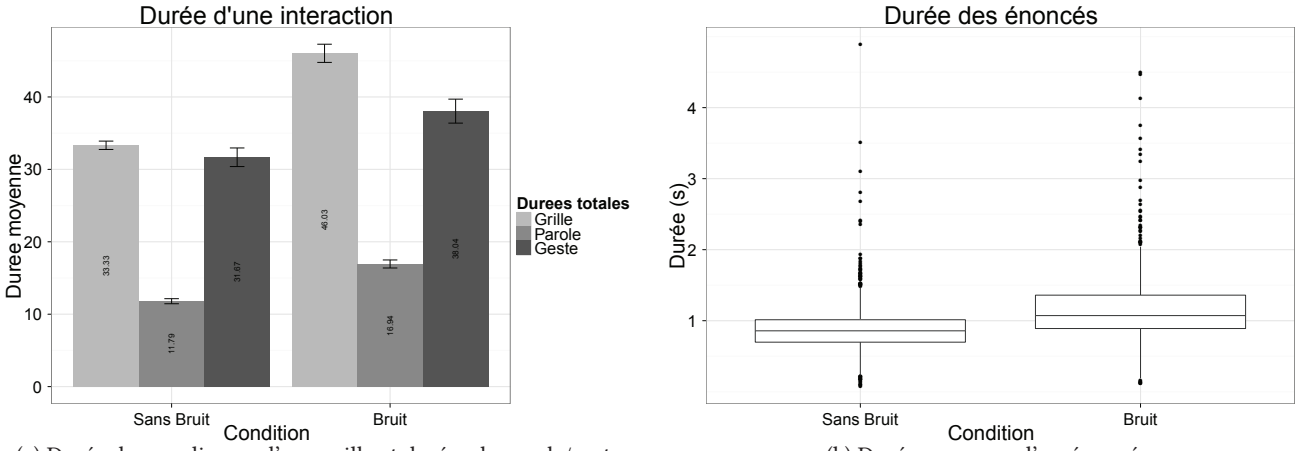
La durée de l'interaction entre les deux interlocuteurs augmente sensiblement entre les deux expériences comme cela est représenté en Figure 4.7a. La durée moyenne de remplissage d'une grille dans la condition non bruitée est de $33,33$ s en condition non bruitée et de $46,03$ s en condition bruitée ($t(19) = -10,63$, $p < 0,001$).

Comme représenté en Figure 4.7b, la durée moyenne d'un énoncé en condition *Bruit* est significativement plus longue qu'en condition *Sans Bruit* ($1,15$ s *vs* $0,89$ s : $t(19) = -8,93$, $p < 0,001$). Cependant, cet allongement de la durée moyenne de la production d'un énoncé ne permet pas d'expliquer à lui seul l'allongement de l'interaction puisqu'en proportion (comme visible sur la Figure 4.7a) la proportion de parole dans l'interaction (par rapport à la durée totale de l'interaction) ne varie pas significativement d'une condition à l'autre ($t(19) = -1,66$, $p = 0,110$).

La principale différence entre les deux conditions, et qui influe beaucoup sur la durée de remplissage de la grille est en fait le nombre d'erreurs commises lors du jeu.

4.4.2.2 Interruptions et corrections

Les interruptions annotées sont les pauses ou hésitations. Les études statistiques montrent une augmentation d'énoncés non valides (*i.e.* globalement, ne respectant pas le modèle de phrase) en passant de la condition *Sans Bruit* à la condition *Bruit* : $+5,28\%$ d'énoncés non valides supplémentaires en condition bruitée ($t(19) = -2,21$, $p < 0,05$)). Cette augmentation du nombre d'énoncés non valides s'accompagne d'une augmentation de la



(a) Durée de remplissage d'une grille et durées de parole/geste associées

(b) Durée moyenne d'un énoncé

FIGURE 4.7 – Durée moyenne d'une interaction et temps moyen passé à faire des gestes manuels / à parler, durée moyenne d'un énoncé dans les conditions *sans bruit* et *avec bruit*

quantité d'énoncés contenant des pauses ($+15,65\%$, $t(19) = -3,83$, $p < 0,01$) bien que la proportion d'énoncés marqués comme contenant une hésitation n'augmente pas ($t(19) = -1,566$, $p = 0,134$).

Le nombre moyen d'énoncés corrigés par grille remplie est représenté dans la Table 4.4. Il est clair que le nombre d'énoncés comportant des corrections est beaucoup plus important en condition bruitée. Dans la table, une autocorrection correspond à un énoncé dans lequel le participant a fait une erreur et se corrige, une correction liée à l'interlocuteur est une correction qui émane soit d'un choix erroné de cible par le complice soit d'un placement erroné par le complice, ces deux types de corrections se répartissent entre corrections portant sur la cible de l'énoncé ou sur l'emplacement de cette cible. Enfin, un dernier type d'énoncé "corrigé" correspond aux énoncés qui sont de simples répétitions d'énoncés précédemment prononcés : aucune erreur (du locuteur ou de l'interlocuteur) n'est corrigée mais l'énoncé est répété à cause d'une incompréhension manifeste du complice.

Block	Total Erreurs	Corrections		Répetitions
		Autocorrections	Interlocuteur	
Sans Bruit	$0,5$	$0,05(10\%)$	$0,4(80\%)$	$0,05(10\%)$
		Cible : $0,35(77,77\%)$		
		Emplacement : $0,10(22,22\%)$		
Bruit	$5,1$	$0,50(9,80\%)$	$4,4(86,27\%)$	$0,20(3,92\%)$
		Cible : $4,35(88,78\%)$		
		Emplacement : $0,55(11,22\%)$		

La nature des corrections est marquée en petite taille sous les corrections effectuées : la correction porte soit sur le mot cible utilisé soit sur l'emplacement désigné.

TABLE 4.4 – Nombre moyen d'"erreurs" par grille et répartition selon les trois catégories autocorrections, erreur due à l'interlocuteur, répétitions

Globalement, le nombre moyen d'énoncés destinés à la correction ou à l'amélioration de la compréhension est multiplié par dix entre les deux conditions. La plupart des corrections sont destinées à corriger une mauvaise compréhension par le complice de l'objet cible à utiliser : la diffusion de bruit perturbe bien l'interaction puisque l'information transmise l'est moins facilement. Par ailleurs, la diffusion de bruit perturbe également la production d'énoncés puisque le nombre d'autocorrections (bien que faible) augmente sensiblement d'une condition à l'autre. Sans surprise, la nature des corrections effectuées est massivement en rapport avec la cible de l'énoncé : la position est le plus souvent exprimée par les gestes de pointage, ce qui réduit le besoin de corrections (les participants prononcent "là" de façon assez automatique). Les corrections sur les cibles dans la condition

bruitée interviennent en général sur les mot-cibles présentant de fortes ressemblances au niveau acoustique : chameau/chapeau (19% et 24% des corrections effectuées dans le bruit) et bonbon/pompon (17% et 25% des corrections dans le bruit) ; en condition non bruitée les erreurs interviennent principalement sur bambou et lama (environ 30% des erreurs chacun).

4.4.2.3 Nombre de mots utilisés

Dans les énoncés prononcés par les participants, les mots utilisés ont été catégorisés en mots lexicaux et mots-outils afin d'avoir une évaluation de la complexité syntaxique des phrases prononcées. Ce décompte est peu intéressant dans le cas d'une expérience avec un corpus contrôlé comme cela est le cas ici. Cependant, bien que le contenu des phrases du corpus soit invariable, ce n'est pas le cas des phrases hors corpus (typiquement, les commentaires, les explications, ...) et le décompte effectué permet donc grossièrement de comparer le contenu de ces phrases. Par ailleurs, il permet d'avoir un décompte précis du nombre de mots utilisés et donc du débit de parole.

En moyenne, le nombre de mots de contenu utilisés est beaucoup plus élevé que le nombre de mots-outils (ce qui est normal, puisque les phrases du corpus sont constituées de trois mots de contenu et d'un mot-outil) : dans la condition *Sans Bruit*, le remplissage d'une grille se fait avec $41,35$ mots-outils et $115,8$ mots de contenu alors que dans la condition bruitée, $48,20$ mots-outils et $127,4$ mots de contenus sont requis. Bien que le nombre de mots soit en augmentation d'une condition à l'autre, cela ne modifie pas la proportion occupée par chaque catégorie (environ 27% de mots-outils et 73% de mots de contenu – $t(19) = 1,699$, $p = 0,106$).

A priori, ces données seraient assez différentes dans le cas d'un corpus non contrôlé : complexité des phrases réduite au minimum dans le cas bruité et possiblement plus élevée dans la condition non bruitée.

4.4.2.4 Directions du regard

La direction du regard a été annotée sur toute la longueur des vidéos enregistrées. Les directions estimées du regard sont classifiées dans 4 catégories distinctes : vers le modèle, vers l'interlocuteur, vers le plateau, vers une direction indéterminée.

La Figure 4.8 représente les directions annotées pour le regard, chaque barre de l'histogramme représentant la proportion de temps passée dans une position de regard donnée. Il est assez clair que le plateau (la zone de jeu) est la partie la plus regardée ($> 40\%$ du temps), en condition non bruitée, les participants ne regardent le complice que $5,97\%$ du temps ce qui est très court (parmi tous les participants, aucun ne regarde l'interlocuteur plus de 15% du temps en condition non bruitée) alors que cette proportion augmente drastiquement en condition *Bruit* puisque les participants fixent l'interlocuteur $27,35\%$ du temps (cet effet de la condition sur la proportion de temps passée à regarder l'interlocuteur est significatif pour chaque participant). Ce résultat confirme le fait que la tâche est perçue comme un jeu collaboratif qui donne lieu à une interaction naturelle.

On observe donc un effet de la condition *Bruit vs Sans Bruit* sur la proportion de temps passé à regarder le plateau ($t(19) = 2,2191$, $p < 0,05$), sur la proportion de temps passé à regarder le modèle ($t(19) = 6,8659$, $p < 0,001$) et sur la proportion de temps passé à regarder le complice ($t(19) = -11,13$, $p < 0,001$).

La direction du regard en fonction de l'instant de production des gestes varie également légèrement parmi les niveaux de la condition BRUIT, sur la Figure 4.9, on observe un comportement similaire du regard pour les instants annotés du geste que sur la durée totale d'une grille : la quantité de regards orientés vers l'interlocuteur augmente sensiblement d'une condition à l'autre. Cependant, il n'y a aucun instant du geste pour lequel un regard systématique dans une direction est remarqué : *a priori* l'augmentation du nombre de regards vers l'interlocuteur et la durée totale de celle-ci est une tendance globale, distribuée sur la durée totale de la tâche et en particulier sur les instants de production des gestes.

4.4.3 Comportements observés pour la parole

La condition BRUIT permet de comparer les productions dans un environnement calme et dans un environnement bruité, il est alors intéressant de voir si les comportements observés suivent les comportements décrits

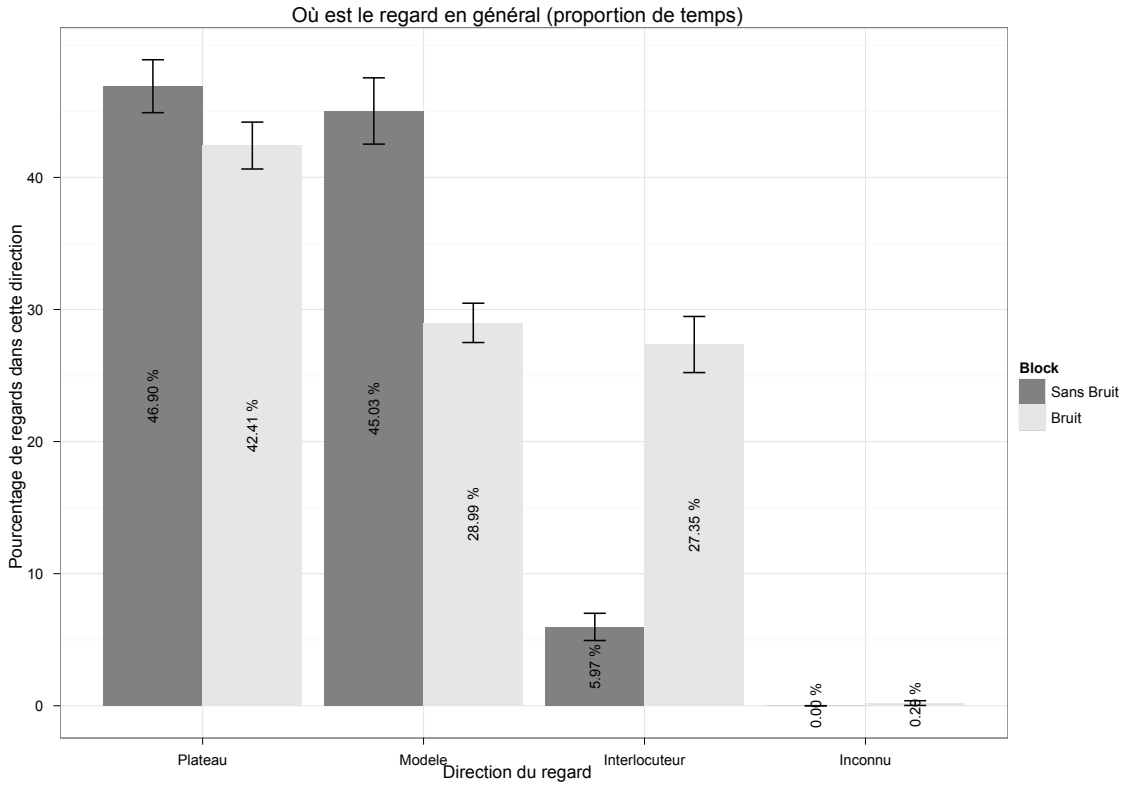


FIGURE 4.8 – Direction du regard sur le temps total de remplissage des grilles

dans la littérature sur l'effet Lombard. Globalement, d'après la littérature, l'effet Lombard peut se caractériser à la fois acoustiquement et articulatoirement. Au niveau acoustique, on trouve en particulier, une augmentation de l'intensité moyenne (toutes les études rapportent cet effet, voir [175] pour une description des spectres en environnement bruité, [57] donne des données précises pour un participant : les augmentations dépendent du type de bruit, de son intensité, du sexe et du phonème produit), une augmentation de F_0/F_1 moyen (ceci est un autre résultat faisant l'unanimité), une augmentation de la durée des productions –*i.e.* un ralentissement du débit de parole moyen– voir [88] par exemple. En articulatoire, peu d'études ont été menée mais les travaux de Kim & Davis présentés en 2005 et 2006 [36] ont montré par des mesures de mouvements faciaux (suivi tridimensionnel par Optotrak) un accroissement de l'amplitude des mouvements articulatoires (en particulier : ouverture aux lèvres, protrusion des lèvres), de même l'article "classique" de Schulman en 1989 [163] montrait une augmentation de l'ouverture de la mâchoire (fortement corrélée à l'ouverture aux lèvres) en voix forte par rapport à une voix non forcée. Enfin, les durées des différentes syllabes ne semblent pas varier avec l'effort vocal supplémentaire fourni (cet effet est trouvé dans certaines études mais ce n'est pas la majorité). Il est à noter que l'étude de la bibliographie sur l'effet Lombard (voir [56] par exemple pour une revue en détail) met régulièrement en avant une grande variabilité dans l'adaptation à une situation de communication dégradée. Ce point est à garder à l'esprit pour toute étude portant sur la parole dans le bruit.

4.4.3.1 Indices acoustiques

Tout d'abord il est évident à l'écoute des signaux, et ceci est conforté par un test statistique, que l'intensité moyenne ainsi que le pic d'intensité sont plus élevés dans la condition bruitée ($77,34$ dB SPL en condition bruitée en moyenne contre $57,56$ dB SPL en moyenne en condition non bruitée –$t(19) = -11,72$, $p < 0,001$ pour l'intensité maximale sur les énoncés). L'augmentation entre les deux conditions n'est pas significativement plus élevée dans une partie de l'énoncé que dans une autre : les augmentations d'intensité observées sur les deux syllabes du mot-cible et sur le démonstratif sont similaires.

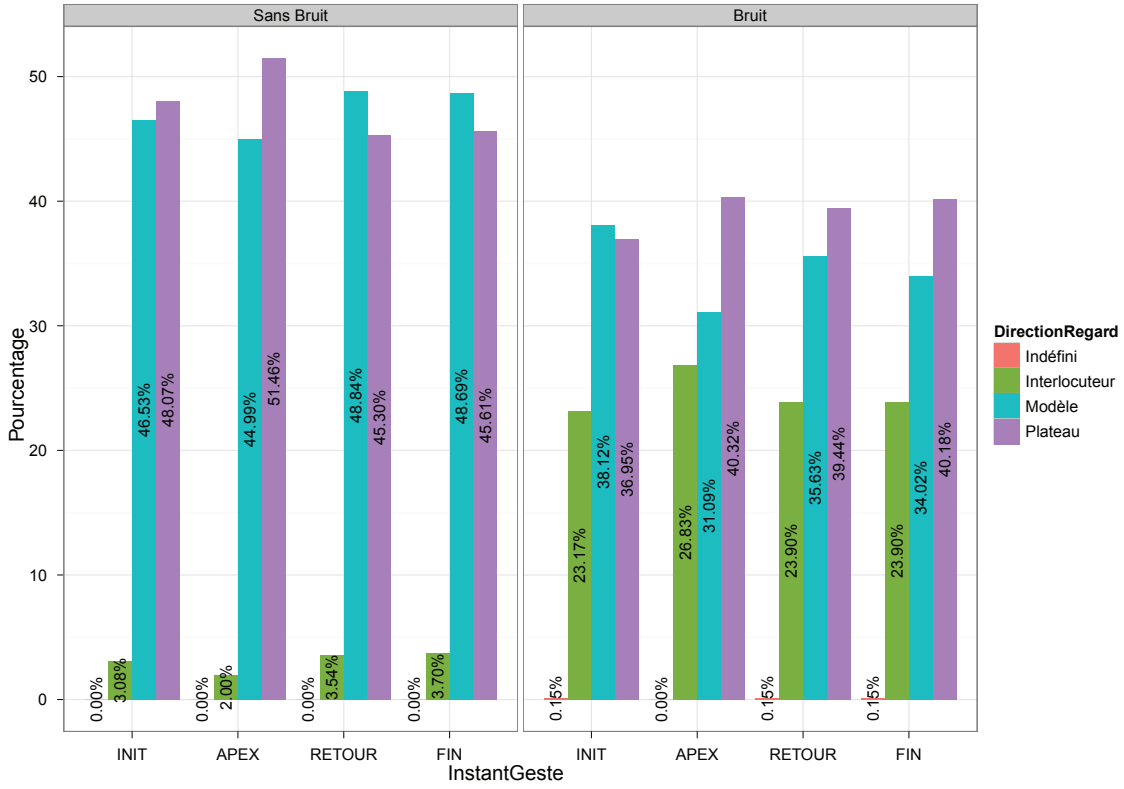


FIGURE 4.9 – Direction du regard aux instants annotés du geste

De même, il est observé que le pic de fréquence fondamentale sur l'énoncé augmente beaucoup d'une condition à l'autre : il passe de $204,17\,\text{Hz}$ en moyenne dans la condition *Sans Bruit* à $303,84\,\text{Hz}$ en condition *Bruit* ($t(19) = -12,11$, $p < 0,001$). Comme présenté dans la littérature, l'influence de la diffusion de bruit n'est pas différente sur les différentes parties de la parole, les pics de fréquence fondamentale ont été mesurés sur les deux syllabes du mot-cible ainsi que sur le démonstratif ; une ANOVA à mesures répétées montre un effet de la condition BRUIT sur l'amplitude du pic de fréquence fondamentale ($F(1,19) = 184$, $p < 0,001$), un effet de l'instant de mesure du pic de F_0 ($F(2,19) = 5,98$, $p < 0,01$ mais pas d'interaction ($F(2,19) = 2,95$, $p = 0,058$) : le pic de fréquence fondamentale est différent pour les trois instants mesurés, il augmente avec la diffusion de bruit mais cette augmentation est similaire en tous points.

4.4.3.2 Débit de parole

Comme montré ci-dessus, le débit de parole est probablement moins élevé en condition *Bruit* puisque les énoncés du corpus ont une durée plus longue. Les mesures de débit donnent raison au raisonnement logique précédent puisque le débit de la parole en condition *Bruit* est de $3,46$ mots/seconde alors qu'il est de $4,52$ mots/seconde dans la condition *Sans Bruit* (les mots sont plutôt courts dans les phrases du corpus).

Dans cette expérience où le corpus est contrôlé et la tâche assez simple, il est assez prévisible que le nombre de mots utilisés ne varie pas entre les conditions avec et sans bruit, et qu'alors la modification du débit de parole soit inversement proportionnelle au temps consacré en moyenne à chaque énoncé. Les mesures de débit montraient l'influence du facteur BRUIT sur le débit de la parole : celui-ci est plus faible en condition *Bruit* que dans l'autre condition. De même, il a été montré en Section 4.4.2 que la parole était ralentie en condition *Bruit*, ce qui correspond à la littérature.

Un dernier point important est le fait que, même si le débit est modifié, cette modification est répartie sur tout l'énoncé et n'est pas localisée sur un mot spécifique. Ainsi, chaque syllabe prononcée occupe la même proportion du temps total de parole dans la condition *Sans Bruit* que dans la condition *Bruit* : l'organisation temporelle de

l'énoncé reste identique entre les deux conditions. Globalement, l'article prend $15,6\%$ du temps, la première syllabe du mot cible $22,3\%$, la seconde syllabe $22,4\%$, le verbe $15,5\%$, le démonstratif $24,2\%$.

4.4.3.3 Indices articulatoires

Les cibles articulatoires annotées l'ont été sur les deux syllabes du mot-cible ainsi que sur le démonstratif et sont de deux types "ouverture" et "protrusion". Il a été proposé une normalisation des amplitudes afin de pouvoir les comparer entre elles précédemment (p.47), mais il est intéressant ici de voir si la condition BRUIT a une influence sur l'amplitude de l'articulation de ces deux types de cibles articulatoires et si oui, si cette influence est différente selon le type de cible. Ici aucune normalisation n'est réalisée et une ANOVA est réalisée pour chaque type de cible articulatoire (ouverture ou protrusion).

Deux ANOVAs à deux facteurs (emplacement de la cible articulatoire $\times$ niveau de la condition BRUIT) ont été donc menées (une pour les données d'ouverture, l'autre pour les données de protrusion). La variable dépendante utilisée est l'amplitude des cibles articulatoires. L'ANOVA sur les données d'ouverture donne des effets principaux pour l'emplacement de la cible articulatoire ($F(2,90) = 5,91$, $p < 0,01$) et du niveau de la condition BRUIT ($F(1,90) = 196,32$, $p < 0,001$) mais aucune interaction significative des deux facteurs ($F(2,90) = 2,41$, $p = 0,0955$) : l'ouverture est plus grande en condition *Bruit* que dans la condition non bruitée ($31,19$ mm *vs* $24,28$ mm) et l'ouverture est plus grande pour la première cible articulatoire ($28,39$ mm) et la cible articulatoire du démonstratif ($28,01$ mm) que pour la cible de la seconde syllabe du mot-cible ($26,48$ mm) en moyenne. L'ANOVA sur les données de protrusion ne donne, au contraire, aucun effet principal (ni de l'emplacement de la cible articulatoire, ni du niveau de la condition BRUIT) ni aucun effet d'interaction.

Finalement, on observe bien une hyper-articulation des voyelles ouvertes en condition bruitée, ce qui n'est pas le cas pour les voyelles protruses (ce résultat est possiblement dû au fait que l'hyper-articulation des voyelles protruse se traduit en pratique par une plus grande ouverture plutôt qu'une plus grande protrusion) *i.e.* il est possible que tous les gestes soient plus ouverts systématiquement dans le bruit (sans forcément exagérer le geste "saillant", par exemple, la protrusion). Les résultats sur l'ouverture sont concordants avec la littérature (par exemple, en vidéo, Garnier *et col.* [57] trouvent effectivement une hyper-articulation qui se traduit par une augmentation de l'ouverture dans une situation de *cocktail-party*).

4.4.4 Comportements observés pour les gestes

4.4.4.1 Quelques éléments préliminaires

Beaucoup de gestes de pointage sont produits lors de la tâche, quelle que soit la condition (*Bruit* ou *Sans Bruit*), ce qui était attendu puisque l'expérience est *construite* pour cela. Bien que tous les participants soient droitiers, il a été remarqué lors de l'annotation des vidéos que différentes stratégies sont utilisées par les participants pour donner les instructions au complice. En effet, les instructions ne sont pas toujours données en utilisant des pointages avec la main droite (certains participants n'utilisent presque pas leur main droite, d'autres utilisent les deux mains indifféremment, ...). La Figure 4.10 montre la répartition des gestes produits par participant. Ces histogrammes contiennent des barres *Pas de geste* : chaque énoncé produit sans geste (typiquement, des commentaires, parfois, des corrections, etc...) est comptabilisé dans cette catégorie.

Il est assez net sur ces graphes que les participants ne changent pas radicalement de façon de procéder entre les deux conditions de BRUIT : la répartition de l'utilisation de la main droite *vs* main gauche est analogue dans les deux conditions pour tous les participants.

De façon similaire, les gestes (de pointage, pour la plupart) ont été classifiés dans plusieurs catégorie afin d'avoir une idée grossière de leur complexité ainsi que de l'enchaînement des gestes. La Figure 4.11 donne la répartition des pointages (et autres gestes) par type annoté. Les types annotés se séparent en deux principales catégories : les gestes enchaînés (sans position de repos intermédiaire) et les gestes séparés par une pause dans le mouvement, les gestes qui ne sont pas des pointages sont classés dans la catégorie "Autre" et enfin, certains énoncés ne sont associés à aucun geste.

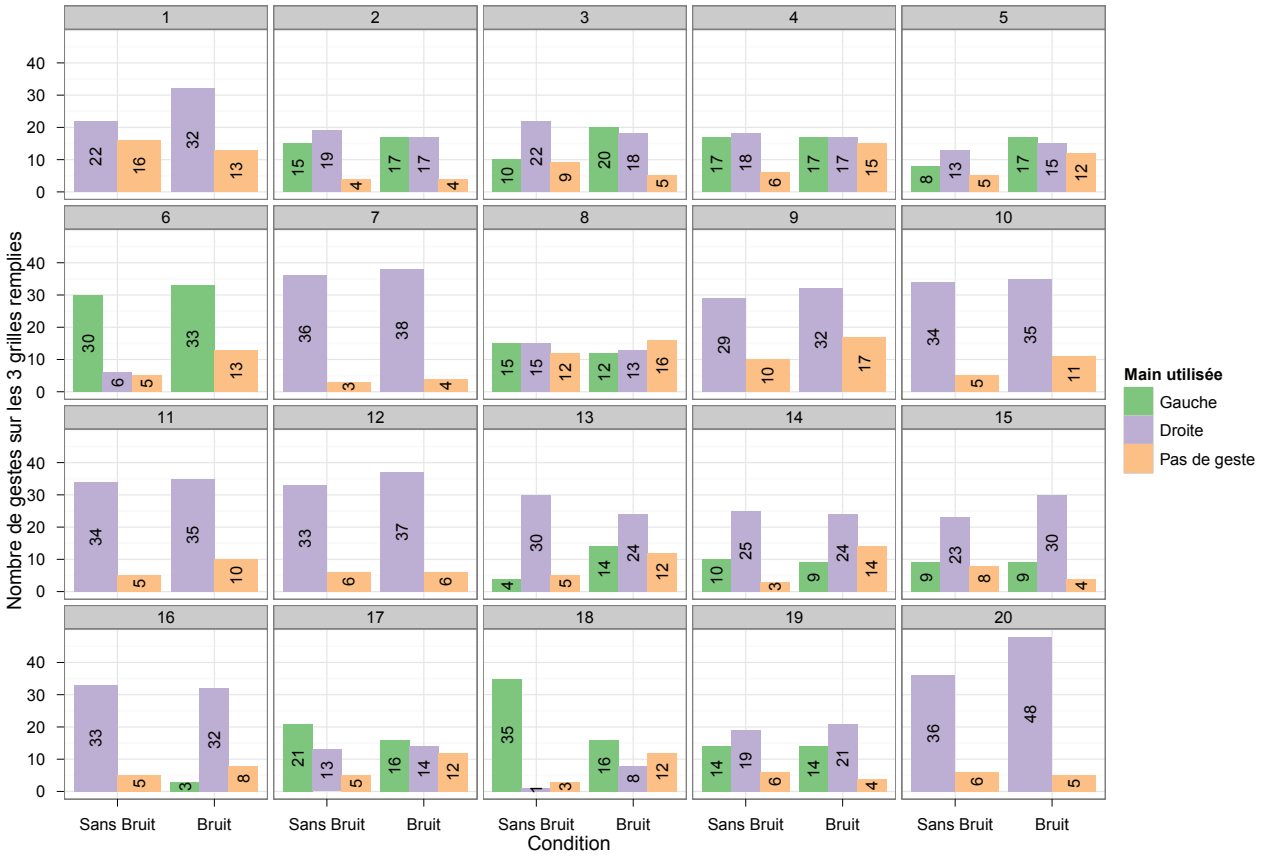


FIGURE 4.10 – Répartition de main utilisée dans la production des gestes manuels par sujet (en nombre de gestes annotés)

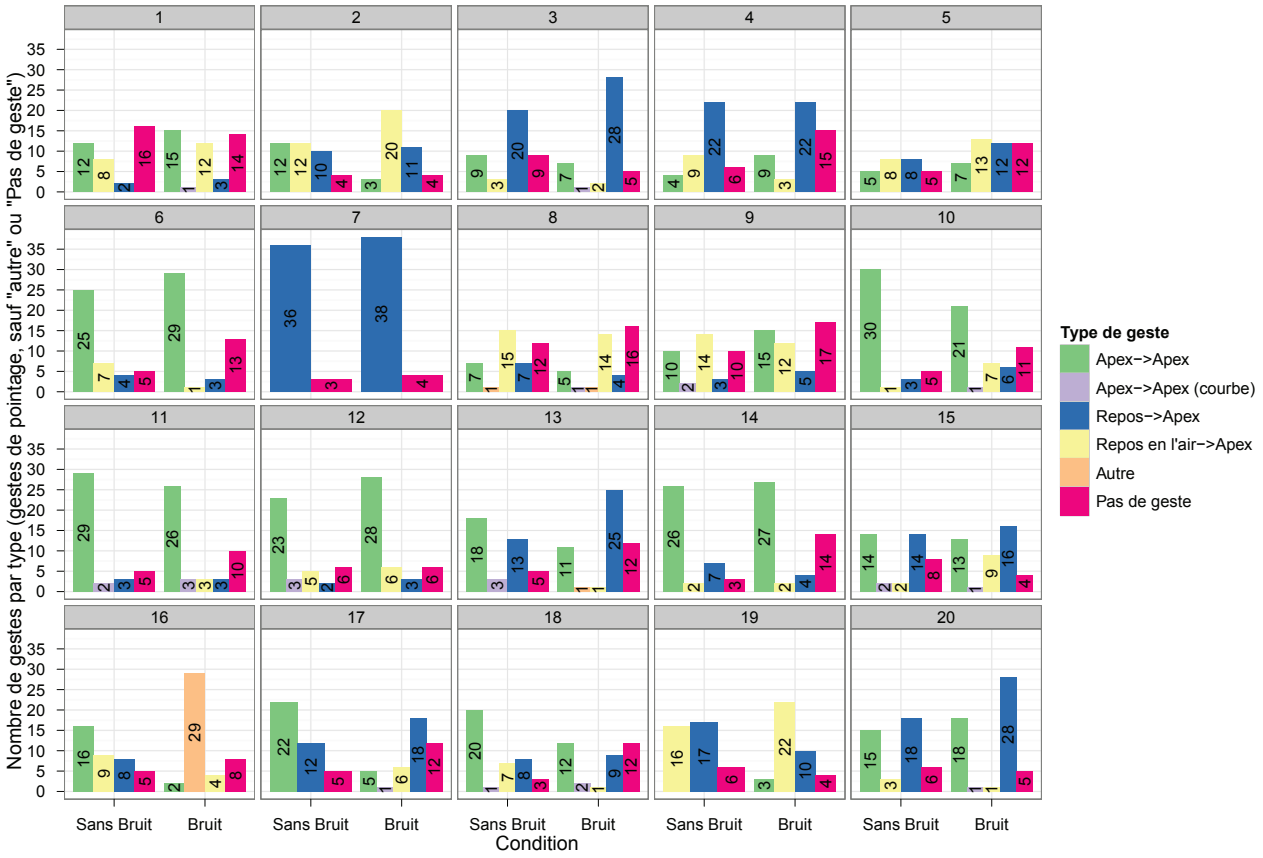
Sur cette figure, on voit assez clairement que, pour une grande majorité des participants, la diffusion de bruit ambiant ne modifie pas réellement la façon dont sont produits leurs gestes de pointage. Seuls les participants n° 16 et 17 semblent avoir des comportements radicalement différents en condition *Bruit*. Le participant n° 16 produit des gestes annotés dans la catégorie "Autre" en condition *Bruit* : dans la condition bruitée, ce participant produit en fait des gestes iconiques pour illustrer son propos, sa stratégie de communication change donc assez radicalement d'une condition à l'autre. Le participant n° 17 ponctue ses gestes avec des pauses entre chaque geste en condition bruitée, ce qu'il ne faisait pas en condition *Sans Bruit*. Dans ces deux cas, les modifications apportées permettent probablement aux participants de "clarifier" le message qu'ils transmettent au complice.

4.4.4.2 Instants de production des gestes

Dans tout ce qui suit, seules les productions d'énoncés valides, ne contenant aucune interruption et auxquelles sont associées un geste sont considérées dans les analyses, en effet, ces productions sont les seules qui sont comparables statistiquement. Le participant n° 9 a produit des pauses lors de tous ses énoncés en condition *Bruit*, ce participant est exclu des analyses dans la plupart des cas puisqu'il est impossible de comparer ses productions dans les deux niveaux de la condition BRUIT. Par ailleurs, tous les instants sont normalisés sur la durée de l'énoncé considéré *i.e.* le temps 0 est le début acoustique de la parole et le temps 1 la fin acoustique de la parole.

Les instants de production des gestes de pointage sont représentés en Figure 4.12, le tableau de la figure permet d'apprécier de manière plus précise les instants de production en relation avec les instants de production de la parole.

Une première chose visible grossièrement sur la figure et confirmée par les tests statistiques est le fait que les instants de production des gestes ne varient pas d'une condition à l'autre, sauf pour l'instant de début de geste



Les catégories représentent les trajectoires des gestes :	
Apex→Apex	Pointages enchaînés
Apex→Apex (courbe)	Pointages enchaînés (trajectoire courbée)
Repos→Apex	Repos sur la table puis pointage
Repos en l'air→Apex	Main immobile en l'air puis pointage
Autre	Tous les autres gestes

FIGURE 4.11 – Répartition des gestes par sujet selon leur type (en nombre de gestes annotés)

qui est produit légèrement plus tard par rapport à la parole en condition bruitée ($t(18) = -2,434$, $p < 0,05$, différence entre les deux conditions : $0,131$). En moyenne, les instants moyens de production d'une condition à l'autre varient très faiblement : le pic de vitesse est produit $0,066$ plus tard en condition bruitée ($t(18) = -1,55$, $p = 0,140$, n.s.), l'apex $0,005$ plus tôt en condition bruitée ($t(18) = -0,126$, $p = 0,900$, n.s.), le retour $0,059$ plus tôt dans la condition bruitée ($t(18) = -0,556$, $p = 0,590$, n.s.) et la fin du geste $0,064$ plus tard en condition bruitée ($t(18) = 0,615$, $p = 0,550$, n.s.). La condition BRUIT a donc une influence minime sur les instants de production moyens des gestes de pointage au sein de l'énoncé.

La figure permet également de voir que les instants de production des gestes ont des variabilités légèrement différentes dans les deux conditions. Ceci est confirmé par les tests statistiques qui montrent une variabilité des instants de production normalisés plus faible pour les instants $\mathrm{P_{On}}$ et $\mathrm{P_A}$ dans la condition bruitée ($\mathrm{P_{On}}$: $t(17) = 2,44$, $p < 0,05$, écart-type plus faible de $0,080$ en condition bruitée ; $\mathrm{P_A}$: $t(17) = 3,182$, $p < 0,01$, écart-type plus faible de $0,118$ en condition *Bruit*). Il faut cependant être prudent pour les résultats concernant la variabilité des données normalisées en condition *Bruit vs Sans bruit* car il a été montré que les durées brutes des productions vocales étaient plus longues en condition bruitée : ainsi si l'écart-type des données brutes est constant entre les deux conditions, celui-ci est "artificiellement" réduit lors de la normalisation par rapport à la durée totale de parole. Sur les données non-normalisées, seul l'instant d'initiation du geste a des écarts-types différents dans les deux conditions ($t(17) = 2,22$, $p < 0,05$, avec un écart-type plus faible en condition *Bruit* de

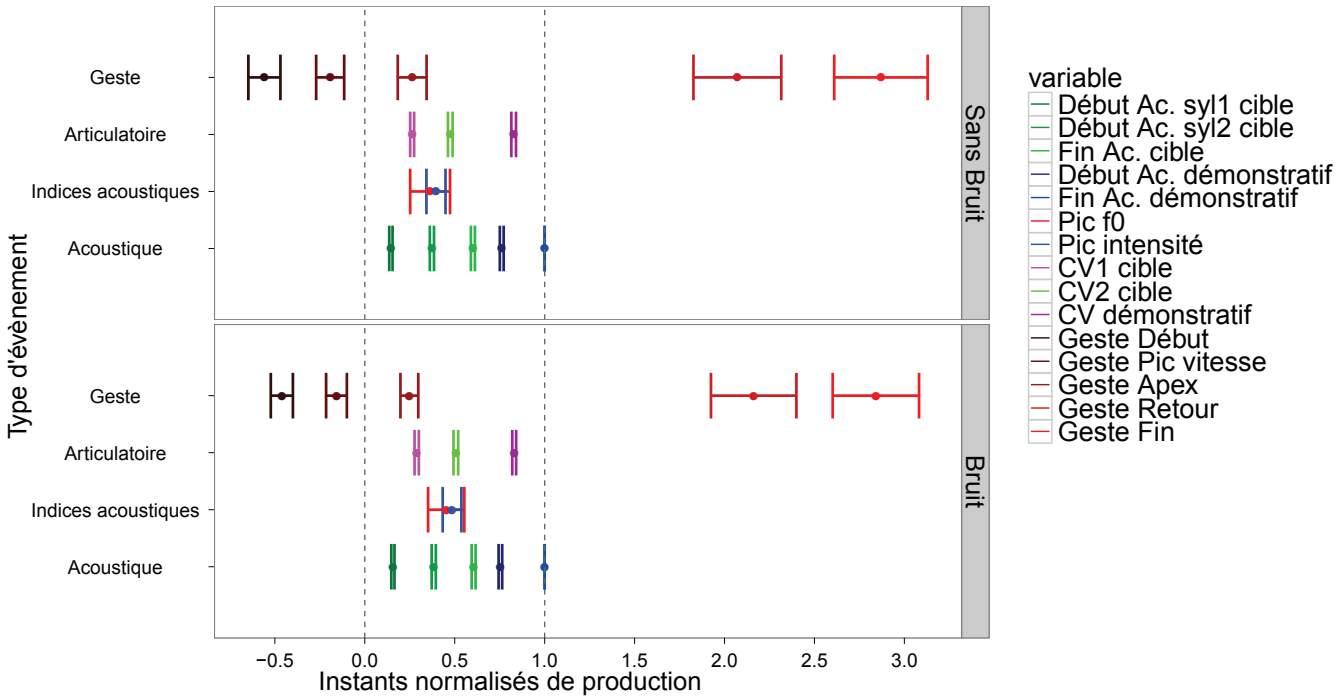


FIGURE 4.12 – Instants de production des gestes en relation avec la parole

42 ms). La variabilité des autres instants ne semble pas être influencée par le niveau de la variable BRUIT (tous les tests t sont non-significatifs et les différences des moyennes sont faibles).

Ainsi, en passant d'une condition à l'autre *i.e.* en perturbant la communication, il semble que l'onset du geste commence plus tard et que les instants d'onset (et d'apex, relativement à la durée de la parole) du geste soient moins variables. Les autres instants annotés sur les gestes ne semblent pas être influencés par la présence ou non de bruit. La réduction de la variabilité (assez importante) des deux instants d'onset et d'apex permet, une fois de plus de penser que ces instants jouent un rôle important en relation avec la parole.

Pour ce qui est de la durée des différentes phases du geste de pointage, les résultats dépendent de la normalisation adoptée. Les durées des gestes (en secondes) sont clairement plus longues en condition bruitée qu'en condition non bruitée ($t(18) = -3,70$, $p < 0,001$, $+451$ ms) et cette différence s'explique par une tenue du geste plus longue en condition bruitée ($t(18) = -3,747$, $p < 0,001$, $+440$ ms). Cependant ces résultats sont totalement différents si on utilise la normalisation sur la durée de l'énoncé : dans ce cas, ni la durée globale du geste ni la durée de la tenue du geste ne sont modifiées en passant d'une condition à l'autre ; néanmoins, on observe une modification des durées du *stroke* d'onset ($t(18) = 4,178$, $p < 0,001$, $+0,129$ en condition sans bruit) et du *stroke* d'offset ($t(18) = 4,606$, $p < 0,001$, $+0,121$ en condition sans bruit). Ceci tend à montrer que la durée de la tenue du geste de pointage est corrélée à la durée de l'énoncé.

Globalement, ce n'est donc pas à un ralentissement généralisé et parallèle des productions de gestes manuels et de parole auquel on assiste, mais à une modification gestuelle qui s'adapte sur les zones importantes. Le geste dure plus longtemps dans la condition bruitée et cet allongement se répercute principalement sur la tenue du geste dont la durée augmente avec la durée de l'énoncé. Ceci est confirmé par une étude de corrélation de ces durées $r = 0,34$, $p < 0,001$. Dans la condition bruitée, le geste commence plus tôt, la durée du *stroke* d'onset est similaire à la condition non bruitée, la tenue rallongée d'un temps similaire à celui de l'énoncé et le *stroke* d'offset a une durée similaire à la condition sans bruit : si le rallongement du temps de parole est dû à un ralentissement global de l'articulation, on ne peut pas en dire autant pour le geste manuel. En effet, le geste commence plus tôt dans le bruit et certaines parties du geste ont des durées similaires à ce qu'on observe sans bruit. Par contre, la seule durée modifiée (la tenue de l'apex) est rallongée d'une durée similaire à celle de la durée supplémentaire de parole. Ceci va donc plutôt dans le sens d'une exécution gestuelle qui s'adapte à l'exécution de la parole et non pas dans le sens d'un ralentissement général de l'exécution du geste : les sujets montrent (tiennent leur geste de

pointage) plus longtemps dans le bruit et cela est visiblement dû à une parole ralentie.

4.4.4.3 Amplitudes manuelles

L'amplitude des gestes manuels n'est pas une variable d'intérêt dans cette étude puisque l'amplitude des gestes est définie par les emplacements pointés par le participant et l'absence de position de repos rend cette variable inadéquate pour une étude statistique : non seulement l'amplitude est contrainte par les positions marquées sur le plateau de jeu, mais il est possible que, selon l'essai considéré, le geste ne "provienne pas" du même emplacement (par exemple : lors d'un essai, le pointage vers la position p est l'enchaînement d'un pointage vers une position $p\prime$; lors d'un autre essai, le pointage vers la position p est placé après un instant de repos ; etc. . .).

Les trajets entre deux positions quelconques sont dans tous les cas des trajets courts et le temps mis à disposition pour réaliser ce trajet n'est pas contraint par quoi que ce soit d'autre que le bon vouloir du participant. Ainsi, pour réaliser n'importe quel mouvement, le participant peut influer sur la trajectoire qu'il choisit (le plus souvent : une trajectoire rectiligne) et sur les paramètres dynamiques qu'il utilise (vitesse, accélération, . . .). Le type de trajectoire utilisée est annoté grâce à la vidéo filmée lors de la passation de l'expérience, cette annotation permet également de dresser quelques catégories de gestes de pointage (gestes enchaînés ou non, trajectoire rectiligne ou non, . . .).

Le pic de vitesse du stroke du geste de pointage est une partie saillante du geste , or pour chaque geste annoté, l'annotation s'est faite à partir des pics de vitesse des gestes, cette variable est donc comparable parmi les gestes. Il est intéressant de voir si la "saillance" des gestes est modifiée dans le bruit, comme cela semble être le cas pour la parole (effet Lombard).

Une analyse de la variance globale ne donne aucun effet principal ni de la condition BRUIT ni de la catégorie de geste de pointage sur l'amplitude du pic de vitesse des gestes annotés. Le bruit n'influence pas l'amplitude du pic de vitesse des gestes.

4.4.5 Instants de production des gestes relatifs à la parole

Comme remarqué ci-dessus (et dans les expériences présentées ci-avant), les gestes sont le plus souvent produits à des instants recouvrant les instants de production de la parole. En particulier, la Figure 4.12 montre que les instants d'onset/pic de vitesse sont généralement produits avant le début de l'énoncé (temps normalisés inférieurs à 0) et que les instants de retour/offset sont généralement produits après la fin de l'énoncé (temps normalisés supérieurs à 1), l'apex étant majoritairement produit au sein de la production vocale.

La Table 4.5 donne un aperçu global de l'ensemble des productions de tous les participants (510 essais dans la condition *Sans Bruit* et 419 dans la condition *Bruit*). Dans cette table "l'objet" est le mot représenté sur une des cartes à disposition du complice *i.e.* un mot du corpus.

Bloc	P_{On} avant la parole	PA dans la parole	P_A dans l'objet	P_A dans le démonstratif	P_R après la parole	Tenue du geste et parole recouvrement non nul
Sans Bruit	96,47%	85,11%	61,01%	2,26%	76,90%	99,34%
Bruit	97,13%	92,91%	58,83%	3,19%	82,90%	100,00%

TABLE 4.5 – Emplacement des productions gestuelles par rapport à la parole

Ce tableau donne quelques résultats intéressants sur les instants de production des gestes en général et plus précisément de l'apex de ceux-ci. Tout d'abord, il est assez évident que dans une large majorité des productions, l'onset du geste se situe avant l'onset de la parole et ce, quel que soit le niveau de la condition BRUIT, le nombre de productions pour lesquelles l'onset du geste commence après la parole est réellement minime. En ce qui concerne l'apex du geste de pointage, celui-ci intervient soit avant l'énoncé (minorité des cas) soit pendant la production de l'énoncé (dans la plupart des cas) mais jamais après l'énoncé.

De manière plus précise, l'apex du geste est produit dans la majorité des cas de manière concurrente à la production du mot correspondant à la cible à placer sur le plateau de jeu, et n'est que marginalement produit de

manière concurrente avec le démonstratif (en fait, la production de l'apex en parallèle du démonstratif n'intervient que pour cinq participants), les productions restantes étant le plus souvent réalisées avant le mot-cible (*i.e.* en parallèle de l'article du début d'énoncé). La cooccurrence de l'apex avec le mot-cible se situe bien au delà du niveau de la chance (pour un tirage uniforme dans l'énoncé, le niveau de chance pour l'apex d'être cooccurrent avec le mot-cible est de 45% environ) : les participants ont tendance à regrouper les information complémentaires. Enfin, le retour du geste intervient dans une très large majorité des cas après la parole, seuls quatre participants produisent souvent le retour des gestes en même temps que la parole.

Il est intéressant de remarquer que la production des gestes, bien que totalement libre, englobe le plus souvent la production de parole : le geste débute le plus souvent avant la parole et se termine après la fin de celle-ci. De manière plus précise, dans tous les cas, le laps de temps pendant lequel a lieu la tenue du geste de pointage a une intersection non vide avec l'intervalle de temps utilisé pour la production de parole.

4.4.5.1 Alignements possibles des instants annotés de production des gestes manuels

L'étude des alignements se fait sur les données non normalisées. En effet, il a été montré précédemment que la durée moyenne des énoncés était plus grande en condition *Bruit* que dans la condition *Sans bruit*, l'utilisation de données normalisées entraînerait donc une réduction artificielle des alignements en condition bruitée et rendrait les deux conditions non comparables.

Apex des gestes Comme vu ci-dessus, l'apex des gestes intervient majoritairement et avec une fréquence supérieure au hasard de manière cooccurrente avec le mot-cible produit. Comme étudié dans les expériences précédentes, il est intéressant ici de voir si cette cooccurrence se fait globalement au niveau du mot-cible ou bien si des corrélats plus fins de coordination peuvent être pris en compte.

La Figure 4.13 représente les densités de probabilité estimées des instants de production de l'apex des gestes de pointage ainsi que des limites acoustiques et indices articulatoires annotés sur les productions. D'après ces graphes, les productions des apex des gestes de pointage interviennent le plus souvent en début de mot-cible (respectivement $36,98\%$ et $46,17\%$ des essais en conditions *Sans Bruit* et *Bruit* ont des instants d'apex du geste de pointage contenus entre les limites acoustiques de la première syllabe du mot-cible) et plus marginalement en fin du mot-cible (respectivement $24,64\%$ et $12,4\%$) et avec l'article (respectivement $12,79\%$ et $25,97\%$). Ceci est différent du hasard qui donnerait 15% de chances de "tomber" sur l'article, $22,5\%$ de chances de sur une des deux syllabes de l'objet et $24,2\%$ de chances de cooccurrence avec le démonstratif.

Cependant, la Figure 4.13 ne montre *a priori* pas de cooccurrence systématique entre la production de l'apex du geste de pointage et les limites acoustiques/les gestes articulatoires annotés. Des tests t sur les différences temporelles entres évènements manuels et indices prosodiques/articulatoires/acoustiques donnent cependant un alignement entre l'instant de production de l'apex et les instants annotés au sein de la première syllabe et ce, dans les deux niveaux du facteur BRUIT : statistiquement l'instant d'occurrence de $\mathrm{P_A}$ n'est pas différent de l'instant d'occurrence de la première cible articulatoire du mot-cible ($t(19) = 0,0179$, $p = 0,980$, différence moyenne : $-0,00075$), ni des pics d'intensité ($t(19) = 0,662$, $p = 0,510$, différence moyenne : $0,026$) et de fréquence fondamentale ($t(19) = 0,387$, $p = 0,700$, différence moyenne : $0,016$) de la première syllabe.

Les alignements observés sont valables pour les deux niveaux du facteur BRUIT, il est cependant intéressant de voir si la perturbation de la situation d'interaction a une influence sur ces alignements. Une analyse statistique ne donne aucun effet de la condition BRUIT sur les alignements possibles observés ci-dessus (CV1 : $t(18) = -1,34$, $p = 0,195$, pic d'intensité de la première syllabe : $t(18) = -0,70$, $p = 0,493$, pic de fréquence fondamentale de la première syllabe : $t(18) = -1,489$, $p = 0,154$). Par ailleurs, le facteur BRUIT n'a aucune influence sur l'écart-type des alignements évoqués ci-dessus (CV1 : $t(17) = -0,143$, $p = 0,888$, pic d'intensité de la première syllabe : $t(17) = 0,145$, $p = 0,887$, pic de fréquence fondamentale de la première syllabe : $t(17) = -0,215$, $p = 0,832$). Ainsi, bien que les productions d'énoncés soient plus lentes en condition *Bruit*, la coordination geste/parole entre l'apex du geste et les instants de la parole reste totalement inchangée : ni la valeur moyenne ni l'écart-type des alignements n'est influencée par la perturbation de la situation d'interaction.

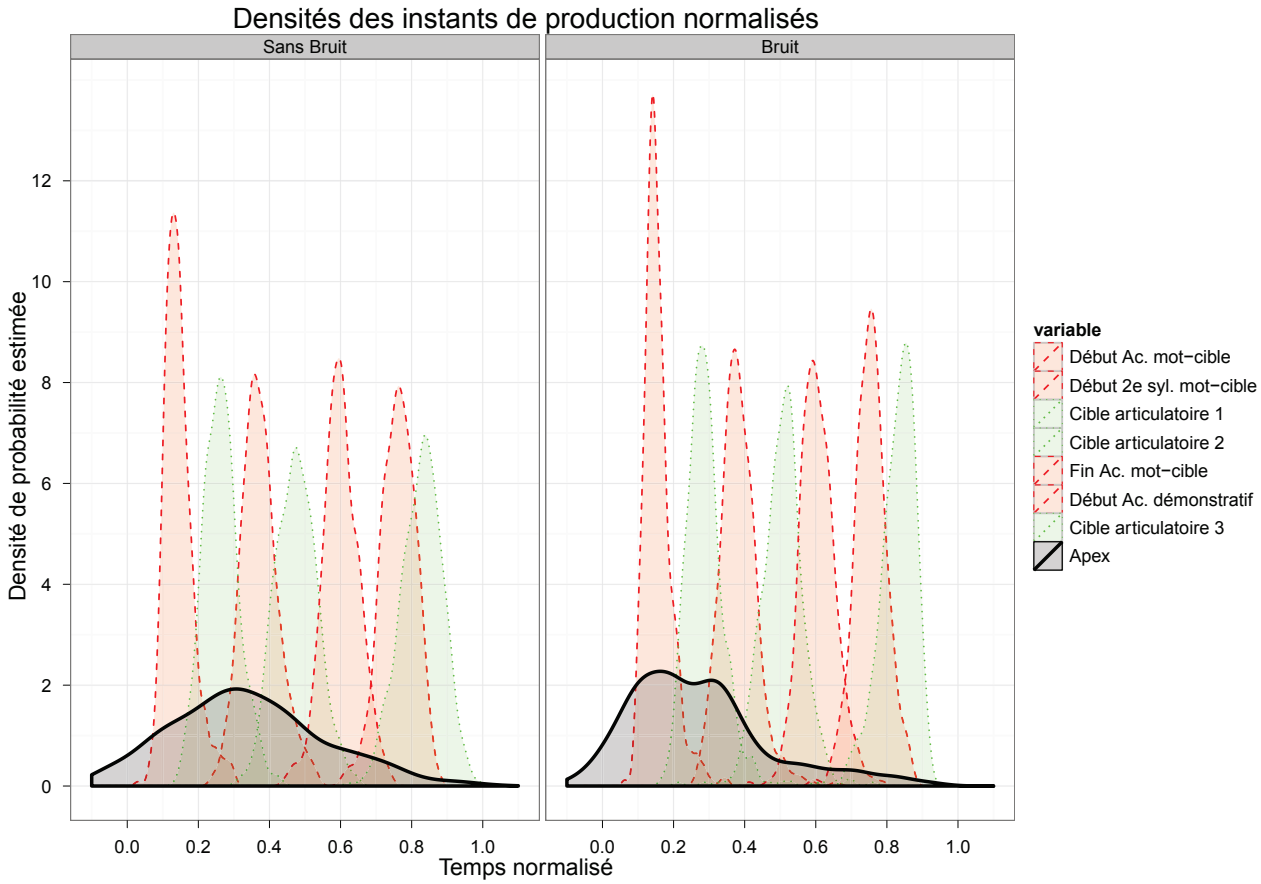


FIGURE 4.13 – Densités de production de l'apex des gestes et des limites acoustiques, cibles articulatoires

Pour donner une vue générale, les différents alignements calculés pour les instants annotés de la parole sont représentés en Figure 4.14 dans les deux niveaux du facteur BRUIT.

Cette figure ainsi que les tests ci-dessus confirment donc un alignement entre l'instant de production de l'apex et les instants de la première syllabe du mot-cible. De façon intéressante, il a donc été montré un alignement entre l'apex du geste de pointage et les pics de fréquence fondamentale et une cible articulatoire du mot correspondant à l'image à placer *i.e.* on observe une coordination entre le mouvement de la main, les mouvements articulatoires et le pic de F_0. Ceci est à rapprocher des travaux de D'Imperio *et col.* [42] qui montraient lors d'enregistrements Optotrak une coordination entre des mouvements supra-laryngés (des lèvres) et le pic de F_0.

Hormis l'apex du geste de pointage, le retour de celui-ci est aussi un instant important puisqu'il signale la fin de la désignation.

Retour des gestes de pointage Comme présenté en Section 4.4.5, le retour des gestes de pointage se fait généralement après la fin de la prononciation de l'énoncé du corpus. Bien qu'il soit tout à fait improbable qu'un alignement soit trouvé avec les instants de la parole, il est possible que ce retour du geste tardif soit lié au fait que le participant attend que son interlocuteur lui "prouve" qu'il a bien compris la consigne qui lui a été communiquée (*i.e.* poser telle carte à tel endroit). La seule façon pour le complice de prouver qu'il a bien compris est de prendre la carte et de la poser au bon endroit. Une fois cette action effectuée (ou quasiment effectuée), le participant est certain que l'information a bien été transmise et que la tâche est effectuée correctement. Cependant, les statistiques ne donnent raison à cette façon de procéder qu'en condition bruitée : lorsqu'il n'y a pas de bruit les deux instants ne sont pas alignés comme représenté sur la Figure 4.15. Ceci est dû au fait que le retour intervient plus tôt par rapport à la pose de carte en condition non bruitée. Cependant, dans les deux niveaux du facteur BRUIT, les instants de retour du geste de pointage et de pose de carte sont statistiquement positivement corrélés

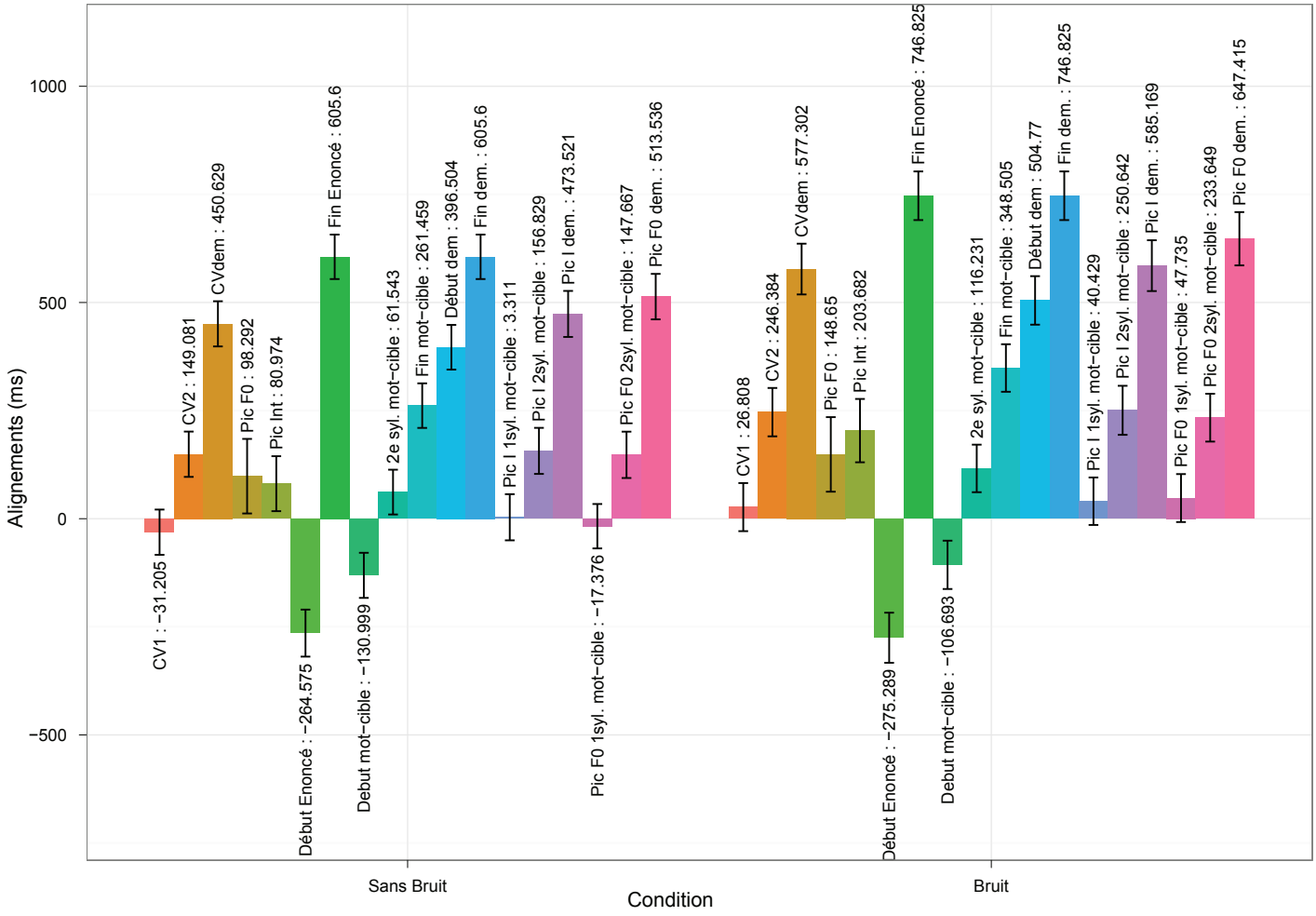


FIGURE 4.14 – Différences temporelles entre l'apex et les différents instants annotés

(bien que la corrélation soit plus grande en condition *Bruit* $-r = 0,52$, $p < 0,001$– qu'en condition *Sans Bruit* $-r = 0,23$, $p < 0,001$).

4.5 Discussion

Cette étude en interaction a permis de voir ce qu'il advenait des comportements observés dans les études précédentes lors d'une tâche interactive plus naturelle. Au cours d'un jeu coopératif, le participant interagissait avec un complice. Ces deux personnes devaient reproduire un modèle de disposition d'objets sur un plateau de jeu : le participant disposait du modèle, le complice des objets. La réalisation de la tâche impliquait donc une interaction. La tâche utilisée élicitait naturellement des gestes de pointage, qui étaient l'objet d'étude. Les interactions ont été étudiées dans des conditions normales et dans des conditions perturbées (par un bruit de type *cocktail-party*) afin de voir l'influence de cette perturbation à la fois sur les stratégies mises en place pour réaliser la tâche, sur les productions dans les deux modalités et sur leurs liens temporels. Enfin, une autre modalité de la deixis langagière était également étudiée ici : les démonstratifs, dont l'utilisation était imposée par les modèles de phrases utilisés.

Les principaux résultats présentés ci-dessus cherchent à montrer l'influence du bruit sur les différentes productions et à caractériser la coordination geste/parole lors d'une interaction plus libre que dans les expériences précédentes. Vingt participants ont pris part à cette étude et les résultats sont tirés de leurs enregistrements. Globalement, il a été montré que l'interaction était modifiée par la présence de bruit : les interactions sont plus longues, contiennent plus d'erreurs et on observe une augmentation sensible du nombre (et de la durée) des

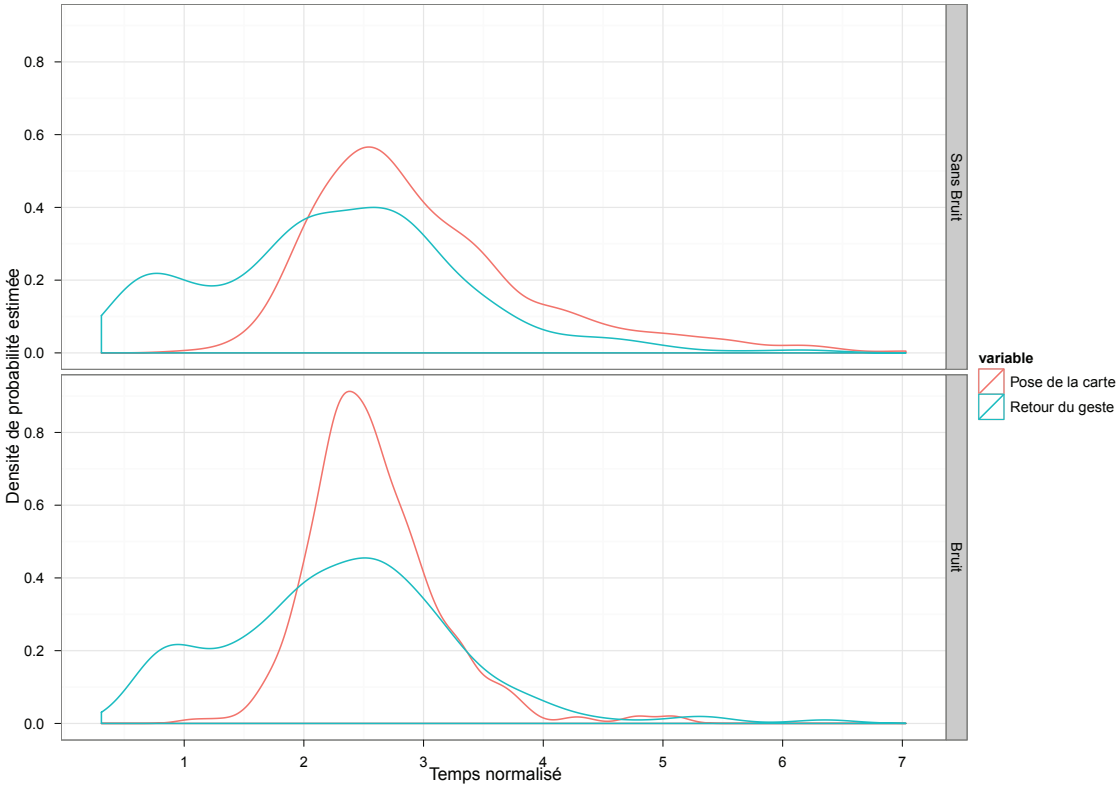


FIGURE 4.15 – Densités de probabilité des instants de retour et de pose de carte

regards vers l'interlocuteur en condition bruitée. Cependant, au niveau des stratégies de remplissage, aucun changement n'est observé (mis à part pour un participant qui change totalement sa stratégie pour les gestes manuels et produit des gestes de type iconique). Au niveau de la parole, on observe simplement un effet de type Lombard avec un débit de parole réduit, une augmentation des pics d'intensité et de fréquence fondamentale et une augmentation de l'amplitude d'ouverture des lèvres pour les voyelles "ouvertes". Pour les gestes manuels, l'influence du bruit se limite à faire baisser la variance des instants de production du début (et d'apex) du geste et a une influence sur l'instant de début du geste qui intervient plus tard par rapport à la parole en condition bruitée. Par ailleurs, il a été montré que les gestes étaient plus longs pour les durées non-normalisées et que l'allongement, localisé sur la tenue du geste, avait une durée proche de l'allongement de la durée des productions vocales. Enfin, l'étude de la coordination a montré que l'apex des gestes est assez largement cooccurrent au mot cible et plus précisément avec la première syllabe de celui-ci. Il a été montré un alignement avec la première cible articulatoire du mot-cible, dans les deux niveaux du facteur BRUIT et *a fortiori*, aucun alignement avec le démonstratif.

4.5.1 L'influence de la perturbation sur les stratégies mises en place

Comme résumé en introduction de cette partie, plonger une diade dans du bruit lors d'une interaction utilisant la parole seule a des conséquences très similaires à celles qui émaneraient d'un écartement physique des deux participants. Plusieurs théories permettent l'explication de l'effet Lombard. Plusieurs chercheurs pensent que cet effet n'est pas un réflexe et qu'une dimension sociale entre en compte dans le "réglage" de l'effet Lombard. C'est ainsi que Lindblom dans son article présentant sa théorie *Hypo&Hyper-articulation* propose que la "tension" menant à l'utilisation de l'hypo/hyper-articulation dépend de facteurs environnementaux (bruit, acoustique environnante) mais également de l'interlocuteur (personne malentendante etc…). D'autres pensent que c'est la modification (dégradation) du retour auditif qui entraîne de manière plus ou moins automatique le forçage vocal (et toutes les modifications qui s'en suivent).

Le cadre de la communication multimodale utilisant la parole et les gestes manuels est différent puisque les gestes manuels pourraient jouer un rôle dans l'amélioration de la transmission du message vocal. Ainsi, si les deux modalités coopèrent totalement dans la régulation, il est possible que les gestes manuels permettent de désambiguïser la parole (si les gestes manuels ont cette capacité), permettant alors à la parole d'éviter le forçage *i.e.* permettant, au prix d'un moindre effort, de conserver une "efficacité" communicative constante. Cette solution est *a priori* peu compatible avec les modèles reposant sur la théorie de McNeill présentés dans le chapitre d'introduction : selon une grande partie de ces modèles, ce qui est exprimé sous forme gestuelle et ce qui est exprimé en parole est "choisi" en fonction des caractéristiques sous-jacentes aux idées à représenter (spatial/non-spatial, typiquement). L'étude de l'évolution depuis la condition contrôle à la condition *Bruit* ne montre pas de différence dans les productions vocales (ce qui découle du protocole expérimental) ni de différence dans le contenu des gestes (totalement libre) sauf pour deux participants. Ces observations vont dans le sens de certains des modèles présentés ci-dessus : la partie du message délivré par la modalité geste est "calculée" pour les concepts spatiaux/dynamiques à transmettre et seulement pour ceux-là (*i.e.* la condition de communication, l'environnement, n'influe pas sur la forme du geste ni sur les concepts qu'il véhicule).

Cependant, comme mentionné ci-dessus, un participant a brutalement changé de stratégie communicative en réalisant des gestes iconiques en parallèle de l'articulation des mots-cibles afin d'illustrer ces mots (et les rendre, ainsi, plus "compréhensibles" par son interlocuteur). Bien qu'à considérer avec modération (puisque n'intervenant que sur un vingtième des participants), ce comportement est intéressant. En effet, il met en avant l'influence de la situation de communication sur la décision de la partie du message à encoder sous forme de parole/geste/les deux. L'environnement a un rôle à jouer dans le processus de décision sous-jacent : ici le participant en question choisit d'exprimer une même idée dans les deux modalités afin de pallier la dégradation de la communication par le bruit, en espérant améliorer la compréhension de son interlocuteur. (Cette stratégie a par ailleurs fonctionné puisqu'elle a évité un bon nombre d'erreurs –en particulier, confusions entre les paires de mots ayant des phonèmes proches : bambou/bonbon/pompon, chapeau/chameau).

Finalement, globalement les stratégies mises en place ne changent pas beaucoup entre les deux conditions *Sans Bruit* et *Bruit*. En particulier, pour la grande majorité des participants, aucun changement dans les formes de gestes réalisés n'est notable. Ceci est en accord avec les modèles selon lesquels l'environnement (ou, de manière plus restrictive, le bruit environnant) n'a pas d'influence sur le contenu des gestes manuels. Il est nonobstant notable qu'un participant s'est adapté à l'environnement bruyant en agrémentant l'articulation des mots-cibles de gestes iconiques afin (probablement) d'améliorer la transmission d'informations vers son interlocuteur. Cette adaptation prouve alors que, même si cela est marginal dans cette étude, l'environnement (et en particulier, le bruit environnant) influe sur la "répartition" de l'information entre les deux modalités. En particulier, l'environnement peut amener à la production de gestes pour désambiguïser la parole qu'ils accompagnent.

4.5.2 L'adaptation des deux modalités et de leur coordination à la perturbation

Il a été montré ci-dessus que les stratégies mises en place pour la réalisation de la tâche ne varient pas beaucoup dans la condition *Bruit* par rapport à la condition *Sans Bruit* (sauf pour deux participants, et en particulier un qui change totalement de stratégie). Ainsi, l'environnement bruyant a un effet limité sur la répartition des concepts entre les deux modalités (bien entendu, il est possible que la contrainte de parole y soit pour quelque chose, et il faudrait tester ces différences dans une tâche totalement non contrainte). Cependant, la littérature prédit un effet d'adaptation de la parole à l'immersion dans un bruit environnant : l'effet Lombard.

Adaptation de la parole La majorité des adaptations observées semblent s'opérer sur la modalité parole. Comme attendu, un effet Lombard est observé pour tous les participants, lorsqu'on plonge les participants dans le bruit, ceux-ci augmentent à la fois les pics d'intensité de la voix dans les énoncés ($+20\,\mathrm{dB}$) mais également les pics de F_0 ($+100\,\mathrm{Hz}$) et un allongement des productions vocales est observé. L'allongement n'est pas localisé mais bien réparti sur toute la production de parole. On observe par ailleurs, une augmentation des mouvements articulatoires d'ouverture pour les voyelles "ouvertes" (ce qui est cohérent avec la littérature) mais pas des mouvements de protrusion pour les voyelles "protruses"… Enfin, au niveau du contenu de la parole, celle-ci contient

plus d'erreurs de production (phrase incorrecte), et d'interruptions (hésitations, pauses,...). Ainsi, que ce soit au niveau temporel ou au niveau des amplitudes de la parole, les résultats trouvés sont les résultats classiques de l'effet Lombard : la production conjointe de gestes manuels ne compense pas cet effet. La production de gestes ne semble pas par ailleurs permettre d'éviter les erreurs de production dans les énoncés.

Adaptation des gestes Contrairement à ce qui est observé dans la modalité parole, les gestes ne subissent que peu de modifications que ce soit au niveau temporel ou spatial/dynamique. Par ailleurs, un résultat important est que, dans les deux conditions du facteur BRUIT, les gestes (de pointage) produits par les participants semblent majoritairement ancrés à la parole par la cooccurrence temporelle entre l'apex du geste et la première syllabe du mot-cible (et plus précisément la cible articulatoire de cette syllabe) *i.e.* la perturbation de l'environnement ne semble pas modifier la coordination entre les deux modalités. Globalement, la parole est ralentie, mais les instants de production des gestes *par rapport à la parole* ne sont pas modifiés (sauf une légère modification pour l'instant de début de geste). Ceci peut s'expliquer de plusieurs façons : soit les deux systèmes de production sont ralentis et la conservation de l'alignement temporel est un effet de bord de ce ralentissement, soit la production des gestes est adaptée en fonction de la modification de la production de la parole (l'inverse n'étant *a priori* pas concevable puisque le ralentissement de la parole dans le bruit est observé même lorsqu'aucun geste manuel n'est produit). Cependant, l'hypothèse d'un ralentissement global des productions n'est pas cohérente avec les résultats observés dans cette étude. En effet, il a été montré que l'instant de production de l'apex ou de tous les instants mesurés avant l'apex étaient *moins* variable en condition *Bruit i.e.* que la variance des productions était plus faible. Or si la production était impactée par un ralentissement général en condition *Bruit*, tous les instants seraient retardés par rapport à la condition *Sans Bruit* (ce qui n'est pas le cas) *et* tous les écarts-types seraient augmentés, ce qui n'est pas le cas. Par ailleurs, relativement à la durée de l'énoncé, les phases de *stroke* et *stroke de retour* sont *plus courtes* en condition bruitée, ce qui est contraire à un ralentissement global de la production.

Finalement, les résultats semblent être en adéquation avec une vision qui considèrerait que la production des gestes s'adapte à la parole mais dans une certaine mesure : dans le cadre du geste de pointage, il semble que les durées du *stroke* et du *stroke de retour* soient constantes même si la durée de la parole est modifiée ; il n'en va pas de même pour le plateau dont la durée semble varier avec la durée de l'énoncé correspondant. Ainsi il semble que les gestes de pointage produits dans l'expérience s'adaptent aux modifications subies par la parole par le biais d'une variable d'ajustement, la durée du plateau qui encadre la cible à désigner dans la parole (les autres durées sont fixées) et par l'instant de début du geste qui est déterminé de manière à respecter un alignement temporel entre l'apex et une cible dans la parole (en majorité, la première cible articulatoire du mot associé à l'objet à placer).

4.5.3 Alignements entre la parole et les gestes

Les deux études présentées dans les chapitres précédents ont mis l'accent sur une coordination forte entre la réalisation du geste (en particulier, le geste de pointage) et la parole qu'il accompagne. Plus précisément, il a été montré une coordination temporelle assez fine entre la production de l'apex du geste de pointage et des cibles articulatoires de l'énoncé accompagné (les indices prosodiques étaient également coordonnés au geste manuel, mais de façon moins fine). Les deux études précédentes se sont principalement intéressées à la coordination geste manuel/parole dans le cadre de la désignation lorsque celle-ci est réalisée par le biais de la focalisation prosodique en parole. Un des points importants soulevé dans la seconde expérience présentée était les mécanismes possiblement sous-jacents à la coordination entre les deux modalités. En particulier, deux hypothèses ont été avancées : il est possible soit que la fonction de désignation guide la coordination (et alors, focalisation et apex du geste de pointage devraient être coordonnés) soit que ce soit l'objet désigné qui guide la coordination (et alors, l'objet désigné et l'apex du geste de pointage devraient être coordonnés). Il a été montré dans le chapitre précédent que les deux stratégies étaient *a priori* possibles, avec une majorité de participants ayant un comportement explicable par une coordination guidée par l'objet désigné plutôt que par la fonction de désignation.

Dans l'étude présentée ici, la désignation est encore présente sous forme manuelle (geste de pointage, dans la plupart des énoncés) et sous forme de parole (utilisation imposée des démonstratifs). Cependant, la désignation

en parole ne désigne pas dans cette expérience l'objet mais se réfère directement à ce qui est montré par le geste (un emplacement sur le plateau de jeu). Dans cette étude, l'objet et la fonction de démonstration sont ici produits à des instants distincts en parole. L'étude temporelle des réalisations multimodales a permis de montrer clairement que la réalisation des gestes manuels n'était pas coordonnée avec la production des démonstratifs en parole. Par contre, pour tous les participants, les apex des gestes de pointage interviennent très majoritairement à des instants concordants à la production vocale de l'objet en question dans l'énoncé. Ainsi, les résultats de cette analyse semblent favoriser une interprétation selon laquelle les informations complémentaires se synchronisent plutôt qu'une vision favorisant le regroupement des fonctions de désignation (apex du geste/démonstratif). Une possibilité intéressante est que la production multimodale d'un geste de pointage et d'une parole associée peut octroyer un rôle important à la complémentarité offerte par les deux modalités. *Dans la tâche présentée dans ce chapitre, il est possible que le geste de pointage ait à la fois un rôle de désignation dans l'espace (à quel emplacement poser la carte sur le plateau de jeu, information indisponible en parole) et un rôle de désignation "temporelle" (à quel instant temporel se trouve l'information importante délivrée en parole).* Ainsi, la réalisation de l'apex du geste de pointage pourrait avoir un rôle "d'amorçage" permettant de focaliser l'attention de l'interlocuteur et ainsi d'améliorer la compréhension et d'éviter les erreurs. Cette interprétation est compatible avec les résultats de l'expérience présentée dans le chapitre précédent puisqu'alors la décision de l'information "importante" en parole pouvait varier selon les participants (dans tous les cas, l'objet ou sa couleur permettaient de distinguer la cible à désigner de la cible perturbatrice).

Enfin, comme dans les deux études présentées précédemment, il semble que la coordination entre la réalisation des gestes manuels (de pointage) et la parole s'opère autour d'une coordination entre l'apex des gestes manuels et les cibles articulatoires de la parole qu'ils accompagnent. Plus précisément, comme précédemment, les apex des gestes de pointage sont produits à des instants qui ne sont pas différents des pics de fréquence fondamentale et d'intensité ainsi que des cibles articulatoires. L'alignement le plus précis est, ici encore, observé entre les cibles articulatoires et l'apex du geste de pointage. Ainsi, la coordination des productions se fait, dans toutes les productions étudiées jusqu'ici, par la cooccurrence d'évènements moteurs saillants. Cela confirme que la vision d'une coordination reposant sur des systèmes dynamiques en interaction modulée par des contraintes communicatives semble être une possibilité intéressante pour considérer les interactions entre les gestes manuels coverbaux et la parole. Finalement, il serait intéressant de voir s'il existe une coordination tripartite entre évènements manuels - laryngés et supra-laryngés puisque d'après les travaux de D'Imperio *et col.*[42], il existe un ancrage fort entre les pics de F_0 (évènements laryngés) et supra-laryngés.

4.5.4 Limites de cette étude et conclusion

Deux limites importantes viennent perturber quelque peu les résultats présentés. Une limite subtile est l'ordre de passage des conditions *Sans Bruit* et *Bruit*. Cet ordre n'a pas été randomisé parmi les participants. Ainsi, chaque participant réalisait d'abord trois grilles dans la condition *Sans Bruit* puis celui-ci était plongé dans le bruit pour trois autres grilles. Les participants n'étaient pas informés de la présence d'une condition bruitée à la suite de la condition *Sans Bruit*. L'intérêt de ce *design* expérimental est principalement de ne pas avoir d'*after-effect* de l'effet Lombard dans les essais sans bruit (ce qui pourrait arriver si les essais bruités et non bruités étaient entremêlés. Cependant, l'absence de randomisation des essais pose le problème de l'habituation à une condition au travers des différentes répétitions (*i.e.* typiquement, il se pourrait que l'effet Lombard soit réduit par une habituation à la perturbation de l'interaction par la diffusion de bruit dans les écouteurs). Cependant, les tests statistiques ne donnent pas raison à un tel effet d'habituation (en tous cas, pour les moyennes portant sur les amplitudes/temps des données vocales) : s'il y a un effet d'habituation dû à la non-randomization des essais, cela n'est pas visible statistiquement. Cette limite est considérée uniquement théorique mais n'est pas confirmée *a posteriori*.

L'autre limite est l'utilisation de modèles de phrases imposés aux participants, cette limitation évidente découle du désir de pouvoir en partie tirer des conclusions statistiques sur les données présentées : l'utilisation de modèles de phrases et ainsi de productions identiques pour un grand nombre d'énoncés a permis d'obtenir un grand nombre de répétitions de phrases accompagnées de gestes de pointage lors de l'interaction, ce qui était le but. Une tâche identique sans limitation au niveau syntaxique est à réaliser afin d'utiliser un protocole

“totalement” non contraint. Un pas dans ce sens est évoqué dans le chapitre suivant.

Chapitre 5

Vers plus de liberté…

Notre objectif dans cette dernière étude est d'explorer quelques directions permettant d'aller vers une configuration expérimentale se rapprochant encore un peu plus d'un cadre d'interaction réellement naturel. Pour ce faire, nous avons exploré trois pistes de développement. D'abord, le mode de diffusion du bruit par casque présente certes de grands avantages en termes de contrôle de l'enregistrement de la parole produite, mais aussi certains inconvénients associés à un mode de communication typiquement « de laboratoire » d'autres solutions de diffusion possibles moins contraintes seront étudiées. Ensuite, le mode de communication dans l'expérience du chapitre précédent reste assez contraint, puisqu'il est basé sur des phrases du type « Le <objet> va là. ». La mise au point d'une expérience similaire sans imposer de modèles de phrases impose de trouver une façon de faire éliciter des gestes de pointage régulièrement aux participants. Cet enjeu n'est pas simple, il constituera le second objectif de cette étude. Enfin, l'annotation vidéo est une partie très importante pour l'étude d'interactions non contrôlées, en particulier, toute la bibliographie étudiant des interactions dyadiques multimodales utilise le signal vidéo comme outil principal d'étude. La mise au point d'une grille d'annotation efficace et complète est une étape complexe mais indispensable, elle fera l'objet du troisième développement de ce chapitre.

5.1 Perturbation de la situation de communication

5.1.1 Modes de diffusion du bruit

Il est naturel de penser que la façon dont on diffuse le bruit pour perturber la situation communicative va influer sur celle-ci. Il y a principalement deux modes opératoires pour diffuser le bruit sans "trop" entraver la situation communicative : il est possible de diffuser le bruit par l'intermédiaire de haut-parleurs, qui auront pour but de rendre l'environnement global plus bruyant, on peut également diffuser le bruit grâce à des casques audio ce qui permettra de limiter la perturbation du signal de parole par le bruit à l'enregistrement en restreignant le bruit à une zone proche des oreilles des interlocuteurs.

Bien entendu, la diffusion par haut-parleurs est le mode de diffusion de bruit le plus naturel puisqu'il n'implique pas le port d'équipements supplémentaires pour les interlocuteurs et ne les isole pas de leur interlocuteur, ne modifie pas l'aspect de la face de l'interlocuteur et ne modifie pas spectralement le retour auditif qu'a le locuteur de sa propre voix (hormis le fait qu'un bruit ambiant est superposé à celle-ci). Le problème de l'utilisation de haut-parleurs pour diffuser du bruit lorsqu'il est nécessaire d'enregistrer la parole vient du fait que le bruit ambiant diffusé par les haut-parleurs est également enregistré par les micros capturant la parole des participants. Les signaux enregistrés sont donc *a priori* inutilisables pour une étude acoustique de la parole. Cependant, quelques études portant sur des méthodes de débruitage ont été publiées (des méthodes d'estimation de canal [174] et des méthodes adaptatives comme présenté dans [147]) et donnent de bons résultats permettant de retrouver le signal de parole au sein d'un signal bruité. Comme l'a montré Garnier [56], les méthodes adaptatives donnent de bons résultats au niveau perceptif mais ne sont pas optimales pour l'étude acoustique des signaux débruités puisque le débruitage introduit des distorsions spectrales (parfois faibles) dans le signal de parole. Leurs performances sont par ailleurs assez limitées pour du bruit de type *cocktail-party*.

Par ailleurs, bien que l'utilisation de casques audio modifie le retour acoustique des participants, modifie les amplitudes des mouvements articulatoires [36, 56] et modifie l'apparence physique des participants par rapport à une présentation par haut-parleurs, elle ne modifie que peu les paramètres acoustiques de la parole qui sont utilisés dans ce manuscrit : dynamiques de l'intensité et de la fréquence fondamentale. La seule modification à garder à l'esprit entre les deux modes de présentation est le fait qu'un effet supplémentaire d'atténuation du retour auditif de la parole produite est dû au casque. Ceci peut se traduire par un effet sur l'amplitude des gestes articulatoires qui s'ajoute à l'effet Lombard par rapport au cas de la présentation du bruit par haut-parleurs.

Dans les deux cas, ces modes de présentation sont "valides" puisque le but principal des travaux présentés dans ce manuscrit est l'étude de la coordination des gestes manuels et de la parole plutôt que l'étude approfondie des amplitudes du signal de parole (étude formantique, intensité, ...). Au niveau temporel, le ralentissement de la parole fait partie de l'effet Lombard, et il est probable que la coordination geste manuel/parole prenne en compte les contraintes temporelles liées à l'effet Lombard : si la coordination est nécessaire pour des raisons communicatives (et/ou motrices), la perturbation d'un système (parole ou gestes manuels) est suivie dans l'autre modalité pour réaliser la coordination. En particulier, si cette coordination est réalisée dans un but communicatif, pour l'interlocuteur, la modification de l'environnement ne devrait pas perturber la coordination. Ceci peut justifier en partie l'utilisation d'un casque pour la diffusion du bruit : si l'utilisation d'un casque rajoute des contraintes temporelles supplémentaires sur la production, il est probable que la coordination geste manuel/parole prenne également en compte ces contraintes, pour les raisons exposées ci-dessus.

5.1.2 Méthode de débruitage pour la diffusion par haut-parleurs

La méthode qui semble donner les meilleurs résultats de débruitage pour l'étude acoustique du signal de parole plongé dans du bruit *cocktail-party* semble être la méthode présentée par Ternström *et col.* [174] avec des paramètres d'optimisation à faire varier, selon le mode opératoire suivi par Garnier [56].

Cette méthode de débruitage est simple et basée sur le principe illustré en Figure 5.1.

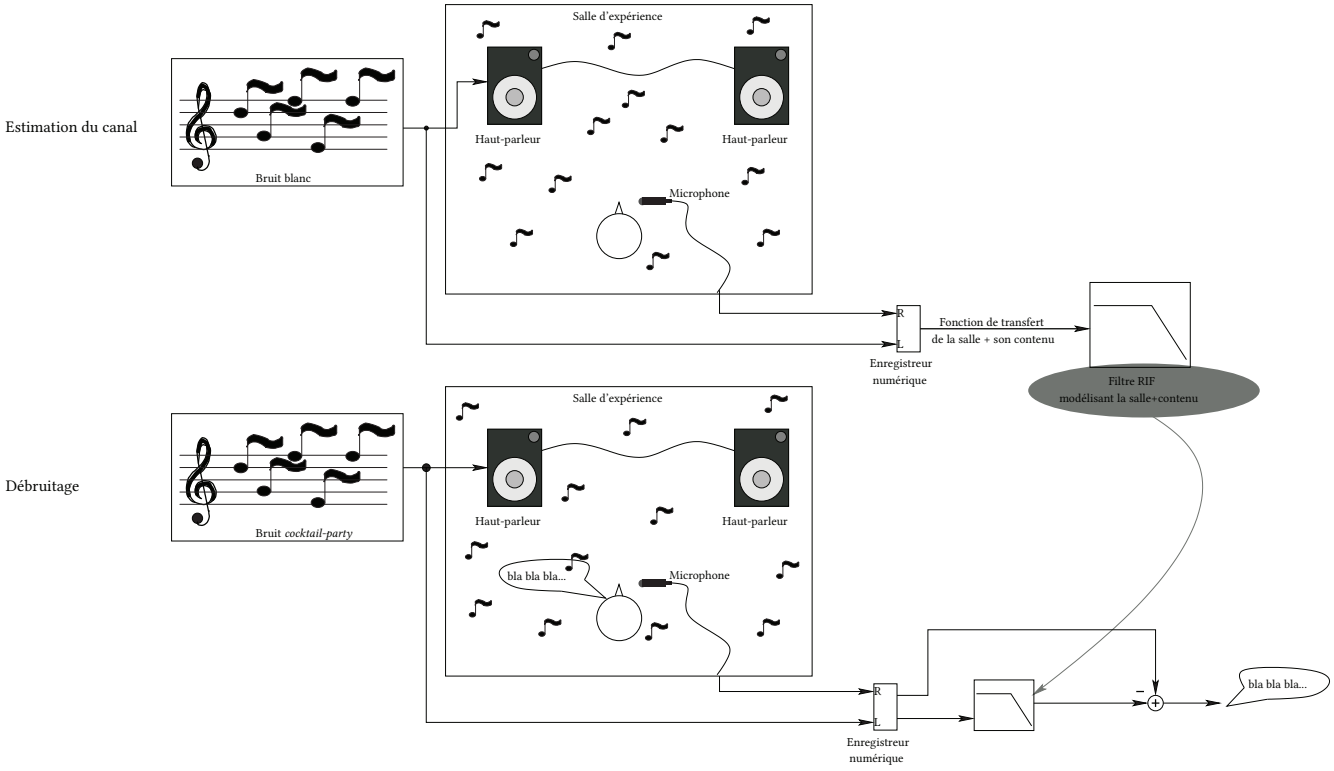


FIGURE 5.1 – Méthode de débruitage de Ternström *et col.* (d'après [174])

On évalue la fonction de transfert de toute la chaîne de diffusion / d'enregistrement d'une part. La fonction de transfert prend alors en compte les caractéristiques des cartes audio des ordinateurs, des haut-parleurs, du microphone, de la salle contenant tout le matériel nécessaire pour l'expérience ainsi que les protagonistes, de la carte d'acquisition de l'enregistreur numérique, de toute la connectique (prises, câbles,...), les temps de diffusion dans la salle. La fonction de transfert est estimée par une "différence" entre le signal de bruit passant par la chaîne de diffusion/acquisition et un signal directement acquis au niveau de l'enregistreur numérique. On génère à partir de cette fonction de transfert un filtre numérique à réponse impulsionnelle finie qui permet de filtrer toute cette chaîne.

Ce filtre à réponse impulsionnelle finie est ensuite utilisé pour filtrer le signal de parole dans le bruit. On enregistre la parole bruitée (le bruit est diffusé en utilisant exactement la même chaîne de diffusion/enregistrement que précédemment) sur une voie de l'enregistreur numérique alors que l'autre voie enregistre le bruit directement. On applique ensuite la fonction de transfert de la chaîne de diffusion/acquisition au bruit seul qu'on soustrait du signal de parole bruité. Il ne reste que le signal de parole en sortie de l'algorithme.

Il est évident, par la description de cet algorithme, qu'une condition *sine qua non* pour que le débruitage fonctionne correctement est l'invariabilité de la fonction de transfert de la chaîne de diffusion/enregistrement. En particulier, il faut absolument que la fonction de transfert de la salle soit invariable entre l'estimation du canal et l'enregistrement.

Cette dernière condition est difficile à réaliser dans une condition où deux personnes interagissent pour réaliser une tâche coopérative. En effet, lorsque les participants bougent dans la salle, leurs mouvements font évoluer la fonction de transfert de celle-ci au cours du temps. *A priori*, ceci rend impossible l'utilisation d'un tel système de débruitage. Cependant, les mouvements utilisés dans les expériences pilotes ci-dessous étant très limités, il est possible que ceux-ci ne modifient que marginalement la fonction de transfert de la salle et rende ainsi ce protocole viable pour le débruitage des signaux. C'est une des questions qui sera testée dans les scenarii d'interaction présenté ci-après.

5.2 Mise au point d'un protocole expérimental

La mise au point d'un protocole complet permettant de respecter les contraintes posées initialement est chose assez complexe. Le but principal de l'étude est la caractérisation de la coordination entre les gestes manuels et la parole lors d'une interaction entre deux personnes. La tâche utilisée est constituée d'un jeu qui impose assez spontanément une interaction entre les deux participants. Dans ce jeu, le participant dispose d'un "modèle" qui représente une organisation spatiale d'objets. Un complice interagit avec le participant et ne peut pas voir le modèle. Par contre, ce complice dispose d'une réserve de cartes (invisible pour le participant) *a priori* infinie représentant les objets à disposer sur la zone de jeu. Une zone de jeu est disposée dans la salle (emplacement variable selon les étude pilotes). Sur cette zone de jeu, le modèle est reproduit mais quelques objets sont absents (par rapport au modèle). Le but du jeu est, pour le participant, d'aider le complice à poser ses cartes sur la zone de jeu de façon à reproduire le modèle.

Les études pilotes testées sont présentées ci-après, elles ont été réalisées en laissant les participants se comporter selon ce qui leur semblait le plus confortable : aucune contrainte (en particulier au niveau des phrases à prononcer) n'était donnée. La seule consigne était de rester assis(e) et d'essayer de résoudre la tâche le plus rapidement possible.

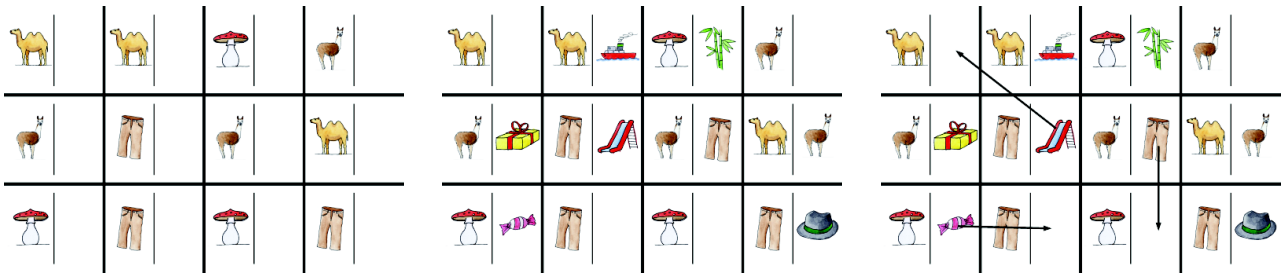
Cette tâche est assez simple à comprendre (d'après les premiers essais effectués, et cela sera confirmé *a posteriori*), ce qui est une des contraintes à remplir. Le but de ces étude pilotes est d'obtenir une tâche coopérative naturelle dans laquelle la production de pointage soit naturelle (*i.e.* non forcée) mais effective.

5.2.1 Première étude pilote

Dans cette première étude pilote, le complice devait réaliser des erreurs de positionnement des cartes volontairement afin que le participant corrige ce positionnement erroné par la production de focalisation prosodique.

Trois “grilles” sont utilisées dans cette étude : la grille servant de plateau de jeu, la grille modèle (uniquement visible par le participant), la grille d’erreur (apprise par cœur par le complice). Trois grilles d’exemples sont données en Table 5.1.

Les grilles sont composées de paires qu’il faut compléter. Sur la grille servant de plateau de jeu, chaque case est séparée en deux par un trait vertical fin, la demi-case gauche contient un dessin et la demi-case droite ne contient rien. Dans le modèle, certaines demi-cases droites contiennent un dessin, le participant doit donc faire remplir ces cases par le complice en lui indiquant l’image à disposer et à quel endroit la disposer. Plusieurs demi-cases gauches sont identiques pour essayer d’éliciter des gestes de pointage. L’utilisation de paires de dessins se justifie par le fait qu’un des desiderata était d’éliciter de la focalisation contrastive prosodique lors de potentielles erreurs de placement : par exemple, selon les erreurs données dans la grille de droite de la Table 5.1, si à une phrase du type “*Le toboggan va avec le chameau*”, la carte correspondant au toboggan était posée à côté du pantalon, on s’attendrait à une correction du type “*avec le **chameau***”, insistant sur l’objet associé à la carte.



Grille projetée sur le tableau | Grille modèle | Grille d’erreurs pour le complice

Table 5.1 – Les 3 types de grilles utilisées dans la première étude pilote

Lors de la réalisation de cette étude pilote, la salle d’expérience était disposée comme représenté grossièrement en Figure 5.2. Le plateau de jeu est posé sur un chevalet à hauteur du regard pour le participant. Le plateau est placé verticalement afin d’éliciter des pointages canoniques, vers l’avant et surtout afin d’éviter des problèmes qui pourraient donner lieu au masquage des diodes Optotrak dans le cas d’un plateau à l’horizontale avec les deux participants de chaque côté du plateau. Le complice se trouve à côté du plateau de jeu et le laisse totalement visible pour le participant. Par l’intermédiaire d’un vidéoprojecteur, une image de la zone de jeu est diffusée sur le plateau de jeu en même temps que s’affiche le modèle sur un écran (invisible pour le complice). Les mouvements sont enregistrés par un système de suivi de mouvement comme dans les expériences précédentes, un signal audio est acquis par l’intermédiaire d’un micro et la vidéo est utilisée pour pouvoir étudier l’étude pilote après enregistrement de celui-ci. Dans les conditions bruitées, le bruit *cocktail-party* est diffusé par l’intermédiaire de haut-parleurs placés deux mètres en avant du participant, et espacés de deux mètres l’un de l’autre. Le niveau de bruit utilisé pour ce mode de diffusion est de 85 dB(C) mesurés au niveau de l’oreille du participant ce qui permet d’avoir un bruit gênant tout en étant non douloureux. La mesure du niveau de bruit a été réalisée avec un sonomètre (Lutron SL-4001).

Le participant est libre de donner les instructions qu’il veut afin que le complice reproduise le modèle. Le complice dispose d’une réserve de cartes qui s’aimantent sur la zone de jeu. Celui-ci feint de se tromper d’emplacement de carte aux endroits correspondants à la grille d’erreur apprise par cœur.

Cette étude pilote a été menée sur un participant.

Conclusions de la première étude pilote Ce premier essai n’a pas été concluant pour plusieurs raisons. Tout d’abord la tâche s’est avérée inefficace pour une élicitation naturelle de gestes de pointage : le participant a mis à profit l’organisation en grille régulière des images pour donner des instructions au complice sous la forme “*Sur la première ligne, dans la troisième case, il y a un bambou*”, rendant inutile tout pointage. L’enseignement tiré de cette absence de pointage est le fait qu’il faut organiser les cibles de manière plus chaotique au sein de la zone de jeu.

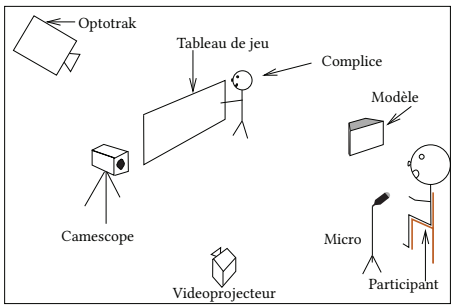


FIGURE 5.2 – Organisation de la chambre sourde pour la première étude pilote

Au cours d'un débriefing suivant l'étude, il a été évoqué le fait que la distance entre le participant et le complice était trop grande et que ceci avait pu avoir eu une influence sur la production de gestes accompagnant la parole et également que la position "de côté" du complice était gênante du fait que le participant pensait que la cible de son geste de pointage pourrait être mal interprétée. Dans les études pilotes suivantes, la distance entre les deux personnes est donc réduite. Il est à noter cependant qu'il est indispensable de conserver le tableau de jeu éloigné du participant pour des raisons techniques : le tableau ne doit pas masquer les diodes émettrices infra-rouge du dispositif de suivi de mouvement.

Par ailleurs, le débriefing du complice a confirmé des craintes sur les "erreurs". Bien que la production d'erreurs volontaires dans la condition bruitée puisse être prise pour des erreurs involontaires, il est très difficile de feindre des erreurs en condition non bruitée. La production d'erreurs volontaires est totalement non naturelle en condition non bruitée (surtout si le nombre d'erreurs à produire est important, comme cela était le cas puisqu'un des buts était d'éliciter beaucoup de focalisation prosodique). Par la suite, l'idée de produire des erreurs volontaires a été abandonnée.

Enfin, bien que peu de données aient été acquises dans le bruit, l'algorithme de débruitage n'a pas donné de bons résultats : le signal de parole débruité est distordu (effet "cathédrale") et contient encore une bonne part de bruit ce qui le rend inexploitable pour une analyse acoustique. Une raison possible est que l'enregistreur numérique utilisé a nécessité de nombreux branchements, ce qui n'est clairement pas optimal. Dans les études pilotes suivantes, l'enregistreur numérique a été changé pour un enregistreur numérique multipiste, permettant facilement de réaliser le traitement de données préconisé dans [174]. Par ailleurs, il est possible que le microphone utilisé ne soit pas assez directif ce qui ne permet pas d'améliorer le rapport signal sur bruit dans les situations bruitées comme le permettrait un microphone directif. Un microphone tour de cou (Shure beta53) beaucoup plus directif a été utilisé dans l'étude pilote suivante afin d'améliorer le rapport signal sur bruit du signal de parole enregistré.

5.2.2 Seconde étude pilote : changement du plateau de jeu et de la disposition de la salle

Afin de prendre en compte les différentes conclusions de l'étude pilote précédente, quelques modifications basiques ont été apportées au protocole initial. Les grilles d'erreurs sont obsolètes et aucune erreur volontaire n'est réalisée par le complice. La zone de jeu a été redessinée afin d'être plus complexe et moins régulière pour avoir plus de chances d'éliciter des pointages. Deux types de zones de jeu ont été testés (voir un exemple sur la Table 5.2) : un où seules des images de référence sont disposées sur la zone de jeu au départ, sans aucun trait supplémentaire dessiné (Type 1) et une où des "zones" ayant des formes non régulières sont dessinées (formes d'écailles, Type 2). Enfin, le participant et le complice sont assis l'un à côté de l'autre afin de réduire la distance entre ceux-ci, facteur qui semblait empêcher le participant de la première étude pilote de produire des gestes de pointage. Ce rapprochement des deux interlocuteurs ne s'accompagnant pas d'un déplacement du plateau de jeu (pour des raisons techniques évoquées ci-dessus), le complice se trouve à distance de la zone de jeu et ne peut donc pas aimanter les images comme précédemment. Une solution informatique au problème a donc été trouvée :

la “grille” de jeu est projetée sur le plateau de jeu par un vidéoprojecteur, le complice dispose d'une souris et d'un écran invisible au participant. Sur l'écran s'affiche sa réserve de “cartes” et il peut déplacer ces cartes vers la zone de jeu grâce à la souris (lorsqu'une carte quitte la réserve de cartes, elle apparaît sur la zone de jeu et est déplaçable grâce à la souris). Une représentation simplifiée du dispositif expérimental est donnée en Figure 5.3.

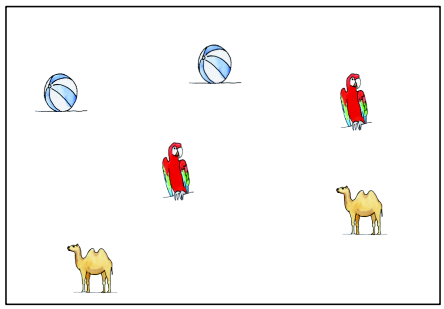
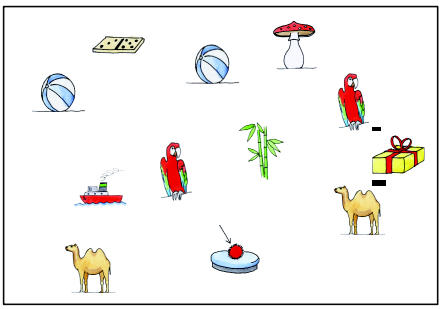
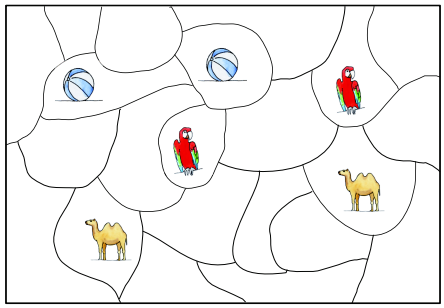
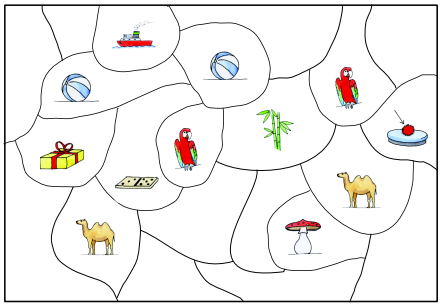
Type 1		
	ou	ou
Type 2		
	Grille projetée sur le tableau	Grille modèle

TABLE 5.2 – Le plateau de jeu et le modèle de la seconde étude pilote

Ici encore, le participant est libre de donner les instructions qu'il désire, aucune consigne ne lui est fournie à part le fait qu'il doit réussir, avec l'aide du complice, à reproduire la “grille” modèle qui est invisible au complice. Les conditions bruitées sont à nouveau réalisées en diffusant le bruit par des haut-parleurs au même niveau sonore que précédemment (85 dB(C)). La chaîne de diffusion/enregistrement a été épurée en comparaison avec l'expérience pilote précédente comme mentionné ci-dessus.

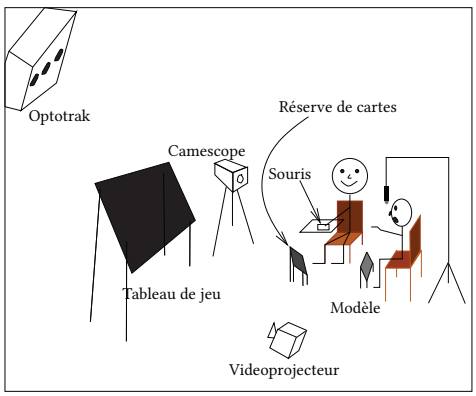


FIGURE 5.3 – Organisation de la chambre sourde pour la seconde expérience pilote

Un seul participant a pris part à cette étude pilote.

Conclusions de la seconde étude pilote Cette seconde étude pilote ne s'est pas révélée beaucoup plus concluante que la première. En effet, ce nouveau dispositif expérimental ne s'est pas révélé suffisant pour éliciter des pointages, mis à part à de très rares occasions.

Le participant s'est appuyé sur les images présentes sur le plateau de jeu au départ comme de références pour faire placer les images au complice. Ainsi, la grande majorité de ses explications étaient du type "*Entre les deux ballons en haut à gauche, il y a un domino*" ou "*Au dessus du perroquet à gauche, il y a un domino*" puis la position précise de l'objet cible était ajustée par des petites précisions du type "*Un peu plus haut...*", le placement des objets étant interactif. Il est possible que le fait d'avoir utilisé des objets ayant des appellations différentes comme objets présents au départ ait rendu l'utilisation de pointages moins utile : si on utilise deux "perroquets" dans la grille de départ, l'un est forcément à gauche de l'autre ou bien l'un est forcément au dessus de l'autre, et on peut ainsi différencier les deux répétitions d'un même dessin facilement puis donner des instructions à partir de cet objet de référence. Une solution simple pour contourner ce problème est d'utiliser un seul et même objet de référence dans la zone de jeu de départ, dans la suite, des objets plus "abstraits" seront utilisés. L'intérêt d'utiliser une multitude d'objets différents était d'essayer d'éliciter de la focalisation prosodique sur ces mots lors des erreurs de placement, mais cela semble trop complexe à mettre en place si on veut garder une durée d'expérience raisonnable.

L'utilisation de motifs sous forme d'écailles, plus complexe *a priori* que les cases rectangulaires des grilles de la première étude pilote car moins organisés, n'a pas eu d'impact sur la façon de procéder du participant. Les plateaux divisés en écailles ont été complétés par des directives du type "*dans l'écaille à gauche du perroquet de gauche, il y a un domino*", ne nécessitant pas la production de pointages pour des raisons similaires à la première étude pilote.

Une fois encore, le debriefing post-expérience a permis de mettre au jour la raison principale, selon le participant, pour laquelle aucun geste de pointage n'est produit : les distances entre le participant, le complice et le tableau *ou* le fait qu'il n'y ait pas d'interaction "réelle" avec le tableau semblent poser problème. Dans la suite, il a été décidé de disposer le plateau de jeu horizontalement sur une table autour de laquelle sont assis les deux interlocuteurs. Ceci permet donc de réduire la distance entre les deux interlocuteurs, la distance entre les interlocuteurs et le plateau et place les deux participants dans une situation d'interaction assez courante, autour d'une table. Cette situation avait été volontairement écartée en premier lieu pour éviter les problèmes liés au masquage des diodes Optotrak par le complice. Une étude de positionnement précis s'avérait donc nécessaire pour perturber le moins possible l'interaction tout en gardant une bonne visibilité des diodes.

Enfin, le nouvel essai de débruitage ne s'est pas révélé satisfaisant, même en utilisant une chaîne de diffusion/enregistrement épurée, un microphone plus directif et en faisant varier les paramètres de l'algorithme de débruitage (taille de la fenêtre de la transformation de Fourier rapide). On observe toujours un effet "cathédrale" et un débruitage de mauvaise qualité. Ceci peut s'expliquer par le fait que la fonction de transfert estimée en début d'expérience l'est alors que tout est immobile dans la salle d'expérience. Or, typiquement, au cours de l'expérience, les deux participants font des mouvements qui, même s'ils ne sont pas forcément très amples, changent la fonction de transfert de la salle. En conséquence, il est évident, d'après la Figure 5.1, que si la fonction de transfert est différente entre l'estimation et la passation de l'expérience, le débruitage n'est pas efficace. L'algorithme ne semble donner de bons résultats que lorsque les mouvements sont réduits au minimum. Dans la suite, cette méthode est donc abandonnée au profit d'une diffusion du bruit par casque fermé. Cette méthode rend l'étape de débruitage inutile.

5.2.3 Troisième étude pilote : changement du plateau de jeu, de la disposition de la salle et du mode de diffusion du bruit

Les conclusions précédentes ont mené le protocole expérimental à évoluer de la façon suivante :

- Le plateau est posé horizontalement sur une table
- Les participants sont face à face autour de la table
- Les éléments présents sur le plateau de départ sont des cubes plutôt que des images
- Le bruit est diffusé par casque dans les conditions bruitées : le participant *et* le complice écoutent le bruit

à un niveau similaire dans des casques similaires. Les deux participants sont informés que leur interlocuteur entend le même bruit qu'eux, les deux personnes sont en quelques sortes "plongées dans le même environnement" afin que la communication soit la plus naturelle possible. Le niveau de bruit utilisé est également de 85 dB(C) mesurés en sortie de casque par un sonomètre Lutron SL-4001 (les deux écouteurs du casque sont collés l'un à l'autre et le sonomètre est inséré entre les coussinets des écouteurs pour réaliser la mesure).

Les deux types de plateaux de jeu sont toujours utilisés, ceux-ci ne sont plus projetés par un vidéoprojecteur du fait de l'organisation horizontale. Les grilles sont dessinées à la main au début de chaque expérience et les cubes sont placés en début d'expérience également. La Figure 5.4 représente grossièrement la mise en place expérimentale de cette troisième expérience pilote.

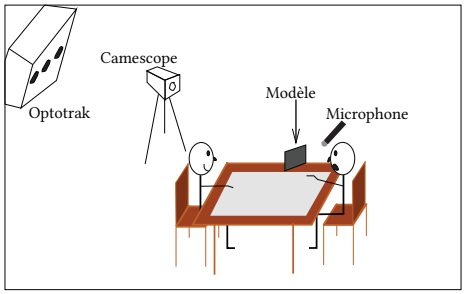


FIGURE 5.4 – Organisation de la chambre sourde pour la troisième étude pilote

Deux positions différentes ont été testées pour le complice : soit en face (comme présenté sur la Figure 5.4) soit sur le côté droit du participant. La condition en face à face est plus courante pour des jeux mais il est possible que le complice cache les diodes Optotrak ; la condition sur le côté permet de ne pas avoir le problème de diodes cachées mais il est possible que cette situation de communication élicite peu de gestes de pointage du fait que les deux interlocuteurs n'aient pas le même "axe" de vision. Ainsi, les deux positions sont testées afin de vérifier si un masquage a lieu et si celui-ci rend les données inexploitables (pour la position face-à-face) et si l'interaction de côté règle ces problèmes et ne gêne pas le comportement des participants.

Deux participants ont pris part à cette étude, aucune consigne n'est donnée à part le but de la tâche qui est de faire reproduire le modèle au complice.

Conclusions de la troisième étude pilote Cette expérience pilote est concluante, les participants utilisent le geste de pointage naturellement – et ce, dans les deux positions du complice –, et le bruit semble à un niveau assez élevé pour gêner la communication, sans être douloureux, même à faible distance. Pour la position en face-à-face, les diodes Optotrak sont trop souvent masquées pour permettre l'étude des données. Dans la position de côté, l'interaction se déroule correctement et aucun masquage des diodes n'est perceptible. Il conviendrait donc que le complice soit placé à une position intermédiaire entre la position en face-à-face et de côté afin d'éviter de masquer les diodes tout en gardant une situation de communication la plus naturelle possible et élicitant le maximum de pointages.

La visibilité des diodes Optotrak est par ailleurs affectée par le dispositif expérimental : les diodes sont également masquées par les cubes posés sur le plateau de jeu. Plus particulièrement, l'apex des gestes est souvent caché par les cubes, ce qui est gênant pour l'étude de la coordination des gestes de pointage avec la parole (*cf.* les parties précédentes). Il serait donc nécessaire de remplacer les cubes par des objets sans relief, ceux-ci seront substitués par des petits points dessinés au stylo sur la zone de jeu. Ceci permettra de signaler des positions (toutes repérées par le même symbole), sans avoir de relief et donc en évitant de masquer les diodes Optotrak. Par ailleurs, l'utilisation d'une grille composée d'"écailles" ne semble avoir aucun intérêt lorsque la zone de jeu est à portée du participant. Cet habillage de la grille devra donc être abandonné par la suite afin d'avoir une zone de jeu la plus épurée possible.

Finalement, la diffusion du son par casque fermé donne de bons résultats : la communication est perturbée et le bruit est inaudible dans le signal de parole acquis avec le micro, aucun débruitage n'est ainsi nécessaire.

5.2.4 Conclusions des différentes études pilotes

Les différentes études pilotes mises en place ont conduit à écarter des mises en places et protocoles expérimentaux qui n'auraient pas permis d'éliciter de pointages ou auraient conduit à des données manquantes. Une dernière étude pilote a été menée sur un participant pour valider le protocole final et faire quelques derniers ajustements. Une personne a pris part à cette étude pilote, le comportement observé s'approche assez fortement du comportement attendu : le participant comprend facilement la tâche et utilise naturellement et très fréquemment le geste de pointage, comme dans l'étude pilote précédente. Les problèmes de visibilité des diodes sont réglés par l'utilisation de points écrits sur le plateau de jeu ne contenant aucun autre repère.

Ces études pilotes sont le reflet de la complexité qui existe lorsqu'il s'agit de mettre en place une expérience à la confluence des études très contrôlées en laboratoire et des études "terrain" : il est nécessaire de procéder à de multiples ajustements pour trouver une tâche naturelle, permettant une interaction non artificielle entre les participants, élicitant des gestes (pourtant très fréquents habituellement) et respectant des contraintes techniques fortes (visibilité des diodes).

5.3 Élaboration d'un processus d'annotation vidéo

La réalisation d'une expérience d'interaction non contrainte impose une étude complexe des énoncés prononcés (*a priori* plus complexes que dans les cas contraints), des gestes réalisés (*a priori* plus complexes que dans une expérience contrainte, bien que la tâche soit ici construite afin d'éliciter un maximum de gestes de pointage) ainsi que de leurs relations. Une large majorité des études modernes se basant sur des enregistrements vidéos annotent ces vidéos en utilisant des grilles d'annotation (voir par exemple les travaux de Colletta *et col.* [32] pour la mise au point d'une grille pour l'annotation multimodale de vidéos d'enfants). Ces grilles permettent de repérer les instants importants des gestes et de la parole et de qualifier chaque laps de temps annoté. Typiquement, il est possible de repérer les gestes, de les qualifier (type de geste sur le continuum de Kendon par exemple), d'en définir les différentes phases, le sens qui leur est associé, etc... On peut également annoter les énoncés en précisant leur contenu lexical, puis en qualifiant chaque unité lexicale plus précisément selon l'étude ultérieure menée (typiquement, qui réalise l'énoncé, à qui il est destiné, quelles sont les fonctions des différents mots, quel est le but de l'énoncé, ...). L'intérêt de l'utilisation des grilles est assez évident : celles-ci permettent de consigner tous les éléments considérés comme variables d'intérêt pour l'étude en cours sous forme compacte (elles évitent *a priori* de visionner un nombre redhibitoire de fois les vidéos afin d'en synthétiser les caractéristiques). Elles permettent en particulier de regrouper les annotations faites sur un grand nombre de vidéos ensemble et d'en tirer des caractéristiques communes, de faire des études statistiques sur certaines grandeurs, etc...

Dans le cadre des études présentées ici, la grille d'annotation joue non seulement un rôle classique de description de l'interaction comme dans la majorité des études mais a également un rôle important dans l'étude de la coordination. En effet, les études s'intéressent à la *coordination* entre les gestes manuels et la parole, or la vidéo est le seul signal enregistré qui permet de lier ce qui se passe au niveau de la parole et au niveau du geste manuel (les enregistrements audio et la capture de mouvements seuls ne permettent pas de faire ce lien). En particulier, seul le visionnage de la vidéo permet de savoir quel geste est associé à un énoncé donné : l'affectation d'un geste manuel à un énoncé (si celle-ci est faisable) émane d'un jugement subjectif lié à la perception multimodale de la scène complète. Finalement, l'annotation vidéo est une étape nécessaire ici pour étudier la coordination, en plus de son rôle important dans la description qualitative des interactions. À ces fins, une grille d'annotation vidéo permettant de repérer les informations d'intérêt a été élaborée. Cette grille est le fruit d'un travail selon la méthode par essais et erreurs. Les contraintes de construction de la grille sont : son but descriptif décrit ci-dessus qui guide les informations à faire apparaître dans la grille et sa nécessité d'être utilisable pour *toutes* les interactions à annoter qui contraint la façon dont cette grille est implémentée (quelles subdivisions pour chaque point à décrire). Ainsi, à partir d'une grille initiale, une procédure de validation a été entreprise (en annotant une dizaine

d'interactions dans des conditions variées et pour des participants différents) afin de voir si la grille permettait l'annotation complète de chaque interaction. Si une incohérence/incomplétude de la grille était détectée, celle-ci était modifiée et la procédure de validation renouvelée jusqu'à l'obtention d'une grille permettant l'annotation des interactions jugées les plus complexes à annoter (gestes aux trajectoires variées, erreurs fréquentes en parole, stratégies originales, ...). La grille résultante est présentée en Figure 5.5. Cette grille a passé les stades de validation et a en particulier été utilisée dans l'expérience présentée au Chapitre 4.

La base constituant la grille d'annotation est constituée de la segmentation acoustique réalisée préalablement (segmentation des interactions en délimitant chaque énoncé et, au sein de ceux-ci, en découpant les mots en syllabes). Sur cette base, la grille d'annotation se décompose en plusieurs catégories qui représentent les points d'intérêt de l'étude. Toutes les catégories annotées sont représentées en Figure 5.5. Les catégories sont découpées en cinq thèmes. Dans la Figure 5.5, chaque table représente un thème, la première colonne est le nom de la catégorie, la seconde colonne représente les valeurs autorisées pour cette catégorie et la troisième colonne est une description de la catégorie. Chacun de ces thèmes est décrit ci-dessous.

L'annotation de ce qui est relatif à la parole prononcée par le participant se fait dans les catégories dont le nom est prefixé par `Utterance` ou `Units`. Ces deux premiers thèmes de catégories englobent des informations qui sont spécifiques à un énoncé ou bien aux unités d'intérêt au sein de cet énoncé. Un troisième thème de catégorie permet l'annotation des gestes de pointage, un quatrième la direction estimée du regard et un dernier permet d'annoter le comportement du complice. Une description rapide des différentes catégories et leur raison d'être (si celle-ci n'est pas évidente) est donnée ci-dessous.

Annotation des propriétés de l'énoncé

- La catégorie `Utterance-Ref` est importante puisqu'elle identifie de manière unique chaque énoncé. Cet identifiant est le seul moyen utilisé par la suite pour lier les gestes manuels et la parole : chaque énoncé a un identifiant unique, et si un geste est associé à cet énoncé, il sera identifié avec le même identifiant. Cette catégorie permet également de relier un énoncé à un autre dans le cas de répétitions/corrections/... par l'intermédiaire d'un suffixe accolé à son numéro de référence.
  Dans le cas d'un énoncé contenant plusieurs cibles ou plusieurs démonstratifs, l'identifiant est suivi d'un suffixe permettant la découpe de l'énoncé en plusieurs parties. Par exemple, l'énoncé numéroté `5` : « *il y a un bonbon là et un ballon ici* » est découpé en deux parties : `5-1` qui correspond à « *il y a un bonbon là* » et `5-2` qui correspond à « *et un ballon ici* ». Dans le cas d'un énoncé complexe, l'utilisation des suffixes indiquant la correction/répétition est toujours possible (par exemple `5-1_C4-3` est l'énoncé identifié `5-1` et qui corrige l'énoncé `4-3`).
- `Utterance-Type` permet de décrire le nombre de cibles citées dans l'énoncé. Ceci permet en particulier de repérer les stratégies mises en place par les participants pour réaliser la tâche (les mêmes objets sont-ils regroupés dans une même phrase, par exemple) et permet également de guider l'annotation/interprétation des gestes associés (si plusieurs objets sont présents dans l'énoncé, il y a de fortes chances que plusieurs gestes soient associés).
- `Utterance-Complexity` permet de classer rapidement les énoncés simples des énoncés corrigés ou répétés (l'information est redondante avec la catégorie `Utterance-Ref` mais permet une recherche rapide dans les annotations).
- `Utterance-Target` permet de désigner la "cible" (le mot-cible du corpus) utilisé dans l'énoncé. Si l'énoncé contient plusieurs cibles, celui-ci est découpé en plusieurs parties, chaque partie contenant une "cible" dans cette catégorie.
- `Utterance-Interruption` donne un moyen rapide de savoir si l'énoncé est interrompu ou non (en particulier, s'il y a des hésitations, de longues pauses, ...) Ceci est important puisque les énoncés contenant des hésitations ou de longues pauses peuvent être étudiées à part.
- `Utterance-Validity` est utile lorsque les phrases sont fixées (*cf.* Chapitre 4 par exemple) puisque cette catégorie permet, lors des analyses suivantes, d'écarter toute production qui ne serait pas exactement tirée du corpus. Cette catégorie est utilisable dans le cas d'une étude contenant des énoncés fixés comme cela était le cas dans l'étude présentée au Chapitre 4.

Utterance-Ref	**Numéro (unique) de référence de l'énoncé**	
	-<numero>	Énoncé en plusieurs parties (plusieurs mots-cibles)
	_Ct<Ref>	Énoncé corrigeant l'objet placé en énoncé <Ref>
	_Cp<Ref>	Énoncé corrigeant l'emplacement utilisé en énoncé <Ref>
	_R<Ref>	Répetition de l'énoncé <Ref>
Utterance-Type	**Contenu de la phrase (nb de cibles à placer)**	
	1	Une cible
	2	Deux cibles identiques
	1+1	Deux cibles distinctes
	+	$>$ 2 cibles
	0	Sans rapport avec la tâche
	end	Fin de grille
Utterance-Complexity	**"propreté" de l'énoncé**	
	Simple	Énoncé simple
	Autocorrection	Énoncé avec autocorrection
	Repetition	Énoncé avec répétition
	...	Combinaison de ci-dessus
Utterance-Target	**Quelle cible désignée par l'énoncé**	
	<corpus>	Un mot du corpus
	None	Pas de mot du corpus
Utterance-Interruption	**Énoncé interrompu ou non**	
	None	Sans interruption
	Pause	Avec pause
	Hesitation	Avec hésitation
	Laugh	Avec rire
	...	combinaison de ci-dessus
Utterance-Validity	**Phrase dans le corpus ou non**	
	Valid	Phrase du corpus
	Not corpus	Production hors corpus
	Error	Phrase du corpus + erreur

Units	**Transcription du discours découpé en unités d'intérêt**	
Units-Type	**Type d'unité**	
	Target	mots-cibles
	Demonstrative	démonstratifs
	Other-InformationPlace	*Info* ∋ emplacement objet (pas démonstratif)
	Other-InformationNumber	*Info* ∋ qté d'objets à placer
	Other-Instruction	*Ordre* (*prends ça*)
	Other-InstructionPlace	*Ordre* ∋ emplacement objet (pas démonstratif)
	Other-InstructionNumber	*Ordre* ∋ nombre d'objets
	Other-Negation	Négation (*non*)
	Other-Conversation	Rien à voir avec la tâche (*j'ai mangé des patates a midi*)
	Other-Comment	Commentaire (*ah oui j'avais oublié le bateau*)
	Other-Confirmation	Confirmation (*Oui c'est ça*)
	Pause	pauses
	Laugh	rires
	Hesitation	hésitation
	Top	top de debut/fin de tâche
	Others	tout le reste
Units-DemonstrativeFunction	**Fonction d'un démonstratif**	
	Place	désignation d'un emplacement
	Object	désignation d'un objet
Units-NbContentWords	**Décompte du nombre de mots de contenu**	
Units-NbFunctionWords	**Décompte du nombre de mots-outils**	
Details	**Segmentation (en "syllabes") effectuée sous Praat**	

Pointing L/R	**Numéro qui relie le geste à un Utterance-Ref**	
	_P	Préparation longue et évidente du geste
	_TR	Réajustement de la trajectoire du geste
	_AR	Réajustement de l'apex du geste
VisualTargetNoL/R	**Emplacement de la cible visée**	
PointingTypeL/R	**Classification du pointage**	
	Rest to apex	Pointage qui part d'une position de repos sur la table
	Rest in the air to apex	Pointage qui part d'une position de repos en l'air
	Apex to apex	Pointage (traj droite) départ : cible précédente
	Apex to apex with curve	Pointage (traj courbe) départ : cible précédente
	Error	Geste contenant une erreur (geste "pas propre")
	Other	Autres gestes ne rentrant pas dans les catégories précédentes

Gaze	**Direction du regard du participant**	
	Board	regard vers le plateau de jeu
	Screen	regard vers l'écran
	Interlocutor	regard vers l'interlocuteur
	Undefined	regard indeterminé (reflets de lunettes, ...)
	Other	regard vers quelque part ailleurs

Interlocutor	**Numéro qui relie le geste à Utterance-ref**	
	_P	Pose de la carte
	_S	Montré de la carte

FIGURE 5.5 – Liste complète des catégories de données annotées sur la vidéo

Annotation des unités d'intérêt

- Les unités d'intérêt considérées sont décrites dans la catégorie `Units-Type`. Globalement, le niveau de granularité utilisé dépend de ce qu'on cherche à étudier sur les données. Les mots-cibles ainsi que les démonstratifs sont les deux catégories d'intérêt principal. Tous les autres éléments sont annotés par "groupe" de mots selon leur fonction dans la phrase (leur type est noté `Other-`), sauf pour les instants d'hésitation/les pauses/les éclats de rire qui sont repérés individuellement.
- Dans une condition libre, il est possible que les démonstratifs ne désignent pas un emplacement mais plutôt un objet (typiquement, en désignant un objet déjà posé sur le plateau de jeu), et il est possible que ceci influe sur la coordination entre la prononciation du démonstratif et l'exécution du geste associé. La catégorie `Units-DemonstrativeFunction` permet de distinguer ces deux "fonctions" pour l'article démonstratif.
- Les catégories `Units-NbContentWords` et `Units-NbFunctionWords` permettent le décompte des mots utilisés au sein des différentes unités d'intérêt. Ce décompte est réalisé en séparant les mots de contenu et les mots grammaticaux. Ceci est utile pour caractériser le débit de parole ou la promptitude des propos selon les conditions.
- La catégorie `Détails` est simplement une recopie de la segmentation en syllabes réalisée sur le signal acoustique (et synchronisée avec la vidéo – dans l'expérience présentée en Chapitre 4, la vidéo a été post-synchronisée manuellement avec le signal acoustique).

Annotation des gestes de pointage Les gestes de pointage sont annotés grossièrement de façon à aider l'annotation automatique des données de suivi de mouvement : les annotations vidéo ne sont qu'un outil pour l'annotation finale. Les annotations réalisées débutent lorsque le geste de pointage commence et se terminent à l'instant de fin d'apex du geste. Ce type d'annotation a été choisi afin de n'avoir aucun recouvrement entre les annotations de deux gestes successifs (ce qui aurait été le cas si le geste complet avait été annoté) : lorsque deux gestes s'enchaînent, l'instant de retour du premier geste correspond exactement à l'instant de départ du second geste.

Les gestes sont annotés par des catégories différentes pour la main droite et la main gauche.

- `PointingL` et `PointingR` permettent de repérer les instants où un geste a lieu (resp. sur la main gauche et la main droite). Les annotations portent le numéro de l'énoncé auquel elles sont associées. Ceci permet dans les analyses ultérieures de lier un geste avec l'énoncé lui correspondant.
- `VisualTargetNo` est un numéro représentant l'emplacement visé par le geste de pointage sur le plateau de jeu (à chaque position sur laquelle est posable un objet est affecté un identifiant). Ceci permet d'étudier les stratégies de remplissage (si le remplissage de la grille se fait en suivant un parcours similaire d'un essai à l'autre par exemple).
- `PointingTypeL` et `PointingTypeR` sont deux catégories classifiant les gestes de pointage dans le but de faciliter l'annotation automatique du suivi de mouvement de ceux-ci. Les différences qui sont importantes à renseigner afin d'obtenir une bonne annotation automatique par la suite sont le fait de savoir si les pointages sont enchaînés ou non et si la trajectoire est courbée ou non.

Annotation du regard La direction du regard est annotée sur les vidéos (catégorie `Gaze`). Les trois directions principales qui sont observées par le participant sont la zone de jeu, l'écran sur lequel est diffusé le modèle et l'interlocuteur. Les annotations commencent lorsque le regard commence à quitter la direction précédemment occupée.

Annotation des mouvements du complice Les instants où le complice dépose une carte sur le plateau de jeu sont annotés. Les annotations portent les numéros des énoncés auxquels elles sont associées. L'annotation débute lorsque la carte touche le plateau et se termine lorsque la main du complice lâche la carte. Le complice n'a utilisé qu'une seule main pour poser les cartes au travers des expériences, ainsi une seule catégorie est nécessaire. *A priori*, dans le cas général, il pourrait être utile d'ajouter une catégorie pour la main gauche...

5.4 Conclusion sur la mise au point de protocoles et l'annotation multimodale

Ce chapitre "technique" a permis de mettre en avant deux points importants.

La première partie a montré la difficulté rencontrée pour la mise en place d'expériences peu contrôlées en laboratoire (en utilisant les systèmes de mesure disponible en laboratoire). Les principales difficultés rencontrées se trouvent à la fois au niveau technique (diffusion de bruit) et au niveau du *design* de l'expérience. Du point de vue technique, hormis les problèmes de masquage rencontrés, un point important à mentionner est la difficulté de diffuser du bruit acoustique afin de perturber une interaction si la voix est une variable d'intérêt dans l'expérience. En effet, les méthodes de débruitage classiques permettant une diffusion assez naturelle du bruit perturbateur (diffusion par haut-parleurs) n'ont pas donné satisfaction, puisque le débruitage ne s'est pas montré suffisant afin de pouvoir étudier la parole des interlocuteurs acoustiquement. Cette méthode semble ainsi bien fonctionner dans des cas proches des études réalisées initialement *i.e.* avec un environnement fixe et un locuteur "le plus immobile possible". La solution de remplacement utilise des casques identiques diffusant un bruit identique ce qui permet d'enregistrer un signal audio exempt de bruit mais modifie l'apparence des participants. Au niveau du *design* expérimental, la difficulté rencontrée a principalement consisté en l'élicitation naturelle de gestes de pointage. Plusieurs essais ont dû être menés afin de trouver une tâche qui ne soit pas trop artificielle et qui élicite "involontairement" des pointages. Finalement, pour éliciter des pointages (dans le cadre d'une tâche consistant à donner des consignes à un interlocuteur et dans laquelle l'interlocuteur doit interagir avec l'endroit désigné), il semble important que les deux interlocuteurs soient assez proches et que la partie de l'espace désignée ne soit pas trop éloignée des interlocuteurs. Enfin, si la désignation se fait sur une partition d'un plan, cette partition doit être assez complexe (ou ne contenir que peu d'indices visuels) pour éviter que les participants indiquent vocalement les emplacements.

La seconde partie s'est intéressée à la mise au point d'une grille d'annotation vidéo afin de décrire qualitativement l'interaction de deux personnes dans une tâche collaborative. Un point important est le fait que cette annotation vidéo constitue le lien nécessaire entre les enregistrements audio et spatiaux. La grille mise au point permet d'annoter les productions vocales (qualification des énoncés –correction, nombre de cibles, "complexité", interruptions–, des "unités" d'intérêt –type d'unité, nombre de mots grammaticaux/de contenu), les gestes des deux mains (type de trajectoire), une estimation de la direction du regard du participant et la pose des cartes pour le complice. Cette grille constitue une base nécessaire pour l'annotation d'interactions non contraintes et permet l'extraction d'informations quantitatives sur l'interaction. Comme mentionné précédemment, cette grille est fonctionnelle puisqu'elle a permis d'annoter plusieurs interactions non contraintes correctement (phase de validation de la grille) et a été utile pour l'annotation de l'étude présentée au Chapitre 4. Finalement, cette grille peut servir de base pour toute recherche étudiant des problématiques similaires à celles exposées dans ce manuscrit.

Chapitre 6

Discussion et conclusion

6.1 Synthèse des résultats et apports de l'étude

Les différentes études présentées dans ce manuscrit s'inscrivent dans une littérature relativement récente traitant des relations qu'il existe entre les différentes modalités prenant part dans la communication entre humains. Comme mentionné en introduction, la modalité gestuelle semble être liée de façon originelle avec la modalité vocale chez l'homme. C'est ainsi que McNeill a mis au point une théorie liant fortement ces deux modalités et permettant d'expliquer de manière plus ou moins extensive le contenu de chaque modalité en relation avec l'idée sous-jacente.

Ainsi, le but principal des études rapportées dans ce manuscrit est de caractériser quantitativement les interactions gestes manuels/parole observées afin de tenter de "remonter à la source" de cette interaction et de comprendre les mécanismes internes menant à ces interactions. Cette section reprend les différents résultats et les points importants mis en avant tout au long du manuscrit.

6.1.1 Une alchimie difficile entre précision et naturel des expériences

Comme mentionné en introduction, un point important à prendre en compte dans les expériences menées (à la fois en production et en perception) est le caractère naturel/écologique de celles-ci. Les études mentionnées en introduction peuvent globalement se scinder en deux catégories : les études écologiques (enregistrement vidéo d'une situation naturelle d'interaction entre plusieurs personnes) et les études de laboratoire (utilisation de corpora très contrôlés, éventuellement, utilisation de systèmes de suivi de mouvement). Le problème soulevé par ces deux types d'études est que chacun présente des avantages et des inconvénients pour l'étude de la coordination des gestes manuels avec la parole dans l'interaction humaine. L'utilisation de la vidéo seule pour acquérir et annoter les données ne permet pas une annotation automatique des données (ou plutôt, cela n'est pas fait dans les expériences de la littérature) et "distord" les données qui sont nécessairement bidimensionnelles (alors que la situation d'interaction se déroule bel et bien dans un espace tridimensionnel !). Cependant, l'utilisation de la vidéo, très peu invasive/intrusive, permet assez aisément d'utiliser des protocoles expérimentaux simples se rapprochant de la réalité de l'interaction naturelle. Inversement, l'utilisation de dispositifs de capture de mouvement permet l'acquisition de données (souvent tridimensionnelles) des mouvements qui permet une annotation objective, sur des signaux représentant totalement les trajectoires d'intérêt. Cependant, les études réalisées dans la littérature utilisant la capture de mouvement (tridimensionnelle) se sont, pour une écrasante majorité, limitées à des productions vocales peu naturelles (mots, groupes de mots ou non-mots en isolation).

Un des points importants des études présentées dans ce manuscrit est la volonté de combiner ces deux modes d'acquisition de signaux afin de tirer le meilleur parti des avantages de ces deux modes opératoires tout en essayant de restreindre leurs limitations. Ainsi, dans toutes les études présentées, il a été fait le choix d'utiliser un dispositif de capture de mouvements tridimensionnel précis (suivi de diodes infrarouges, de type Optotrak). Un signal vidéo a également été acquis mais n'a pas été utilisé dans l'annotation effective des gestes manuels/articulatoires. Le signal vidéo sert globalement à décrire l'interaction à la fois de façon qualitative et quantitative (mais

ne sert pas pour les mesures de grandeurs physiques).

Outre le fait que l'annotation des données en utilisant des données tridimensionnelles soit plus "objective" que de regarder la vidéo et l'annoter, l'annotation des trajectoires issues de capteurs tridimensionnels permet de spécifier (et vérifier) des "seuils" de position/vitesse pour scinder les mouvements en différentes phases. De façon importante, cette façon de procéder permet également une annotation semi-automatique des données (annotation automatique puis vérification manuelle rapide *a posteriori*) ce qui rend le travail beaucoup plus rapide.

Enfin, un autre apport de l'étude est l'utilisation de phrases complètes, grammaticalement correctes dans les productions. Les productions sont plus ou moins contrôlées, mais dans tous les cas, les productions sont "naturelles" et ne se limitent pas à un (groupe de) mots ou non-mots en isolation.

6.1.2 Les interactions gestes manuels/parole : un phénomène complexe

De façon consensuelle dans la littérature, il est admis qu'il est possible de dresser une typologie des gestes (suite aux travaux de Kendon et McNeill, la typologie présentée en Section 1.2.1 semble rencontrer un accord quasiment unanime). La classification proposée par McNeill ordonne les gestes selon leur lien entretenu avec la parole. Par ailleurs, McNeill (et une multitude d'études prônent un fonctionnement du même type) défend des liens internes complexes entre réalisation de la parole et réalisation des gestes manuels cooccurrents avec la parole. Ainsi, il semble assez naturel de penser que, si différents gestes ont différentes relations avec la parole, et que leurs réalisations dépendent de ces relations, alors chaque "type" (voire, sous-type) de geste est réalisé différemment en relation avec la parole lorsqu'il accompagne celle-ci.

Par conséquent, l'étude et la qualification extensive des interactions gestes manuels/parole semble difficilement réalisable. Les études présentées dans ce manuscrit ont confirmé en partie ces hypothèses en montrant (dans les deux premières expériences présentées) que les réalisations temporelles des gestes par rapport à la parole étaient différentes pour différents types de gestes dans un même contexte de parole (gestes communicatifs déictiques *vs* gestes communicatifs non déictiques *vs* gestes non communicatifs dans un contexte de désignation utilisant la focalisation prosodique en parole). L'étude des différences entre types de gestes s'est réalisée en contrôlant assez fortement la modalité parole : des modèles de phrases (correctes grammaticalement) sont utilisés, au sein desquels la structure reste inchangée parmi les essais, seuls deux mots (le sujet ou le complément) peuvent varier et les productions sont totalement guidées par les stimuli présentés.

La décision a été prise dans la troisième expérience présentée de caractériser plus précisément les relations temporelles des gestes de type déictique avec la parole. Cette contrainte (qui n'était pas ouvertement donnée aux participants mais qui a guidé la création du protocole) sur la modalité gestuelle a permis de libérer des contraintes sur la modalité parole. En effet, l'interaction réalisée dans la troisième expérience présentée ne contient aucune contrainte (hormis l'utilisation d'un modèle de phrases lors du placement des cartes et la nécessité de rester assis) sur le déroulement de l'expérience, ainsi les interactions sont assez naturelles (aucun ordre n'est imposé dans les placements, aucune production de gestes n'est imposée, aucun "type" de geste n'est imposé etc…).

Les résultats trouvés à la fois lorsqu'une contrainte est fixée sur les deux modalités ainsi que lorsque la contrainte est levée sur les gestes manuels (et dans une condition réaliste d'interaction) sont des prémisses importants à la réalisation d'une tâche totalement non contrainte. En effet, le relâchement progressif des contraintes imposées permet de valider le processus d'acquisition des données, d'annotation des données et de s'assurer de la faisabilité d'une étude totalement naturelle s'appuyant sur un protocole similaire.

Un autre point important présenté dans le chapitre précédent est le fait que le *design* d'expériences permettant l'élicitation "naturelle" d'un type de gestes n'est pas évident : comme mentionné dans la littérature et comme observable au quotidien, les gestes produits en cooccurrence avec la parole sont idiosyncrasiques et rendent donc la mise au point de protocoles "naturels" complexe si le but de ces protocoles est l'élicitation de gestes face à une situation donnée.

6.1.3 Annotation conjointe d'interactions multimodales

Dans la littérature, les interactions multimodales ont majoritairement été –lorsqu'elles l'ont été– annotées uniquement sur les signaux vidéo et audio (souvent extrait de l'enregistrement vidéo). L'utilisation de l'annota-

tion vidéo présente plusieurs points positifs qui sont liés au fait que la vidéo contient à la fois des informations sur les gestes mais également sur la parole et, plus généralement, sur la situation environnante, les émotions véhiculées par les expressions faciales, la posture, etc... Cependant, comme mentionné plus haut, l'annotation vidéo souffre de quelques limitations qu'il semble important de relever. Au niveau de la parole, l'acquisition doit nécessairement être réalisée par un micro d'excellente qualité et positionné proche du locuteur si une analyse acoustique est voulue par la suite, ce qui n'est pas faisable avec des caméras classiques. De manière plus importune, l'annotation des gestes n'est pas réalisable précisément sur un signal vidéo bidimensionnel : les articulations permettent aux bras/main/doigts de bouger dans les trois dimensions de l'espace, le mouvement est donc par essence tridimensionnel.

Le paradigme utilisé dans les expériences du manuscrit et généralisé dans le Chapitre 5 a permis de présenter une méthode d'annotation systématique pour les données multimodales mêlant gestes manuels et parole. Selon cette méthode, chaque modalité est annotée grâce à des signaux acquis représentant correctement les productions. Le signal acoustique est acquis par l'intermédiaire d'un microphone placé à proximité du participant ce qui permet une bonne qualité d'enregistrement de la voix. Les trajectoires spatiales d'intérêt sont suivies *tridimensionnellement* par un système infrarouge, ce qui permet d'avoir accès à la trajectoire complète (et non simplement à une projection dans un plan bidimensionnel). Enfin, un enregistrement vidéo est effectué mais cet enregistrement *ne sert pas* à l'annotation des deux modalités vocale et manuelle : il est utilisé dans le but de caractériser qualitativement l'interaction et de manière fondamentale, il sert à lier les enregistrements audio et spatial.

Au cours du manuscrit, il a été présenté la méthode d'annotation originale mise en place. Dans les deux premières expériences, la vidéo n'est pas utilisée pour l'annotation, dans la troisième expérience, celle-ci permet l'annotation qualitative de l'interaction et le guidage de l'annotation automatique des gestes manuels. Basiquement, l'annotation se déroule alors en plusieurs étapes :

1. Segmentation audio et extraction des paramètres acoustiques d'intérêt : cette phase permet de scinder le signal audio en syllabes. Idéalement, la segmentation se fait de façon semi-automatique (alignement automatique en utilisant la théorie des chaînes de Markov cachées) avec un ajustement manuel *a fortiori*. Il est possible d'annoter les variables quantitatives liées uniquement à la parole directement à ce niveau.
2. Annotation vidéo : cette phase permet non seulement de décrire qualitativement l'interaction, mais également de repérer grossièrement l'emplacement des gestes (et de qualifier leur "forme" afin de choisir parmi différents patrons d'annotation automatique par la suite) pour guider l'annotation automatique des gestes, cette annotation globale permet de repérer les relations entre les segments de parole et les gestes associés (qui est un problème impossible à traiter sans le signal vidéo).
3. Annotation des gestes et des indices prosodiques : cette étape permet d'annoter les différents signaux (acoustique, articulatoire, manuel) en prenant en compte les annotations faites dans les deux étapes précédentes. Toute l'annotation est semi-automatique (automatique puis vérifiée –et corrigée le cas échéant– manuellement).

Enfin, une grille d'annotation pour la vidéo a été mise en place et est facilement utilisable pour une expérience qui utiliserait le même protocole sans contrainte au niveau de la parole –et bien entendu, facilement adaptable à un autre travail s'intéressant à des questions proches de celles évoquées auparavant.

Par delà les apports relatifs à la mise en place des expériences, le manuscrit apporte quelques points intéressants pour les différentes expériences menées. Ces résultats, déjà résumés dans les différents chapitres sont rapportés ci-dessous dans un souci de synthèse.

6.1.4 La coordination geste/parole dans des conditions contraintes

L'expérience présentée au Chapitre 2 a étudié la production conjointe de gestes (de différents types) et de parole dans le cadre de la désignation (focalisation contrastive prosodique au niveau de la parole). Il a été montré un effet principal de la position de la focalisation au sein de l'énoncé : plus la focalisation est produite tard, plus le geste est également produit tard *i.e.* la focalisation attire le geste manuel. Cet effet est valable pour tous les types de gestes étudiés (battement, pointage, appui sur un bouton) mais est particulièrement saillant pour le geste de pointage.

Par ailleurs, les gestes ne sont pas toujours cooccurrents avec la partie focalisée de la parole. En particulier, le geste d'appui sur un bouton ne semble pas coordonné avec la focalisation sauf lorsque celle-ci se trouve en début de production vocale. De même, l'exécution du geste de battement, bien que plus influencée par la position de la focalisation que l'exécution du geste d'appui sur le bouton, n'intervient majoritairement en cooccurrence avec la focalisation que lorsque celle-ci se trouve en fin d'énoncé. Ces comportements sont différents de ce qui est observé pour le geste de pointage dont l'apex intervient majoritairement au sein de la focalisation prosodique et ce, dans les deux positions de focalisation.

Lorsque les gestes sont produits de façon cooccurrente avec la parole, on observe assez naturellement des "alignements" temporels entre les instants importants de leur réalisation et la réalisation de la parole. En particulier, ces alignements sont assez fins en ce qui concerne le geste de pointage pour lequel l'apex est généralement produit temporellement très proche des cibles articulatoires des mots focalisés (seconde cible articulatoire lorsque la focalisation est en début d'énoncé, première cible lorsque la focalisation est en fin d'énoncé). On observe également des alignements temporels (moins fins et moins réguliers) avec les pics de fréquence fondamentale et d'intensité de la partie focalisée. Pour le geste de battement, on observe aussi un tel alignement mais seulement lorsque la focalisation est en fin d'énoncé, de même pour le geste d'appui sur un bouton lorsque la focalisation est en début d'énoncé.

Ces résultats ont été interprétés comme reflétant le fonctionnement de deux mécanismes modifiant les productions respectives des deux modalités : des contraintes motrices se couplent à des contraintes communicatives. Les contraintes communicatives étant plus fortes pour le gestes de pointage (celui-ci a un rôle très proche de la focalisation dans la tâche de désignation), ceci peut expliquer les liens forts qui existent au niveau temporel.

Enfin l'étude a montré que les adaptations mutuelles entre les deux modalités étaient de rigueur la plupart du temps. En particulier, la production des gestes s'adapte à la production de parole (les gestes ne sont pas produits aux mêmes instants selon la position de la focalisation en parole) et la production de la parole s'adapte à celle des gestes (la production de gestes de pointage ralentit le début de production de parole par rapport à une condition sans production de geste).

6.1.5 Importance du lien communicatif geste manuel/parole

La seconde expérience présentée en Chapitre 3 s'est appuyée sur un protocole totalement similaire à la première expérience. Une simple modification a été effectuée : le lien (appelé lien communicatif) entre ce qui est désigné par le geste de pointage et ce qui est désigné par la focalisation a été changé. Ainsi, selon les prédictions, ce changement ne devrait affecter que la production conjointe de parole et de gestes de pointage. Il s'est avéré que ce simple changement a eu un impact assez important sur les productions des participants.

En effet, il a été montré la division des participants en deux sous-groupes utilisant des stratégies coordinatives dans l'exécution de la parole et des gestes manuels différentes (les deux stratégies diffèrent en particulier sur la coordination des exécutions du geste de pointage et de la focalisation). Globalement, pour les gestes de battement et d'appui sur un bouton, la modification de la tâche n'implique pas de changements majeurs dans leur production relativement à la production de la parole. Comme attendu, ceci n'est pas le cas pour le geste de pointage. Une (faible) majorité des participants change sa stratégie de production pour les gestes de pointage pour lesquels la production fait intervenir en parallèle l'apex du geste et l'objet désigné en début d'énoncé (quelle que soit la position de la focalisation). Un deuxième groupe réalise des productions similaires à l'expérience précédente en coordonnant l'apex des gestes avec la partie focalisée de l'énoncé. Ainsi, le lien communicatif entre le geste et la parole a une influence sur la coordination geste manuel/parole et cette influence dépend des participants : il y a beaucoup de variation interparticipants.

Cette expérience a permis de mettre au jour des mécanismes fins et complexes qui entrent en jeu dans la coordination geste manuel/parole. Par ailleurs, il a été montré, tout comme dans l'expérience précédente, la non-influence de la production de gestes manuels sur l'organisation temporelle de la parole associée et aucune modification des amplitudes des évènements manuels/articulatoires mesurés dans les différentes conditions. Enfin, un point intéressant montré par cette étude porte sur l'instant de début de production de la parole : en comparant les deux premières expériences, il est possible de tirer des conclusions sur la façon dont se font les planifica-

tions des gestes et de la parole, et en particulier, les données ici présentées vont dans le sens d'une planification en parallèle. L'instant du début des productions semble par ailleurs être "calculé" en fonction de l'avancement de la planification dans chacune des modalités afin de respecter des contraintes communicatives ("synchronie" d'évènements "saillants" dans les deux modalités lors de la production).

6.1.6 La coordination geste/parole lors d'une interaction contrôlée et l'influence d'une perturbation de l'interaction

Une troisième expérience a permis de voir ce qu'il advenait des résultats précédents dans une condition d'interaction plus naturelle. Cette expérience s'est concentrée sur le geste de pointage dans le cadre d'une tâche de désignation. Ici la désignation en parole n'était plus réalisée par la focalisation prosodique mais par l'usage de démonstratifs. Outre les questions de coordination similaires aux expériences précédentes, l'influence d'une perturbation de la situation d'interaction sur cette coordination a été étudiée.

De manière similaire à ce qui est observé dans les deux expériences antérieures, il a été montré dans cette expérience que l'occurrence des gestes de pointage semble être assez largement coordonnée avec l'occurrence en parole du mot correspondant à la carte à placer sur l'emplacement désigné. D'après cette expérience, comme dans la précédente, il semble que l'apex du geste de pointage se synchronise temporellement avec l'objet désigné en parole plutôt qu'avec le processus démonstratif en parole (focalisation ou démonstratif). Ceci a permis d'attribuer au geste de pointage un double rôle de désignation (à la fois temporelle et visuelle/spatiale) : le geste de pointage "montre" un emplacement dans l'espace et il signale à l'interlocuteur l'endroit de l'énoncé où une attention maximale est requise pour comprendre parfaitement le message.

Par ailleurs, la perturbation de l'interaction par la diffusion de bruit dans l'environnement affecte la production de parole comme cela est mentionné dans la plupart des études sur l'effet Lombard (augmentation d'intensité, de fréquence fondamentale, du nombre d'erreurs, abaissement du débit, . . .). Au niveau des gestes, les productions sont modifiées mais semblent suivre les modifications de la parole (au moins, au niveau temporel) : il semble que l'exécution (ou la planification. . .) des gestes manuels s'adapte à l'exécution de la parole. De manière intéressante et en accord avec la littérature, il a été montré une invariance de la durée du "lancer" du geste de pointage et une adaptation de la durée du geste à la durée de la parole localisée sur la durée de la tenue du geste de pointage. Le geste ne subit pas un ralentissement général comme la parole, mais quelques phases (pour le geste de pointage, la tenue du geste) sont modifiées pour s'adapter au ralentissement de la parole.

6.2 Axes de discussion et ouverture

6.2.1 Les modèles de production conjoints geste/parole

L'introduction de ce manuscrit a présenté les quelques modèles influents de la littérature traitant de la production conjointe de gestes manuels et de parole. Les trois expériences menées au cours de la thèse et présentées précédemment permettent une approche novatrice de ces modèles par le biais de la deixis multimodale. En effet, la plupart de ces modèles ont été conçus (et testés) pour les gestes iconiques, or ces modèles prétendent à l'explication de la coordination des deux modalités pour les "gesticulations" en général (sauf le battement, pour le modèle de de Ruiter). Dans nos deux premières expériences, le geste de contrôle n'est pas concerné par ces modèles (le geste n'est pas une gesticulation) et il a été montré que le geste de battement élicité ne semblait pas être un geste de battement "naturel" (il est alors plus prudent d'écarter le cas des battements obtenus dans les deux premières expériences dans cette discussion).

Globalement, les résultats trouvés dans toutes les expériences rapportées ci-dessus sont en accord avec l'adage bien connus de la littérature (de Kendon et McNeill) : la règle de synchronie phonologique, selon laquelle le *stroke* (ou apex) d'une gesticulation précède ou est synchrone avec le "pic phonologique" de l'énoncé en parallèle du geste. Cette règle est assez largement respectée par le geste de pointage dans les données (dans les trois expériences). Un point important à rapporter est le fait que bien que cette règle soit *majoritairement* respectée, elle ne l'est pas dans tous les cas *i.e.* la règle est plus souple qu'initialement formulé par Kendon.

Par ailleurs, il a été montré assez clairement au cours des trois expériences qu'une coordination existait entre la réalisation des gestes manuels (en particulier, l'apex et plus généralement, le *stroke*) et l'articulation des syllabes des mots affiliés aux gestes. Afin de réaliser une telle coordination, il est possible qu'au sein des modèles, la "branche du modèle" en charge de la mise au point du programme moteur des gestes manuels soit en lien avec la "branche du modèle" responsable de la mise au point du plan articulatoire (l'encodeur phonologique chez Levelt [115]). Bien entendu, la relation n'est pas forcément directe (comme cela est le cas dans le modèle d'interfaces de Kita & Ozyürek), mais elle est nécessaire pour pouvoir prendre en compte cette contrainte motrice. Une telle connexion suppose que le plan articulatoire soit au point pour adapter le timing du geste l'accompagnant et impose donc une interaction tardive entre les timings des deux modalités. Une autre façon possible pour expliquer cette adaptation peut venir d'une attraction motrice qui aurait lieu lors de l'exécution : les mouvements demandant le plus d'effort dans chaque modalité jouent le rôle d'attracteurs et s'attirent mutuellement donnant naturellement une cooccurrence de ces évènements.

Il a été montré tout au long du manuscrit une adaptation bidirectionnelle à la fois de la parole aux gestes (la production d'un geste de pointage ralentit le début de la parole dans la première expérience) mais aussi des gestes à la parole (dans la troisième expérience, par exemple, la modification de la production des gestes suit exactement la modification de la production de la parole afin de garder des alignements similaires). Différentes solutions sont envisageables pour relier les deux "branches" du modèle (quel qu'il soit) : un lien soit unidirectionnel soit bidirectionnel (parole $\rightarrow$ geste, geste $\rightarrow$ parole, parole $\leftrightarrow$ geste), ou une attraction par les instants de plus grand effort moteur. La solution la plus simple est un lien bidirectionnel à cause de la double adaptation. Il est possible de penser que le lien fasse transiter entre les deux systèmes des informations complexes de timing des productions afin d'aboutir à une synchronisation des productions. Cependant, ceci semble être exagéré à la vue des résultats présentés dans ce manuscrit. En effet, il a été montré que ① le lancer du geste de pointage semble avoir une durée invariante dans différentes conditions de pointage (et la durée du plateau du geste est allongée lorsque la parole est ralentie) ② l'apex du geste de pointage intervient sur une cible articulatoire de l'affilié lexical ③ la focalisation attire le geste ④ la position de l'affilié lexical dans l'énoncé influe sur la position de l'apex au sein de l'affilié lexical ⑤ dans la seconde expérience, deux groupes de participants sont trouvés.

Une amélioration possible des modèles temporels : les systèmes dynamiques Si, comme le suggèrent les études de la littérature, on considère que le temps mis pour planifier le geste est négligeable en comparaison du temps mis pour planifier la parole (le premier processus étant supposé beaucoup plus simple que le second), il est possible que lors de la production d'une proposition multimodale, le geste soit d'abord planifié puis interagisse avec la production de parole au niveau du conceptualiseur (comme le défend de Ruiter). Ainsi le geste est planifié et la production de parole se fait en relation avec le geste planifié (dont la phase de préparation commence dès que le geste est planifié). Cependant cette solution n'est pas totalement satisfaisante à la vue des données présentées ci-dessus puisqu'il a été montré que le geste s'adaptait également à la parole. Une solution évoquée consiste en un décours temporel comme il suit : le geste est planifié (et exécuté directement), la planification phonétique est alors lancée puis la parole articulée, les parties de la parole demandant le plus grand effort (*i.e.* les parties focalisées de la parole) attirent la partie du geste demandant le plus grand effort (*i.e.* le stroke, l'apex du geste).

Cette façon de fonctionner pourrait expliquer en partie les différents résultats mis en avant précédemment : le conceptualiseur "choisit" le terme sur lequel le geste intervient en parallèle (et ce choix est idiosyncrasique, tel que démontré dans la seconde expérience). Lorsque le geste est prêt, celui-ci démarre et la parole est planifiée, le temps de planification de la parole et son exécution laissent au geste le temps d'arriver à l'instant de stroke au moment voulu (temps proche de l'affilié lexical choisi par le conceptualiseur). L'instant précis de production du stroke (plus grand effort moteur) du geste manuel se situe à l'instant où l'effort articulatoire est le plus grand *i.e.* sur un *stroke* articulatoire de la parole, avant l'atteinte d'une cible. *A priori*, ainsi, si le temps disponible pour exécuter la préparation du geste est suffisant avant que la parole soit prête à être articulée, l'attirance entre stroke et cible articulatoire est plus grande pour les cibles articulatoires en début de mot, si la parole est plus longue à planifier, l'apex du geste a plus de chance d'être attiré par les cibles articulatoires en fin de mot. Ceci pourrait expliquer les différences de coordination observées dans les deux conditions de focalisation dans la première expérience (synchronisation de l'apex avec la seconde cible articulatoire en condition *Foc1* et avec la première

cible en condition *Foc2*).

Deux résultats ne sont cependant *a priori* pas expliqués par ce mode de fonctionnement : l'adaptation de l'instant de début de geste dans le cas de la situation bruitée et le fait que la synchronisation entre l'apex du geste de pointage et les cibles articulatoires ne se fasse pas sur les mêmes cibles articulatoires dans les trois expériences. Le premier point peut être expliqué par un retard des productions induit par un retard dans le conceptualiseur : la perturbation de la situation de communication rend la répartition des concepts dans les différentes modalités plus complexe mais n'influe pas sur la génération précise de chaque mouvement (articulatoire/manuel). Le second point est plus complexe à expliquer : les résultats des deux premières expériences sont en accord avec ce qui est évoqué ci-dessus, par contre, dans l'expérience 3, on observe un alignement avec la première cible articulatoire et l'apex du geste de pointage alors que le début de phrase est *a priori* aussi long à planifier dans les deux expériences (*le ballon est vert vs le bonbon va là*). Cette différence dans les alignements peut-être expliquée par la consigne qui imposait une réponse rapide dans les deux premières expériences, mais pas dans la troisième, ainsi dans la troisième expérience, le stroke du geste arrive au début de l'affilié lexical et est donc attiré par la première cible articulatoire.

L'affilié lexical Par ailleurs, un résultat important évoqué ci-dessus par l'intermédiaire des deux types de coordination dans la seconde expérience est le fait que le choix de l'affilié lexical est idiosyncrasique. Bien que la plupart des modèles s'appuient sur des données simples où l'affilié lexical est évident et non ambigu, ici dans la seconde expérience, l'affilié lexical n'est pas évident et deux stratégies de coordination sont mises en place par différents participants.

Dans le même ordre d'idée, contrairement à ce à quoi on pourrait s'attendre, dans la troisième expérience, le geste de pointage n'est pas coordonné avec l'articulation du démonstratif mais plutôt avec celle de l'objet à placer. Ces résultats sont à rapprocher de ceux de Levelt et coll. [114] dans lesquels, lors d'une tâche de désignation, des énoncés du type "cette lampe" avaient conduit à une coordination avec le démonstratif "cette". Ainsi, selon la tâche, le geste de pointage ne semble pas être synchronisé avec le même affilié (celui-ci semble être surtout synchronisé avec le mot apportant le plus d'information dans l'énoncé plutôt qu'avec un mot correspondant à la même représentation/fonction).

L'environnement Un dernier point à prendre en compte est l'influence de l'environnement. La troisième expérience montre que l'environnement bruité change légèrement les productions temporelles des deux modalités. Plus précisément, il semble assez nécessaire de prendre en compte l'environnement dans les deux modalités. Il a été montré que le bruit modifiait principalement les évènements en parole et que, globalement, le geste "suivait" les modifications de la parole afin de produire des alignements temporels similaires en condition bruitée et en condition non bruitée. Ce "suivi" permet d'expliquer en particulier l'allongement du plateau du pointage. D'autres éléments ne peuvent pas être expliqués par cela, en particulier, la baisse de la variabilité des instants de début des gestes ne peut être expliquée que par une influence de l'environnement sur la "branche" du modèle responsable de la génération du geste (ou sur une partie du modèle commune aux deux modalités – possiblement, le conceptualiseur dans la majorité des modèles).

Une modification du modèle Sketch dans le cadre de la désignation Le modèle Sketch de de Ruiter ([38]) présenté en introduction est un des seuls modèles de productions conjointes parole/gestes manuels permettant de faire des prédictions sur les instants relatifs de production dans les deux modalités. Le modèle Sketch est reproduit en Figure 6.1. Comme mentionné en introduction, ce modèle permet de rendre compte selon son auteur, de la production des gestes coverbaux (sauf battement) et de la parole. Les données des différentes expériences menées permettent de compléter légèrement le modèle de de Ruiter dans le cadre de la désignation.

Dans le cadre des expériences menées et dans une perspective de modélisation de la fonction de désignation, les prédictions/hypothèses de Levelt et de Ruiter sont majoritairement corroborées par les données :

- Selon Levelt (et de Ruiter), le message préverbal (sortant du conceptualiseur) contient une marque permettant à l'encodeur phonologique de générer phonologiquement la focalisation sur un mot (si une telle

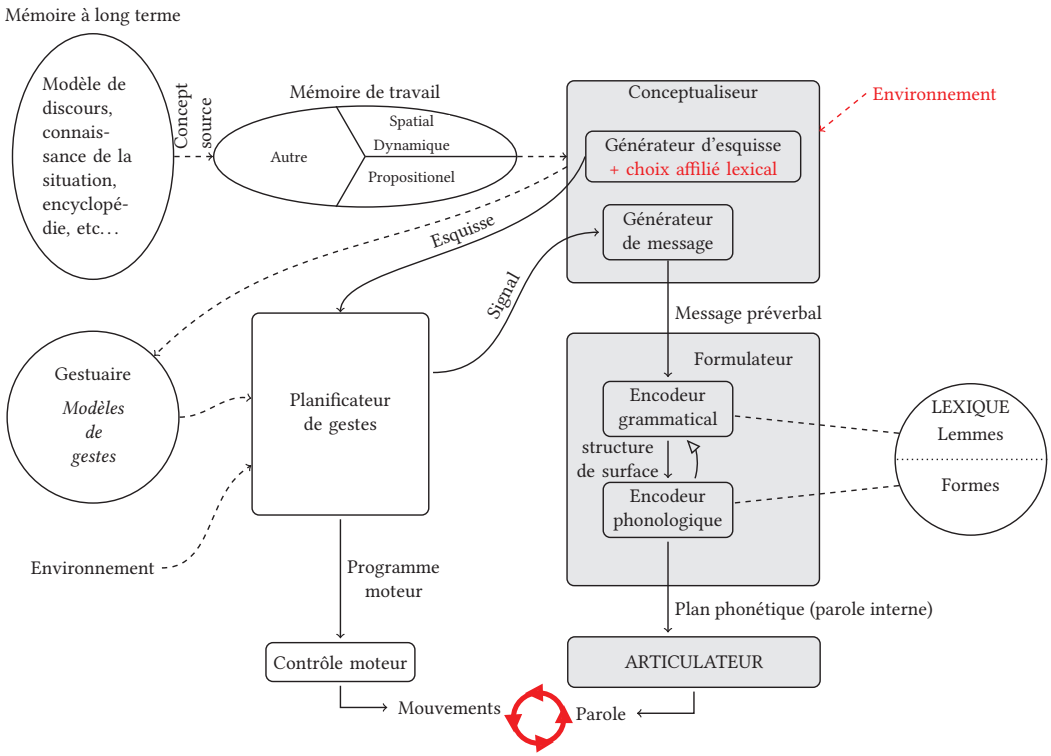


FIGURE 6.1 – Modèle Sketch et modifications proposées (en rouge)

focalisation est à réaliser). Ainsi, l'information indiquant quelle partie d'un groupe nominal (ou plus généralement, un énoncé) porte la focalisation est disponible dans l'étape du conceptualiseur. Ceci est intéressant dans le cas des premières expériences où il a été montré une attirance du geste par la focalisation : le conceptualiseur peut faire passer l'information de l'emplacement de la focalisation au planificateur de gestes, ainsi le modèle permet de rendre compte des données du manuscrit sur ce point.

- Le conceptualiseur a par ailleurs un rôle dans la durée du geste iconique, puisque c'est lui (selon de Ruiter) qui commande la tenue du geste, une fois que le message préverbal a été généré correctement –ceci est possible grâce au monitoring du modèle Blueprint de Levelt–. Ceci permet de s'assurer que le geste recouvre l'affilié lexical, résultat qu'on observe dans les données ci-dessus.
- De Ruiter montre une covariation entre les instants de la syllabe comportant le pic de fréquence fondamentale ainsi que l'exécution du geste : le geste s'adapterait à la parole en ceci que le *stroke* du geste est exécuté avec le pic de fréquence fondamentale. Cette adaptation n'est pas visible dans le modèle (puisqu'il n'y a aucun retour de la branche "parole" sur la branche "geste") et de Ruiter l'explique par un effet venant d'un système externe de synchronisation (par exemple, la respiration : l'expiration la plus forte pourrait correspondre au pic de fréquence fondamentale et être synchronisée avec le mouvement de la main). L'idée de systèmes dynamiques a également été abordée dans ce manuscrit et semble être une piste possible pour expliquer les détails de la coordination.
- Dans le cas des gestes de pointage, de Ruiter propose que le timing *prévu* des gestes doit être rendu disponible au conceptualiseur avant exécution. Selon le modèle Sketch final, le planificateur de geste fournit ces informations au conceptualiseur qui envoie par la suite le messsage préverbal au formulateur. Le formulateur peut alors planifier la parole.

Quelques points peuvent être apportés par les données du manuscrit afin de préciser le modèle dans le cadre de la désignation :

- Selon de Ruiter, dans le cas d'une désignation, c'est dans le conceptualiseur que se fait le choix du terme déictique à utiliser (par ex. ici *vs* là pour des objets proches *vs* lointains). Cependant, le modèle Sketch ne fait pas de prédiction sur le choix de l'affilié lexical au sein d'un énoncé. Dans les expériences menées ci-

dessus, les choix étaient *a priori* multiples (dans la seconde expérience, il est possible que l'affilié soit soit l'objet, soit son attribut, dans la troisième expérience, l'affilié peut être soit l'objet à placer soit le démonstratif – ou les deux), le choix de l'affilié doit se faire dans le conceptualiseur et ce choix est idiosyncrasique (*cf.* les résultats de la seconde expérience et la constitution de deux groupes).

- Dans le modèle Sketch, l'environnement a une influence directe sur le planificateur de gestes. Or il semble assez logique de déplacer cette interaction au niveau du conceptualiseur. En effet, il est probable que l'environnement dans lequel se déroule une interaction ait un effet sur la coordination gestes manuels/parole et en particulier sur le choix de l'affilié lexical. Par ailleurs, le conceptualiseur peut décider selon la situation d'interaction de représenter certaines choses sous forme gestuelle ou non (il est des cas –par exemple, un environnement bruité– où les gestes apportent une aide précieuse à la compréhension qui ne serait pas nécessaire sans perturbation). Enfin, l'environnement influence naturellement la production de parole (typiquement : effet Lombard) et doit donc apparaître au niveau conceptuel dans le modèle Sketch.
- Comme discuté plus haut (et comme avancé par de Ruiter), il est possible que les effecteurs de la parole et des gestes manuels soient couplés lors de la réalisation d'un message multimodal. Ainsi, à l'exécution, une dynamique se met en place et la réalisation d'action dans chaque modalité joue un rôle d'attracteur moteur. Dans les expériences menées, il a été montré majoritairement un alignement entre l'apex des gestes (de pointage) et les cibles articulatoires (maximum d'ouverture/protrusion) des affiliés lexicaux et l'alignement de l'apex ne se fait pas avec la même cible articulatoire selon la position de l'affilié dans l'énoncé.

Globalement, peu de modifications sont à apporter au modèle à la vue des résultats du manuscrit : le conceptualiseur est influencé par l'environnement, celui-ci sélectionne l'affilié lexical d'un geste et l'utilité de produire un geste ou non, et enfin les productions dans les deux modalités interagissent à un niveau moteur (attracteurs dynamiques). Une proposition d'ajustement du modèle est proposée en Figure 6.1, en rouge sont représentées les modifications principales.

6.2.2 Les agents conversationnels

La mise au point d'agents conversationnels/avatars est un axe de recherche qui s'est développé assez tôt. Les têtes parlantes ont graduellement pris en compte les mouvements de la face pour un meilleur rendu visuel puis certaines têtes parlantes ont été adaptées sur des structures de corps complètes afin de modéliser, par exemple, la posture, les gestes manuels etc... De manière assez évidente, le but premier de ces agents est de paraître le plus naturel possible, d'imiter le comportement humain de manière parfaite de façon à susciter l'interaction par son interlocuteur humain. L'imitation du comportement humain a pour but de rendre l'agent conversationnel expressif, outre la mise au point d'un système permettant la synthèse vocale, ceci passe par une animation de l'agent naturelle. À l'heure actuelle, la mise au point d'expressions faciales permettant de faire transparaître des émotions, la prise de postures et l'animation des membres supérieures sont les points à l'étude (voir [143] pour un état de l'art des travaux, voir Figure 6.2 pour quelques exemples d'agents). Ainsi, un point essentiel pour permettre à un agent conversationnel de paraître "naturel" est l'organisation temporelle des évènements et en particulier la synchronie des évènements acoustiques et visuels. Ceci est en particulier vrai pour les mouvements des articulateurs modélisés, mais les études présentées en introduction (*cf.* par exemple [68]) montrent également une importance capitale du timing respectif de la parole et des gestes manuels.

Comme montré en introduction, les gestes manuels font partie intégrante d'une situation d'interaction : une portion conséquente de la parole est accompagnée de gestes. Ainsi, l'intégration des gestes manuels dans les agents conversationnels est devenue durant les années 1990 un point important de la recherche dans ce domaine. Les travaux de Cassel et collègues [26] ont mis au point un cadre théorique novateur permettant la génération de gestes coverbaux (et en particulier, de leur timing dans le flux de parole) pour des avatars. Selon le modèle mis en place par cette équipe, la position temporelle des gestes au sein du discours peut se calculer grâce aux similarités entre les gestes manuels et l'intonation dans leur rôle de structuration de l'information. Suivant la citation de Bolinger selon laquelle l'intonation a "plus à voir" avec les gestes qu'avec la grammaire [1], l'équipe pense que l'intonation d'un énoncé permet la prédiction de l'instant précis où apparaît le geste dans l'énoncé

1. "Intonation belongs more with gesture than with grammar"(Bolinger, 1983)

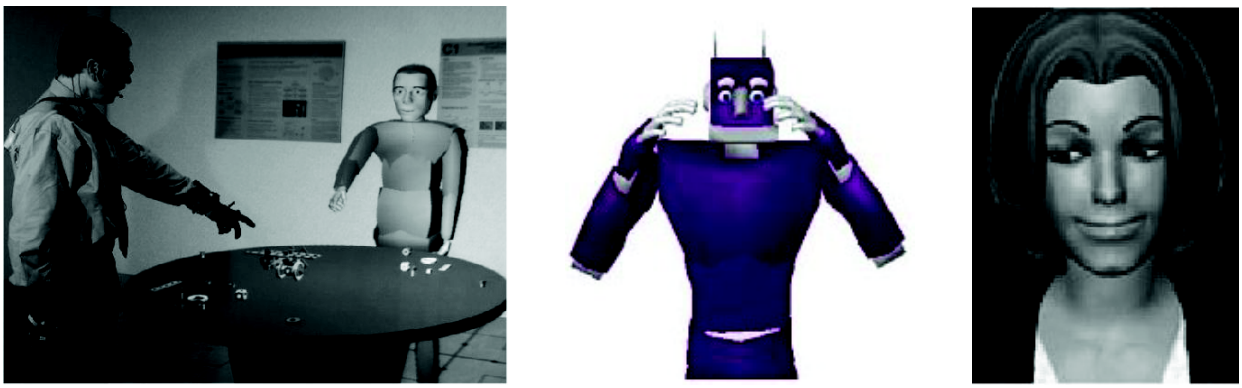


FIGURE 6.2 – Exemples d'Agents Conversationnels Animés (exemples de Bielefeld, Carnegie Mellon, ParisTech)

(et que la structure du discours permet de prédire sur quels énoncés un geste sera produit). Globalement, les gestes interviennent sur les rhèmes (*i.e.* les parties représentant l'information nouvelle) d'un discours et leur type est totalement défini par la parole qu'ils accompagnent. Leur occurrence temporelle est par ailleurs totalement définie par rapport à la prosodie de la parole : les gestes de battements coincident avec la syllabe accentuée de l'énoncé, les autres gestes sont réalisés en trois étapes ① début de la préparation avant l'articulation de l'énoncé ② le *stroke* est cooccurrent avec l'accent d'énoncé et ③ le retour du geste a lieu autour de la fin du syntagme intonatif.

Une approche plus récente, utilisée à Bielefeld par Kopp et Wachsmuth [102] sur l'avatar "Max" de l'institut Max Planck, reprend les travaux de Cassel en évoquant ses limites (peu de flexibilité, en particulier pour modifier l'exécution d'un geste –ce qui est dû au fait que l'exécution est totalement ballistique–, difficulté à enchaîner des gestes). Cette approche s'appuie sur une hypothèse selon laquelle chaque élément d'idée est exprimé par de la parole et un geste associé (possiblement, aucun geste). Ainsi, les auteurs définissent une "unité de production geste-parole" comme étant une paire constituée de la production d'un syntagme intonatif et d'un geste manuel. Selon le cadre mis en place, lorsqu'une idée est exprimée par les deux modalités, le *stroke* du mouvement manuel est cooccurrent avec le mot affilié au geste (*i.e.* le stroke commence avant l'affilié lexical et se termine parfois après). Plus précisément, si l'affilié lexical est focalisé, le *stroke* commence au moment où l'articulation de l'affilié commence, si l'affilié n'est pas focalisé, le *stroke* commence avant l'articulation de l'affilié lexical (typiquement, 300 ms) ; le retour du geste débute lorsque la "prononciation" de l'affilié lexical est terminée. L'apport de cette approche par rapport à celle de Cassel et coll. est que les gestes sont planifiés par un "planificateur de mouvements" et qu'un ordonnanceur global gère la production des gestes successifs et permet à un geste d'être exécuté *avant* que le retour complet du geste précédent soit réalisé. Ici encore, la synchronie entre production de gestes manuels et production de parole se fait entre l'acoustique (et éventuellement l'intonation) de la phrase synthétisée et les phases "classiques" des gestes.

L'apport du manuscrit dans le cadre des agents conversationnels est important. En effet, au sein des travaux mentionnés ci-dessus, les hypothèses sous-jacentes à la génération des productions multimodales synchronisées pour agents conversationnels sont pour la plupart issues de travaux non quantitatifs, l'ajustement des timings précis des productions est donc *a priori* fait sur des bases empiriques. Les mesures présentées dans ce manuscrit permettraient l'ajustement temporel des productions dans les deux modalités.

Par ailleurs, un point important démontré par les mesures est la prise en compte de la dimension articulatoire dans la synchronisation des productions manuelles et vocales. Bien que les systèmes permettant l'animation des avatars fournissent un accès beaucoup plus direct aux données intonatives qu'aux données articulatoires (qui sont générées à partir des spécifications intonatives), le manuscrit présente assez clairement une synchronisation reposant sur une coordination articulatoire. Bien entendu, ces résultats sont valables pour le geste de pointage seulement, aucun résultat n'est avancé pour les gestes iconiques dans le manuscrit.

Enfin, les résultats présentés au Chapitre 3 montrent de façon encore plus évidente que les autres parties l'importance de la variabilité des productions au sein des participants. En particulier, pour le geste de pointage,

l'expérience a montré la constitution de deux groupes ayant des comportements coordinatifs geste manuel/parole très différents pour une même tâche. Cette variabilité semble intéressante à prendre en compte dans le cadre du *design* des agents conversationnels puisque la variabilité est nécessaire pour "humaniser" un système automatique, la variabilité étant le propre des productions réelles. Deux types de variations sont à prendre en compte : les variations intralocuteurs (par exemple, les alignements ne sont pas toujours *exactement* les mêmes) et les variation interlocuteurs (par exemple, dans la seconde expérience de ce manuscrit, les alignements entre le *stroke* du geste de pointage et la parole ne se font pas au même endroit pour deux "groupes" de participants).

6.2.3 Aide pour les pathologies ?

Une contribution possible des études présentées ci-dessus pourrait être l'aide au repérage précoce de pathologies entraînant des difficultés communicatives. En effet, de nombreux travaux ont montré un rôle des gestes dans les développements atypiques.

Au niveau de l'identification précoce des pathologies, en particulier pour l'autisme, un rôle plus efficace de l'étude des productions gestuelles que des productions vocales (voir par exemple Mitchell et coll. [132]) a été montré (les productions gestuelles fournissent un indicateur plus précoce que la parole pour l'identification des troubles autistiques). De façon comportementale, il a également été montré un rôle compensatoire pour les fonctions cognitives/langagières des gestes manuels chez les enfants atteints de trisomie 21 (comme indiqué par Pirchio, Caselli et Volterra ou Stefanini, Caselli et Volterra [145, 169]) ou de troubles spécifiques du développement du langage (selon Evans, Alibali et McNeill [50]) : les gestes sont plus nombreux (que dans d'autres populations ayant un développement typique) et expriment parfois des idées plus complexes que la parole.

Par ailleurs, il a été montré un rôle important du geste à la fois dans l'apprentissage du langage (*cf.* l'introduction, avec les publications de Goldin-Meadow et collaborateurs) mais également dans l'amélioration de la communication avec des enfants présentant un retard langagier (Brady et coll., Calculator [18, 23]) et dans l'amélioration des interactions sociales chez les jeunes enfants autistes en leur apprenant les gestes "corrects" dans certaines situations (voir Buffington et coll. [20]).

Les résultats présentés dans ce manuscrit décrivent surtout les relations temporelles entre les gestes et la parole dans différentes situations. En particulier, le geste de pointage, très utilisé par les jeunes enfants lors de l'acquisition du langage, a été étudié dans plusieurs situations. L'introduction (et en particulier, la Section 1.4.4) a montré un couplage fort entre les zones cérébrales liées à la production de parole multimodale et celles liées à la perception de cette parole. Il est possible de faire l'hypothèse –à tester !– que la constatation d'une désynchronisation des productions geste/parole (au moins, pour le geste de pointage, dans le cadre de ce qui a été montré ci-dessus) puisse mener à l'observation d'une moins bonne intégration multimodale (ou du moins, une intégration "différente" de l'intégration typique) des messages contenant les deux modalités. Une intégration déficiente des messages multimodaux pourrait être très dommageable lors de l'acquisition du langage, puisqu'à cette période, l'utilisation de la multimodalité est exacerbée (à la fois chez l'enfant mais également chez les parents). De façon préliminaire, pour tester cette hypothèse, il pourrait ainsi être intéressant de voir si les populations ayant des difficultés à intégrer les messages multimodaux ont effectivement des problèmes de synchronisation avec la parole lors de la production de gestes (typiquement, de pointage). Parmi les pathologies pour lesquelles ont été relevées des difficultés d'intégration multimodale, on peut mentionner les troubles du spectre autistique (*cf.* Carvalho Da Silva You [25, pp. 193-95]) pour lesquels le traitement des données multimodales est déficient : ce déficit peut s'expliquer selon les théories de plusieurs façons (déficits attentionnels, ou déficits d'intégration audio-visuelle –possiblement à cause d'un dysfonctionnement du sulcus temporal supérieur chez les personnes atteintes d'autisme). Enfin, pour les troubles autistiques, le désordre du traitement temporel des évènements est reconnu et peut jouer un rôle dans cette difficulté. La schizophrénie semble également affecter le traitement d'informations de parole multimodale (voir par exemple l'article de de Gelder et coll. [58] au titre évocateur) : les populations schizophrènes présentent des difficulté de traitement des sons de la parole (phonèmes). Bien que dans l'étude mentionnée, les problèmes soient *a priori* seulement liés à l'intégration d'une information (de type phonétique) venant de plusieurs sources (mouvement des lèvres et signal acoustique), il est possible que la co-occurrence temporelle des évènements joue un rôle exacerbé dans le cas de cette pathologie, tout comme pour

l'autisme. Si tel est le cas, l'étude de la synchronie des productions pourrait, par exemple, aider à diagnostiquer des problèmes plus complexes menant à des difficultés d'intégrations multimodales et donc à une interaction "sociale" plus difficile.

6.3 Ouvertures possibles

Bien entendu, les études présentées ci-avant sont loin d'être complètes et la démarche quantitative engagée (timidement) depuis quelques années semble nécessaire à poursuivre afin de pouvoir valider/falsifier les modèles mis au point, les affiner et ainsi comprendre de façon plus complète comment se déroule l'interaction gestes manuels/parole. Hormis les études complémentaires qu'il est possible de mener sur les données déjà acquises (par exemple, étude des énoncés contenant des pauses, des hésitations, caractérisation de la coordination dans ces cas particuliers, ...), il est possible d'utiliser les travaux présentés dans ce manuscrit comme une base de travail. L'étendue des expériences à mener est considérable, il s'agit donc de cibler les ouvertures possibles à ce travail.

6.3.1 Études de cas plus naturels

Un premier point qui a été amélioré en passant des deux premières expériences à la troisième est le fait que la tâche est plus naturelle dans la dernière expérience présentée. Cependant, il est évident que, bien que plus naturelle, cette tâche n'est clairement pas complètement naturelle, en particulier, une contrainte assez forte portant sur l'utilisation de modèles de phrases est toujours présente. Par ailleurs, le Chapitre 5 a montré la mise au point d'une méthode permettant l'annotation de vidéos d'interactions non contraintes (permettant de rendre compte des différentes compositions multimodales possibles).

L'étape suivante est donc assez claire du point de vue de l'évolution des expériences menées ci-dessus : libérer la dernière contrainte qui subsiste sur les productions *i.e.* supprimer la contrainte de modèle de phrase. *A priori*, la parole produite devrait être différente (du moins, les phrases "porteuses"), mais il y a de grandes chances que les gestes de pointages restent de mise (le design de l'expérience ayant été réalisé au sein du Chapitre 5 dans le but d'éliciter des pointages sans contrainte de parole). Cependant, il est également possible que d'autres types de gestes (par exemple, iconiques) fassent leur apparition dans la condition bruitée.

L'étude d'un cas totalement non contraint a des avantages certains en permettant l'étude de la coordination gestes/parole dans une situation d'interaction totalement libre tout en fournissant des mesures quantitatives des différents mouvements (à la fois articulatoires et manuels). Cependant, ce type d'étude ne permet pas de mener des analyses statistiques comme cela a été fait précédemment (puisqu'*a priori*, les productions de parole ne seront pas comparables à la fois pour un participant et surtout entre les participants).

Quelques enregistrements ont été acquis avec succès auprès des participants ayant réalisé l'expérience présentée au Chapitre 4. L'annotation et l'étude de ces productions reste à réaliser.

6.3.2 Amélioration du dispositif expérimental

Toujours dans un souci de rendre la tâche plus naturelle, il est possible d'améliorer le dispositif expérimental afin de le rendre moins "invasif" et ainsi de libérer les mouvements des participants au maximum. Les deux points à améliorer dans le dispositif expérimental de ce point de vue ont été évoqués plus haut dans le manuscrit : le casque audio pour la diffusion du bruit dans la dernière expérience et l'ensemble des fils qui relient les diodes infrarouge du système de capture de mouvement aux boîtiers permettant leur alimentation.

L'utilisation du casque audio est remplaçable par l'utilisation de haut-parleurs comme ceci a été essayé dans le Chapitre 5. Cependant, il a été montré dans le même chapitre que les méthodes de débruitage simples testées ne permettaient pas un débruitage suffisamment correct pour pouvoir étudier le signal de parole (initialement dégradé par le bruit environnant). Ainsi, pour rendre possible la diffusion de bruit par l'intermédiaire d'enceintes, il semble nécessaire de trouver une méthode de débruitage qui puisse s'adapter à un environnement qui varie constamment (les personnes bougent). L'utilisation d'un microphone plus directif peut également être une piste

à explorer : l'utilisation d'un microphone (très) directif se déplaçant avec la tête du locuteur peut possiblement suffire à atténuer assez le bruit environnant pour rendre l'analyse acoustique possible. Par ailleurs, l'utilisation du système de suivi de mouvements Optotrak a obligé, dans toutes les expériences, à coller des diodes infrarouge sur la face et les mains des participants. Les diodes infrarouge émettent des rayonnements infrarouges auxquels les caméras Optotrak sont sensibles, il faut donc aux diodes une alimentation électrique. Ceci n'est (pour l'instant) réalisable que par l'utilisation de câbles électriques qui, aussi fins soient-ils, peuvent produire une sensation de gêne pour les participants. Cette gêne peut évidemment se traduire par des mouvements plus étriqués, moins naturels. L'utilisation de la vidéo n'est pas une solution satisfaisante comme montré plus haut, en effet, la vidéo simple ne permet pas un suivi *tridimensionnel* des données et demande un temps de prétraitement des données rédhibitoire. Ainsi, une alternative possible est l'utilisation de système de suivi vidéo tridimensionnel de points (passifs). Les systèmes comme ceux proposés par Qualisys (`www.qualisys.com`) permettent de réaliser ce genre de suivi : trois caméras filment une scène et un logiciel permet le suivi de pastilles réflectives dans la vidéo. L'intérêt est ici d'économiser les fils partant des diodes et les reliant à une alimentation mais cet avantage est contrebalancé par un inconvénient qui est lié au fait que les points suivis soient passifs : le suivi de mouvement est moins correct (plus de données manquantes) et il faut assez régulièrement corriger *a posteriori* les données acquises. Par ailleurs, cette technique nécessite un bon éclairage de la scène ce qui peut parfois être gênant si l'expérience dure un temps conséquent et qui peut rendre la situation moins naturelle pour certaines personnes. Globalement, tous les systèmes reposant sur de la vidéo et un suivi de points passifs se heurteront aux mêmes problèmes (il en va donc de même pour des systèmes de type Kinect développé par Microsoft : `http://www.xbox.com/fr-FR/Kinect`). Ainsi, le choix du système de suivi de mouvement est un compromis à trouver entre la complétude/précision des données et le confort d'utilisation pour l'utilisateur.

Un dernier point à relever est le fait que dans les deux premières expériences, les gestes se faisaient pour des stimuli situés sur un écran en avant du participant, et plus particulièrement, dans le cas du pointage, le pointage se faisait vers cet écran en avant, alors que dans la troisième expérience, le geste de pointage est réalisé vers un plateau de jeu situé à l'horizontale devant le participant. Ainsi, il serait bon de trouver un dispositif expérimental permettant à la fois une interaction entre deux participants et des pointages similaires à ce qui était étudié dans les premières expériences (pointages plus courants). Ces dispositifs ont été testés mais abandonnés car produisant trop de données manquantes au niveau du suivi de mouvement, une solution possible serait d'utiliser deux systèmes Optotrak couplés afin de s'affranchir du masquage des données dû à une zone de jeu verticale (les systèmes Optotrak sont alors déportés de chaque côté de la "zone de jeu" et permettent de suivre les diodes sans être cachés par la zone de jeu).

6.3.3 Élargissement de ce type d'études à d'autres types de gestes

La caractérisation temporelle des productions conjointes gestes manuels/parole est complexe et les études présentées ci-dessus se sont intéressées à un nombre réduit de types de gestes. Plus particulièrement, un geste a pu être étudié sous plusieurs de ses aspects : le geste de pointage. Dans le cadre de la désignation dans lequel les expériences ont été menées, ce geste était primordial à étudier pour le rôle tout particulier qu'il entretient avec la parole. Par ailleurs, ce geste est important à étudier pour son rôle unique dans l'apprentissage du lexique et son utilisation quasi-quotidienne dans les interactions entre humains.

Cependant il paraît intéressant d'étudier ce qu'il se passerait pour d'autres types de gestes, en particulier pour les gestes iconiques, souvent analysés dans la littérature et assez facilement élicitables. En effet, nombreuses sont les études ayant porté sur ces gestes et la plupart des modèles ont été réalisés à partir d'expériences portant sur ce type de gestes. L'étude de la coordination gestes/parole (en utilisant un dispositif de mesure aussi précis que celui employé dans les expériences précédentes) permettrait de valider certains points avancés sur les mécanismes de coordination (en particulier, les phénomènes d'attraction motrice, l'influence des mouvements articulatoires sur la coordination). Pour les gestes iconiques, il est plus difficile (voire impossible) de définir un "apex" du geste mais les résultats ci-dessus sont également valables pour le stroke complet du geste.

Enfin, l'idée d'étudier à nouveau le geste de battement semble attrayante et en particulier, le geste de battement dans une tâche peu contrainte. En effet, les participants ont émis la remarque selon laquelle la production

contrainte de gestes de battement était une tâche complexe à réaliser (il est possible que la difficulté de cette tâche dépende de la culture –selon Krahmer et Swerts, au cours de leur étude présentée dans [103] demandant aux participants de produire des gestes de battement à des instants plus ou moins “naturels” au sein d’un énoncé, aucun participant n’a trouvé la tâche peu naturelle). Or il est attendu pour ce type de geste un rôle important de la synchronie temporelle avec la parole puisqu’un des rôles (assez consensuel) du battement est le marquage des parties importantes du discours.

Ces différents points peuvent être abordés (au moins en partie) dans une étude similaire à la troisième expérience dans laquelle aucune contrainte en parole n’est imposée.

Bibliographie

[1] Z. K. Agnew, S. Brownsett, Z. Woodhead et X. de Boissezon. « A Step Forward for Mirror Neurons ? Investigating the Functional Link between Action Execution and Action Observation in Limb Apraxia ». Dans : *The Journal of Neuroscience* 28.31 (2008), p. 7726–7727.

[2] M. A. Arbib. « From monkey-like action recognition to human language : an evolutionary framework for neurolinguistics ». Dans : *Behavioral and Brain Sciences* 28.2 (2005), p. 105–167.

[3] M. A. Arbib. « Interweaving protosign and protospeech : Further developments beyond the mirror ». Dans : *Interaction Studies* 6.2 (2005), p. 145–171.

[4] M. A. Arbib. « Mirror system activity for action and language is embedded in the integration of dorsal and ventral pathways ». Dans : *Brain and Language* 112.1 (jan. 2010), p. 12–24.

[5] D. F. Armstrong, W. C. Stokoe et S. E. Wilcox. *Gesture and the Nature of Language*. Cambridge University Press, 1995.

[6] P. Badin, Y. Tarabalka, F. Elisei et G. Bailly. « Can you 'read' tongue movements ? Evaluation of the contribution of tongue display to speech understanding ». Dans : *Speech Communication* 52.6 (2010), p. 493–503.

[7] F. Barbieri, A. Buonocore, R. D. Volta et M. Gentilucci. « How symbolic gestures and words interact with each other ». Dans : *Brain and Language* 110.1 (juil. 2009), p. 1–11.

[8] A. M. Barrett, L. S. Dore, K. A. Hansell et K. M. Heilman. « Speaking while gesturing : the relationship between speech and limb praxis. » Dans : *Neurology* 58.3 (fév. 2002), p. 499–500.

[9] E. Bates et F. Dick. « Language, gesture, and the developing brain ». Dans : *Developmental Psychobiology* 40.3 (2002), p. 293–310.

[10] J. B. Bavelas. « Gestures as part of speech : Methodological implications ». Dans : *Research on Language and Social Interaction* 27.3 (1994), p. 201–221.

[11] J. Bavelas, J. Gerwing, C. Sutton et D. Prevost. « Gesturing on the telephone : Independent effects of dialogue and visibility ». Dans : *Journal of Memory and Language* 58.2 (fév. 2008), p. 495–520.

[12] G. Beattie et H. Shovelton. « Why the spontaneous images created by the hands during talk can help make TV advertisements more effective. » Dans : *British journal of psychology (London, England : 1953)* 96.Pt 1 (fév. 2005), p. 21–37.

[13] C. Benoit, T. Mohamadi et S. Kandel. « Effects of Phonetic Context on Audio-Visual Intelligibility of French ». Dans : *Journal of Speech and Hearing Research* 37.5 (oct. 1994), p. 1195–1203.

[14] A. C. Berthoud. « Deixis, thématisation et détermination ». Dans : sous la dir. de M. A. Morel et L. Danon-Boileau. La Sorbonne, 1990, p. 527–542.

[15] R. L. Birdwhistell. « Some relations between American kinesics and spoken American English ». Dans : *Communication and culture*. Sous la dir. d'A. G. Smith. New York : Holt, Rinehart et Winston, 1966, p. 182–189.

[16] P. Boersma et D. Weenink. *Praat : doing phonetics by computer*. 1995-2010.

[17] D. Bolinger. « Intonation and gesture ». Dans : *American Speech* 58 (1983), p. 156–174.

[18] N. C. Brady, J. Marquis, K. Fleming et L. McLean. « Prelinguistic predictors of language growth in children with developmental disabilities. » Dans : *Journal of speech, language, and hearing research : JSLHR* 47.3 (juin 2004), p. 663–677.

[19] R. Breckinridge Church et S. Goldin-Meadow. « The mismatch between gesture and speech as an index of transitional knowledge ». Dans : *Cognition* 23.1 (juin 1986), p. 43–71.

[20] D. M. Buffington, P. J. Krantz, L. E. McClannahan et C. L. Poulson. « Procedures for teaching appropriate gestural communication skills to children with autism. » Dans : *Journal of autism and developmental disorders* 28.6 (déc. 1998), p. 535–545.

[21] C. Butcher et S. Goldin-Meadow. « Gesture and the transition from one- to two-word speech : when hand and mouth come together ». Dans : (). Sous la dir. de D. McNeill, p. 235–258.

[22] B. Butterworth et G. Beattie. « Gesture and silence as indicators of planning in speech ». Dans : *Recent advances in the psychology of language : Formal and experimental approaches* 4 (1978), p. 247–360.

[23] S. N. Calculator. « Use of Enhanced Natural Gestures to Foster Interactions Between Children With Angelman Syndrome and Their Parents ». Dans : *Am J Speech Lang Pathol* 11.4 (nov. 2002), p. 340–355.

[24] J. Call et M. Tomasello. « Production and comprehension of referential pointing by orangutans (Pongo pygmaeus). » Dans : *Journal of Comparative Psychology* 108.4 (1994), p. 307–317.

[25] R. S. Carvalho Da Silva You. « Etude de la perception catégorielle des stimuli auditifs et visuels de parole sans contenu émotionnel chez les enfants du Spectre Autistique ». Thèse de doct. Paris, France : Université Paris-Diderot (Paris 7), mar. 2012.

[26] J. Cassel. « A Framework For Gesture Generation And Interpretation ». Dans : sous la dir. de R. Cipolla et A. Pentland. Cambridge University Press, 1998, p. 191–215.

[27] J. Cassell, D. McNeill et K. E. McCullough. « Speech-gesture mismatches : Evidence for one underlying representation of linguistic and nonlinguistic information ». Dans : *Pragmatics 38 ; Cognition* 7.1 (1999), p. 1–34.

[28] C. Cavé, I. Guaïtella, R. Bertrand, S. Santi, F. Harlay et R. Espesser. « About the relationship between eyebrow movements and F_0 variations ». Dans : *Proceedings ICSLP 96* (1996). Sous la dir. de H. T. Bunnell et W. Idsardi, p. 2175–2178.

[29] S. Chieffi, C. Secchi et M. Gentilucci. « Deictic word and gesture production : Their interaction ». Dans : *Behavioural Brain Research* (mai 2009).

[30] R. B. Church, S. Ayman-Nolley et S. Mahootian. « The Role of Gesture in Bilingual Education : Does Gesture Enhance Learning ? » Dans : *International Journal of Bilingual Education and Bilingualism* 7.4 (2004), p. 303–319.

[31] N. Cocks, L. Dipper, R. Middleton et G. Morgan. « What can iconic gestures tell us about the language system ? A case of conduction aphasia ». Dans : *International Journal of Language & Communication Disorders* (nov. 2010), p. 101112063947042+.

[32] J.-M. Colletta, R. N. Kunene, A. Venouil, V. Kaufmann et J.-P. Simon. « Multimodal corpora ». Dans : sous la dir. de M. Kipp, J.-C. Martin, P. Paggio et D. Heylen. Berlin, Heidelberg : Springer-Verlag, 2009. Chap. Multi-track annotation of child language and gestures.

[33] J.-M. Colletta, C. Pellenq et M. Guidetti. « Age-related changes in co-speech gesture and narrative : Evidence from French children and adults ». Dans : *Speech Communication* 52.6 (juin 2010), p. 565–576.

[34] M. Corballis. « On the evolution of language and generativity ». Dans : *Cognition* 44.3 (1992), p. 197–226.

[35] M. C. Corballis. *From Hand to Mouth : The Origins of Language*. Princeton University Press, sept. 2003.

[36] C. Davis, J. Kim, K. Grauwinkel et H. Mixdorff. « Lombard Speech : Auditory (A), visual (V) and AV effects ». Dans : *Speech Prosody*. Dresden, Germany, mai 2006, p. 361–365.

[37] J. P. De Ruiter. « Can gesticulation help aphasic people speak, or rather, communicate ? » Dans : *International Journal of Speech-Language Pathology* 8.2 (2006), p. 124–127.

[38] J. de Ruiter. « Gesture and speech production ». Thèse de doct. Netherlands : Catholic University of Nijmegen, 1998.

[39] A. Di Cristo. « Vers une modélisation de l'accentuation du français (seconde partie) ». Dans : *Journal of French Language Studies* 10.01 (2000), p. 27–44.

[40] A. S. Dick, S. Goldin-Meadow, U. Hasson, J. I. Skipper et S. L. Small. « Co-speech gestures influence neural activity in brain regions associated with processing semantic information ». Dans : *Hum. Brain Mapp.* 30.11 (2009), p. 3509–3526.

[41] H. Diessel. « Demonstratives, joint attention, and the emergence of grammar ». Dans : *Cognitive Linguistics* 17.4 (déc. 2006), p. 463–489.

[42] M. D'Imperio, R. Espesser, H. Loevenbruck, C. Menezes, N. Nguyen et P. Welby. « Are tones aligned with articulatory events ? Evidence from Italian and French ». Dans : *Papers in Laboratory Phonology 9*. Sous la dir. de J. Cole. Mouton de Gruyter, 2007, p. 577–608.

[43] M. Dohen. « Deixis prosodique multisensorielle : production et perception audiovisuelle de la focalisation contrastive en français ». (Cognitive Science). Thèse de doct. France : Institut National Polytechnique de Grenoble, 2005.

[44] M. Dohen et H. Lœvenbruck. « Pre-focal Rephrasing, Focal Enhancement and Post-focal Deaccentuation in French ». Dans : *Interspeech-ICSLP 2004*. 2004, p. 1313–1316.

[45] J. E. Driskell et P. H. Radtke. « The effect of gesture on speech production and comprehension ». Dans : *Human Factors : The Journal of the Human Factors and Ergonomics Society* 45.3 (2003), p. 445–454.

[46] D. Efron. *Gesture and environment*. Sous la dir. de K. C. Press. Oxford, England, 1941.

[47] J. J. Egan. « Psychoacoustics of the Lombard voice response. » Dans : *Journal of Auditory Research* 12.4 (oct. 1972), p. 318–324.

[48] P. Ekman et W. V. Friesen. « The repertoire of nonverbal behavior : Categories, origins, usage, and coding ». Dans : *Semiotica* 1.1 (1969), p. 49–98.

[49] N. J. Enfield. « Lip-pointing : A discussion of form and function with reference to data from Laos ». Dans : *Gesture* (2001), p. 185–211.

[50] J. L. Evans, M. W. Alibali et N. M. McNeil. « Divergence of verbal expression and embodied knowledge : Evidence from speech and gesture in children with specific language impairment ». Dans : *Language and Cognitive Processes* 16.2-3 (avr. 2001), p. 309–331.

[51] J. Fast. *Body Language*. 1st edition. London : Pan Books, mai 1971.

[52] P. Feyereisen. « The Competition between Gesture and Speech Production in Dual-Task Paradigms ». Dans : *Journal of Memory and Language* 36.1 (1997), p. 13–33.

[53] P. Feyereisen. « Further investigation on the mnemonic effect of gestures : Their meaning matters ». Dans : *European Journal of Cognitive Psychology* 18.2 (mar. 2006), p. 185–205.

[54] A. Floel, T. Ellger, C. Breitenstein et S. Knecht. « Language perception activates the hand motor cortex : implications for motor theories of speech perception ». Dans : *European Journal of Neuroscience* 18.3 (août 2003), p. 704–708.

[55] P. Garber et S. Goldin-Meadow. « Gesture offers insight into problem-solving in adults and children ». Dans : *Cognitive Science* 26.6 (2002), p. 817–831.

[56] M. Garnier. « Communiquer en environnement bruyant : de l'adaptation jusqu'au forçage vocal ». Français. Thèse de doct. Université Pierre et Marie Curie - Paris VI, juin 2007.

[57] M. Garnier, L. Bailly, M. Dohen, P. Welby et H. Lœvenbruck. « An Acoustic and Articulatory Study of Lombard Speech : Global Effects on the Utterance ». Dans : *INTERSPEECH 2006*. Pittsburgh, Pennsylvania, sept. 2006, p. 2246+.

[58] B. de Gelder, J. Vroomen, L. Annen, E. Masthof et P. Hodiamont. « Audio-visual integration in schizophrenia. » Dans : *Schizophrenia research* 59.2-3 (fév. 2003), p. 211–218.

[59] M. Gentilucci et M. C. Corballis. « From manual gesture to speech : a gradual transition. » Dans : *Neuroscience and biobehavioral reviews* 30.7 (2006), p. 949–960.

[60] S. Goldin-Meadow, M. Alibali et R. Church. « Transitions in concept acquisition : Using the hand to read the mind. » Dans : *Psychological review* 100.2 (1993), p. 279.

[61] S. Goldin-Meadow, H. Nusbaum, S. D. Kelly et S. Wagner. « Explaining math : gesturing lightens the load. » Dans : *Psychological science* 12.6 (nov. 2001), p. 516–522.

[62] S. Goldin-Meadow. *Hearing Gesture : How Our Hands Help Us Think*. Belknap Press of Harvard University Press, oct. 2005.

[63] S. Goldin-Meadow. « Learning through gesture ». Dans : *Wiley Interdisciplinary Reviews : Cognitive Science* 2.6 (2011), p. 595–607.

[64] S. Goldin-Meadow et C. Butcher. « Pointing Toward Two-Word Speech in Young Children ». Dans : *Pointing : Where Language, Culture and Cognition Meet*. Sous la dir. de S. Kita. Lawrence Erlbaum Associates, 2003. Chap. 5, p. 85–108.

[65] S. Goldin Meadow, D. Wein et C. Chang. « Assessing Knowledge through Gesture : Using Children's Hands to Read Their Minds ». Dans : *Cognition and Instruction* 9.3 (1992), p. 201–219.

[66] C. Gonseth, C. Vilain et A. Vilain. « Encodage de la distance et coopération parole/geste : étude développementale du pointage multimodal ». Dans : *Journées d'Etudes en Parole, Traitement Automatique des Langues Naturelles (JEP-TALN)*. Grenoble, juin 2012, p. 681–688.

[67] M. Guidetti. « The emergence of pragmatics : forms and functions of conventional gestures in young French children ». Dans : *First Language* 22.3 (oct. 2002), p. 265–285.

[68] B. Habets, S. Kita, Z. Shao, A. Özyürek et P. Hagoort. « The Role of Synchrony and Ambiguity in Speech-Gesture Integration during Comprehension ». Dans : *Journal of Cognitive Neuroscience* (jan. 2011), p. 1–10.

[69] U. Hadar. « Two types of gesture and their role in speech production ». Dans : *Journal of Language and Social Psychology* 8.3-4 (1989), p. 221–228.

[70] U. Hadar, D. Wenkert-Olenik, R. Krauss et N. Soroker. « Gesture and the Processing of Speech : Neuropsychological Evidence, » dans : *Brain and Language* 62.1 (mar. 1998), p. 107–126.

[71] O. Hauk, I. Johnsrude et F. Pulvermüller. « Somatotopic Representation of Action Words in Human Motor and Premotor Cortex ». Dans : *Neuron* 41.2 (jan. 2004), p. 301–307.

[72] G. W. Hewes. « Primate Communication and the Gestural Origin of Language [and Comments and Reply] ». Dans : *Current Anthropology* 14.1/2 (1973), p. 5–24.

[73] D. Holender. « Interference between a vocal and a manual response to the same stimulus ». Dans : *Tutorials in Motor Behavior* (1980). Sous la dir. de G. E. Stelmach et J. Requin.

[74] H. Holle et T. C. Gunter. « The Role of Iconic Gestures in Speech Disambiguation : ERP Evidence ». Dans : *Journal of Cognitive Neuroscience* 19.7 (juin 2007), p. 1175–1192.

[75] H. Holle, J. Obleser, S.-A. Rueschemeyer et T. C. Gunter. « Integration of iconic gestures and speech in left superior temporal areas boosts speech comprehension under adverse listening conditions ». Dans : *NeuroImage* 49.1 (jan. 2010), p. 875–884.

[76] A. B. Hostetter. « When do gestures communicate ? A meta-analysis. » Dans : *Psychological Bulletin* 137.2 (2011), p. 297–315.

[77] A. B. Hostetter et M. W. Alibali. « Raise your hand if you're spatial : Relations between verbal and spatial skills and gesture production ». Dans : *Gesture* 7.1 (avr. 2007), p. 73–95.

[78] A. B. Hostetter et M. W. Alibali. « Visible embodiment : gestures as simulated action. » Dans : *Psychonomic bulletin & review* 15.3 (juin 2008), p. 495–514.

[79] A. B. Hostetter et M. W. Alibali. « Language, gesture, action ! A test of the Gesture as Simulated Action framework ». Dans : *Journal of Memory and Language* (mai 2010).

[80] Y. H. Hsieh. « Gestural Beats in Chinese Narrative Discourse ». Mém.de maîtr. Taiwan : National Chengchi University, 2007.

[81] A. L. Hubbard, S. M. Wilson, D. E. Callan et M. Dapretto. « Giving Speech a Hand : Gesture Modulates Activity in Auditory Cortex During Speech Perception ». Dans : *Human Brain Mapping* 30 (2009). Sous la dir. d'I. Wiley-Liss, p. 1028–1037.

[82] J. M. Iverson et E. Thelen. « Hand, mouth and brain. The dynamic emergence of speech and gesture ». Dans : *Journal of Consciousness Studies* 6.11-12 (1999), p. 19–40.

[83] J. M. Iverson et S. Goldin-Meadow. « Why people gesture when they speak ». Dans : *Nature* 396.6708 (1998), p. 228–228.

[84] J. M. Iverson et S. Goldin-Meadow. « Gesture Paves the Way for Language Development ». Dans : *Psychological Science* 16.5 (mai 2005), p. 367–371.

[85] N. Jacobs et A. Garnham. « The role of conversational hand gestures in a narrative task ». Dans : *Journal of Memory and Language* 56.2 (fév. 2007), p. 291–303.

[86] S. Jannedy et N. Mendoza-Denton. « Structuring information through gesture and intonation ». Dans : *Interdisciplinary Studies on Information Structure* 3 (2005), p. 199–244.

[87] R. E. Johnson. « A comparison of the phonological structures of two northwest sawmill sign languages ». Dans : *Communication and Cognition* 11 (1978), p. 105–132.

[88] J.-C. Junqua. « The Lombard reflex and its role on human listeners and automatic speech recognizers ». Dans : 93.1 (jan. 1993), p. 510–524.

[89] S. D. Kelly, A. Özyürek et E. Maris. « Two Sides of the Same Coin ». Dans : *Psychological Science* 21.2 (fév. 2010), p. 260–267.

[90] J. Kelso. « Haken-Kelso-Bunz model ». Dans : *Scholarpedia* 3.10 (2008), p. 1612+.

[91] J. Kelso, B. Tuller et K. Harris. « A "dynamic pattern" perspective on the control and coordination of movement ». Dans : *The production of speech* (1983), p. 137–173.

[92] A. Kendon. « Gesture and speech : How they interact ». Dans : *Nonverbal interaction* (1983), p. 13–45.

[93] A. Kendon. « Some relationships between body motion and speech ». Dans : *Studies in Dyadic Communication*. Sous la dir. d'A. Seigman et B. Pope. Elmsford, New York : Pergamon Press, 1972, p. 177–216.

[94] A. Kendon. « Gesticulation and speech : Two aspects of the process of utterance ». Dans : *The relation between verbal and nonverbal communication*. Sous la dir. de M. R. Key. The Hague : Mouton, 1980, p. 207–227.

[95] A. Kendon. « Do Gestures Communicate ? A Review ». Dans : *Research on Language & Social Interaction* 27.3 (juil. 1994), p. 175–200.

[96] A. Kendon. *Gesture : Visible Action as Utterance*. Cambridge University Press, 2004.

[97] T. Kircher, B. Straube, D. Leube, S. Weis, O. Sachs, K. Willmes, K. Konrad et A. Green. « Neural interaction of speech and gesture : Differential activations of metaphoric co-verbal gestures ». Dans : *Neuropsychologia* 47.1 (jan. 2009), p. 169–179.

[98] S. Kita. « How Representational Gestures Help Speaking ». Dans : *Language and Gesture : Window into Thought and Action*. Sous la dir. de D. McNeill. Cambridge Univ. Press, 2000. Chap. How Representational Gestures Help Speaking, p. 162–185.

[99] S. Kita. *Pointing, Where Language, Culture, and Cognition Meet*. Lawrence Erlbaum Associates, 2003.

[100] S. Kita et A. Özyürek. « What does cross-linguistic variation in semantic coordination of speech and gesture reveal ? : Evidence for an interface representation of spatial thinking and speaking ». Dans : *Journal of Memory and Language* 48.1 (2003), p. 16–32.

[101] S. Kita, A. Özyürek, S. Allen, A. Brown, R. Furman et T. Ishizuka. « Relations between syntactic encoding and co-speech gestures : Implications for a model of speech and gesture production ». Dans : *Language and Cognitive Processes* 22.8 (2007), p. 1212–1236.

[102] S. Kopp et I. Wachsmuth. « Synthesizing multimodal utterances for conversational agents ». Dans : *Computer Animation and Virtual Worlds* 15.1 (mar. 2004), p. 39–52.

[103] E. Krahmer et M. Swerts. « The effects of visual beats on prosodic prominence : Acoustic analyses, auditory perception and visual perception ». Dans : *Journal of Memory and Language* 57.3 (2007), p. 396–414.

[104] R. M. Krauss, Y. Chen et R. F. Gottesman. « Lexical gestures and lexical access : A process model ». Dans : *Language and Gesture*. New York : Cambridge University Press, 2000, p. 261–283.

[105] R. M. Krauss et U. Hadar. « The role of speech-related arm/hand gestures in word retrieval ». Dans : *Gesture, speech, and sign* (1999), p. 93–116.

[106] R. M. Krauss. « Why Do We Gesture When We Speak ? » Dans : *Current Directions in Psychological Science* 7.2 (1998), p. 54–60.

[107] R. M. Krauss, Y. Chen et P. Chawla. « Nonverbal behavior and nonverbal communication : What do conversational hand gestures tell us ? » Dans : *Advances in Experimental Social Psychology* 8 (1996). Sous la dir. de M. Zanna, p. 389–450.

[108] R. M. Krauss, R. A. Dushay, Y. Chen et F. Rauscher. « The Communicative Value of Conversational Hand Gesture ». Dans : *Journal of Experimental Social Psychology* 31.6 (nov. 1995), p. 533–552.

[109] A. C. Kuntay et A. Özyürek. « Joint Attention and the Development of the Use of Demonstrative Pronouns in Turkish ». Dans : *Proceedings of the 26th annual Boston University Conference on Language Development*. T. 1. Cascadilla Press. 2002, p. 336.

[110] S. Kurumi. « A Comparative Study on Grammar of Alternate Sign Language : Monastic Sign Language and Sawmill Sign Language ». Dans : *Study report of the Japan College of Social Work* 46 (1999-12), p. 3–44.

[111] D. Laurent. « Pointage et démonstratifs : Exploration de la grammaire par la coordination geste-parole ». Mém.de maîtr. Grenoble, France : Université Sthendhal, 2007.

[112] A. Lemasson, K. Ouattara, E. Petit et K. Zuberbuhler. « Social learning of vocal structure in a nonhuman primate ? » Dans : *BMC Evolutionary Biology* 11.1 (déc. 2011), p. 362+.

[113] T. Leonard et F. Cummins. « The temporal relation between beat gestures and speech ». Dans : *Language and Cognitive Processes* (2010).

[114] W. Levelt, G. Richardson et W. La Heij. « Pointing and Voicing in Deictic Expressions ». Dans : *Journal of Memory and Language* 24.2 (1985), p. 133–164.

[115] W. J. M. Levelt. *Speaking : From Intention to Articulation*. Cambridge, MA : MIT Press, 1989.

[116] D. P. Loehr. « Gesture and Intonation ». Thèse de doct. Faculty of the Graduate School of Arts et Sciences of Georgetown University, 2004.

[117] A. Lücking, K. Bergmann, F. Hahn, S. Kopp et H. Rieser. « The bielefeld speech and gesture alignment corpus (saga) ». Dans : *LREC 2010 Workshop : Multimodal CorporaAdvances in Capturing, Coding and Analyzing Multimodality* (2010). Sous la dir. de M. Kipp, J. P. Martin, P. Paggio et D. Heylen, p. 92–98.

[118] F. Maricchiolo, A. Gnisci, M. Bonaiuto et G. Ficca. « Effects of different types of hand gestures in persuasive speech on receivers' evaluations ». Dans : *Language and Cognitive Processes* 24.2 (fév. 2009), p. 239–266.

[119] E. Mathiot, M. Leroy, F. Limousin et A. Morgenster. « Premiers pointages chez l'enfant entendant et l'enfant sourd-signeur : deux suivis longitudinaux entre 7 mois et 1 an 7 mois ». Dans : *Acquisition et interaction en langue étrangère* (2009), p. 141–168.

[120] MathWorks. *Matlab*. 1984-2009.

[121] R. I. Mayberry et J. Jaques. « Gesture production during stuttered speech : Insights into the nature of gesture-speech integration ». Dans : *Language and Gesture : Window into Thought and Action*. Sous la dir. de D. McNeill. Cambridge Univ. Press, 2000, p. 199–213.

[122] R. I. Mayberry, J. Jaques et G. DeDe. « What Stuttering Reveals about the Development of the Gesture-Speech Relationship. » Dans : *New Directions for Child Development* (1998).

[123] E. McClave. « Gestural beats : The rhythm hypothesis ». Dans : *Journal of Psycholinguistic Research* 23.1 (1994), p. 45–66.

[124] E. Z. McClave. « Pitch and Manual Gestures ». Dans : *Journal of Psycholinguistic Research* 27.1 (1998), p. 69–89.

[125] H. McGurk et J. MacDonald. « Hearing lips and seeing voices ». Dans : *Nature* 264.5588 (déc. 1976), p. 746–748.

[126] D. McNeill, L. Quaeghebeur et S. Duncan. « IW-"The Man Who Lost His Body" ». Dans : *Handbook of Phenomenology and Cognitive Science* (2010), p. 519–543.

[127] D. McNeill. « So you think gestures are nonverbal ? » Dans : *Psychological Review* 92.3 (1985), p. 350–371.

[128] D. McNeill. *Hand and Mind : What Gestures Reveal about Thought*. University Of Chicago Press, 1992.

[129] D. McNeill. *Language and Gesture*. Cambridge University Press, août 2000.

[130] D. McNeill et S. D. Duncan. « Growth Points in Thinking-for-Speaking ». Dans : *Language and Gesture*. 2000, p. 141–161.

[131] D. McNeill, S. D. Duncan, J. Cole, S. Gallagher et B. Bertenthal. « Growth points from the very beginning ». Dans : *Interaction Studies* 9.1 (avr. 2008), p. 117–132.

[132] S. Mitchell, J. Bian, L. Zwaigenbaum, W. Roberts, P. Szatmari, I. Smith et S. Bryson. « Early Language and Communication Development of Infants Later Diagnosed with Autism Spectrum Disorder ». Dans : *Journal of Developmental & Behavioral Pediatrics* 27.2 (2006).

[133] P. Morrel-Samuels et R. M. Krauss. « Word familiarity predicts temporal asynchrony of hand gestures and speech. » Dans : *Journal of Experimental Psychology : Learning, Memory, and Cognition* 18.3 (1992), p. 615–622.

[134] E. Morsella et R. M. Krauss. « The role of gestures in spatial working memory and speech. » Dans : *The American journal of psychology* 117.3 (2004), p. 411–424.

[135] Neurobehavioral Systems. *Presentation*. 2009.

[136] S. Nobe. « Representational gestures, cognitive rhythms, and acoustic aspects of speech : A network/threshold model of gesture production ». Thèse de doct. The Faculty of the Division of the Social Sciences, 1996.

[137] H. Nølke. *Linguistique modulaire : de la forme au sens*. T. 28. Peeters Pub & Booksellers, 1994.

[138] C. Obermeier, H. Holle et T. C. Gunter. « What Iconic Gesture Fragments Reveal about Gesture–Speech Integration : When Synchrony Is Lost, Memory Can Help. » Dans : *Journal of Cognitive Neuroscience* 23.7 (2011), p. 1648 –1663.

[139] R. C. Oldfield. « The assessment and analysis of handedness : The Edinburgh inventory ». Dans : *Neuropsychologia* 9.1 (mar. 1971), p. 97–113.

[140] A. Özyürek. « Do Speakers Design Their Cospeech Gestures for Their Addressees ? The Effects of Addressee Location on Representational Gestures ». Dans : *Journal of Memory and Language* 46.4 (mai 2002), p. 688–704.

[141] A. Özyürek, S. Kita, S. E. M. Allen, R. Furman et A. Brown. « How does linguistic framing of events influence co-speech gestures ? : Insights from crosslinguistic variations and similarities ». Dans : *Gesture* 5.1 (2005), p. 219–240.

[142] A. Özyürek, R. M. Willems, S. Kita et P. Hagoort. « On-line Integration of Semantic Information from Speech and Gesture : Insights from Event-related Brain Potentials ». Dans : *Journal of Cognitive Neuroscience* 19.4 (mar. 2007), p. 605–616.

[143] C. Pelachaud. « Studies on gesture expressivity for a virtual agent ». Dans : *Speech Commun.* 51.7 (juil. 2009).

[144] M. Perry, R. Breckinridge Church et S. Goldin-Meadow. « Transitional knowledge in the acquisition of concepts ». Dans : *Cognitive Development* 3.4 (oct. 1988), p. 359–400.

[145] S. Pirchio, M. C. Caselli et V. Volterra. « Gestes, mots et tours de parole chez des enfants atteints du syndrome de Williams ou du syndrome de Down ». Dans : *Enfance* 55.3 (2003), p. 251+.

[146] A. S. Pollick et F. B. de Waal. « Ape gestures and language evolution ». Dans : *Proceedings of the National Academy of Sciences* 104.19 (mai 2007), p. 8184–8189.

[147] J. Prado et E. Moulines. « Frequency-domain adaptive filtering with applications to acoustic echo cancellation ». Dans : *Annals of Telecommunications* 49 (7 1994). 10.1007/BF02999430, p. 414–428.

[148] F. Pulvermüller, O. Hauk, V. V. Nikulin et R. J. Ilmoniemi. « Functional links between motor and language systems. » Dans : *The European journal of neuroscience* 21.3 (fév. 2005), p. 793–797.

[149] M. F. Quintilianus. *Institutio Oratoria*. T. XI/3. ˜50A.D.

[150] R Development Core Team. *R : A Language and Environment for Statistical Computing*. ISBN 3-900051-07-0. R Foundation for Statistical Computing. Vienna, Austria, 2010.

[151] L. R. Rabiner. « A tutorial on hidden Markov models and selected applications inspeech recognition ». Dans : *Proceedings of the IEEE* 77.2 (fév. 1989), p. 257–286.

[152] F. H. Rauscher, R. M. Krauss et Y. Chen. « Gesture, Speech, and Lexical Access : The Role of Lexical Movements in Speech Production ». Dans : *Psychological Science* 7.4 (juil. 1996), p. 226–231.

[153] G. Rizzolatti, L. Fadiga, V. Gallese et L. Fogassi. « Premotor cortex and the recognition of motor actions. » Dans : *Brain research. Cognitive brain research* 3.2 (mar. 1996), p. 131–141.

[154] G. Rizzolatti et M. A. Arbib. « Language within our grasp ». Dans : *Trends in Neurosciences* 21.5 (1998), p. 188–194.

[155] J. Robert-Ribes, J.-L. Schwartz et P. Escudier. « A comparison of models for fusion of the auditory and visual sensors in speech perception ». Dans : *Artificial Intelligence Review* 9.4 (oct. 1995), p. 323–346.

[156] A. Rochet-Capellan. « De la substance à la forme : rôle des contraintes motrices orofaciales et brachio-manuelles de la parole dans l'émergence du langage ». Thèse de doct. France : Institut National Polytechnique de Grenoble, 2007.

[157] A. Rochet-Capellan, R. Laboissière, A. Galvan et J.-L. Schwartz. « The speech focus position effect on jaw-finger coordination in a pointing task ». Dans : *Journal of Speech, Language, and Hearing Research* 51 (2008), p. 1507–1521.

[158] L. D. Rosenblum. « Speech Perception as a Multimodal Phenomenon ». Dans : *Current Directions in Psychological Science* 17.6 (déc. 2008), p. 405–409.

[159] J. P. de Ruiter et D. P. Wilkins. « Cross-cultural psycholinguistic evidence for the modularity of speech and gesture (The synchronization of gesture and speech in Dutch and Arrernte) ». Dans : *Oralité et gestualité. Communication multimodale, interaction*. Sous la dir. de S. I. Santi, I. Guaïtella, C. Cavé et G. Konopczynski. Paris, 1998, p. 603–607.

[160] H. L. Rusiewicz. « The role of prosodic stress and speech perturbation on the temporal synchronization of speech and deictic gestures. » Thèse de doct. Pittsburg : University of Pittsburg, 2010.

[161] H. L. Rusiewicz. « Synchronization of prosodic stress and gesture : a dynamic systems perspective. » Dans : *Gesture and Speech in Interaction (GESPIN'11)*. Bielefeld, 2011.

[162] J. L. Russell, S. Braccini, N. Buehler, M. J. Kachin, S. J. Schapiro et W. D. Hopkins. « Chimpanzee (Pan troglodytes) intentional communication is not contingent upon food ». Dans : *Animal Cognition* 8.4 (oct. 2005), p. 263–272.

[163] R. Schulman. « Articulatory dynamics of loud and normal speech ». Dans : 85.1 (jan. 1989), p. 295–312.

[164] J.-L. Schwartz, F. Berthommier et C. Savariaux. « Seeing to hear better : evidence for early audio-visual interactions in speech identification ». Dans : *Cognition* 93.2 (2004), B69–B78.

[165] E. A. Sheehan, L. L. Namy et D. L. Mills. « Developmental changes in neural activity to familiar words and gestures. » Dans : *Brain and language* 101.3 (juin 2007), p. 246–259.

[166] J. I. Skipper, S. Goldin-Meadow, H. C. Nusbaum et S. L. Small. « Speech-associated gestures, Broca's area, and the human mirror system. » Dans : *Brain and language* 101.3 (juin 2007), p. 260–277.

[167] D. I. Slobin. « From "thought and language" to "thinking for speaking" ». Dans : *Rethinking linguistic relativity* 17 (1996), p. 70–96.

[168] W. C. So, C. Sim Chen-Hui et J. Low Wei-Shan. « Mnemonic effect of iconic gesture and beat gesture in adults and children : Is meaning in gesture important for memory recall ? » Dans : *Language and Cognitive Processes* (mai 2011), p. 1–17.

[169] S. Stefanini, M. C. Caselli et V. Volterra. « Spoken and gestural production in a naming task by young children with Down syndrome ». Dans : *Brain and Language* 101.3 (juin 2007), p. 208–221.

[170] A. Sueyoshi et D. M. Hardison. « The Role of Gestures and Facial Cues in Second Language Listening Comprehension ». Dans : *Language Learning* 55.4 (déc. 2005), p. 661–699.

[171] W. H. Sumby et I. Pollack. « Visual Contribution to Speech Intelligibility in Noise ». Dans : 26.2 (mar. 1954), p. 212–215.

[172] Q. Summerfield. « Some preliminaries to a comprehensive account of audio-visual speech perception ». Dans : *Hearing by eye : The psychology of lip-reading*. Sous la dir. de Dodd. Hillsdale, New Jersey : Lawrence Erlbaum Associates, 1987.

[173] M. Tellier. « The effect of gestures on second language memorisation by young children ». Dans : *Gesture* 8.2 (2008), p. 219–235.

[174] S. Ternstrom, M. Sodersten et M. Bohman. « Cancellation of Simulated Environmental Noise as a Tool for Measuring Vocal Performance During Noise Exposure ». Dans : *Journal of Voice* 16.2 (juin 2002), p. 195–206.

[175] S. Ternström, M. Bohman et M. Södersten. « Loud speech over noise : Some spectral attributes, with gender differences ». Dans : 119.3 (mar. 2006), p. 1648–1665.

[176] H. Traunmüller et A. Eriksson. « Acoustic effects of variation in vocal effort by men, women, and children ». Dans : *The Journal of the Acoustical Society of America* 107.6 (2000), p. 3438–3451.

[177] K. Tuite. « The production of gesture ». Dans : *Semiotica* 93.1-2 (1993), p. 83–106.

[178] E. Vatikiotis-Bateson, I.-M. Eigsti, S. Yano et K. G. Munhall. « Eye movement of perceivers during audiovisualspeech perception ». Dans : *Attention, Perception, & Psychophysics* 60.6 (sept. 1998), p. 926–940.

[179] R. Wesp, J. Hesse, D. Keutmann et K. Wheaton. « Gestures Maintain Spatial Imagery ». Dans : *The American Journal of Psychology* 114.4 (2001), p. 591–600.

[180] H. Wickham. *ggplot2 : elegant graphics for data analysis*. Springer New York, 2009.

[181] R. M. Willems et P. Hagoort. « Neural evidence for the interplay between language, gesture, and action : A review ». Dans : *Brain and Language* 101.3 (juin 2007), p. 278–289.

[182] R. M. Willems, A. Özyürek et P. Hagoort. « When Language Meets Action : The Neural Integration of Gesture and Speech ». Dans : *Cerebral Cortex* 17.10 (oct. 2007), p. 2322–2333.

[183] K. A. Wilmes. « Hands in Focus : Focus Marking by Speech-Accompanying Gestures ». Mém.de maîtr. Germany : Cognitive Science dpt, University of Osnabrück, 2009.

[184] P. Wittenburg, H. Brugman, A. Russel, A. Klassmann et H. Sloetjes. « Elan : a professional framework for multimodality research ». Dans : *Proceedings of Language Resources and Evaluation Conference (LREC)*. 2006.

[185] P. Wolff et J. Gutstein. « Effects of induced motor gestures on vocal output. » Dans : *The Journal of communication* 22.3 (sept. 1972), p. 277–288.

[186] K. Yerian. « Directions in intonation and gesture research : A Review ». Thèse de doct. Washington DC : Georgetown University, 1995.

[187] S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev et P. C. Woodland. *The HTK Book*. Cambridge, UK : Cambridge University Engineering Department, 2006.

[188] J. Zeiliger, J. F. Serignat, D. Auteserre et C. Meunier. « BDBRUIT, une base de données parole de locuteurs soumis à du bruit ». Dans : *10th Journées d'Etudes sur la Parole*. Lannion, France, juin 1994, p. 287–290.

Appendices

Annexe A

Consignes données aux participants

A.1 Consignes pour la première expérience

Vous allez donc entendre deux locuteurs se parler et vous allez corriger ce que dit le second locuteur en fonction de ce qu'a dit le premier.
Parlez clairement avec un débit normal (ni trop lent, ni trop rapide).
Vous allez voir des images s'afficher en face de vous. Celles-ci correspondent aux phrases entendues.
Pendant cette phase, vous n'avez rien à faire avec votre main et vous la laisserez donc sur la marque bleue de repos.
Nous allons commencer par quelques essais d'entraînement.

(a) Exemple de consigne : Parole seule

Vous allez entendre deux locuteurs se parler et vous allez corriger ce que dit le second locuteur en fonction de ce qu'à dit le premier.
Parlez clairement avec un débit normal (ni trop lent, ni trop rapide).
Vous allez voir des images s'afficher en face de vous. Celles-ci correspondent aux phrases entendues. Pendant cette phase, vous effectuerez un geste <TYPE DE GESTE> en même temps que vous parlerez. <DESCRIPTION RAPIDE DU GESTE>
Veillez à bien partir de la marque de repos et à y revenir quand votre mouvement est terminé.
Nous allons commencer par quelques essais d'entraînement.

(b) Exemple de consigne : Parole + geste

Figure A.1 – Consignes de la première expérience présentée au Chapitre 2

A.2 Consignes pour la seconde expérience

Vous allez entendre un locuteur dire une phrase du type OBJET est COULEUR puis vous verrez deux images s'afficher qui vous permettront de corriger ce qu'a dit le locuteur.
Parlez clairement avec un débit normal (ni trop lent, ni trop rapide).
Pendant cette phase vous ne bougerez pas la main qui devra rester sur la marque bleue de repos.
Nous allons commencer par quelques essais d'entraînement.

(a) Exemple de consigne : Parole seule

Vous allez entendre un locuteur dire une phrase du type OBJET est COULEUR puis vous verrez deux images s'afficher qui vous permettront de corriger ce qu'a dit le locuteur.
Parlez clairement avec un débit normal (ni trop lent, ni trop rapide) et effectuez un geste <TYPE DE GESTE> en même temps que vous parlez. <DESCRIPTION SOMMAIRE DU GESTE>
Veillez à bien partir de la marque de repos et à y revenir quand votre mouvement est terminé.
Nous allons commencer par quelques essais d'entraînement.

(b) Exemple de consigne : Parole + geste

Figure A.2 – Consignes de l'expérience présentée en Chapitre 3

A.3 Consignes pour l'expérience en interaction

Lire les consignes à voix haute.
Vous êtes complètement libre et votre tâche consiste uniquement à faire reproduire le modèle sur le plateau par votre interlocuteur.
N'oubliez pas, le but du jeu est d'aller le plus vite possible et votre interlocuteur ne peut vous poser aucune question. Il dispose simplement d'images qu'il peut disposer sur la table selon vos instructions.
Vous pouvez bouger librement dans la mesure où vous restez assis(e) sur la chaise.
Quand vous êtes prêt(e) dites à votre interlocuteur d'appuyer sur le bouton.
Dès que la grille est finie, dites à votre interlocuteur de rappuyer sur le bouton.
A vous de jouer…

(a) Consignes d'entraînement

Lire les consignes à voix haute.
La règle du jeu reste la même sauf que vous n'allez pouvoir dire que des phrases du type « Un X va là. ».
Par exemple : Un ballon va là.
Vous ne pouvez dire aucune autre phrase, ni aucun autre mot.
Vous pouvez par contre bouger librement dans la mesure où vous restez assis(e) sur la chaise.
Quand vous êtes prêt(e) dites à votre interlocuteur d'appuyer sur le bouton.
Dès que la grille est finie, dites à votre interlocuteur de rappuyer sur le bouton.
A vous de jouer…

(b) Exemple de consigne : Condition non bruitée

Lire les consignes à voix haute.
Cette phase va se dérouler exactement de la même façon que la précédente (« Un X va là. ») sauf que cette fois, vous allez jouer dans le bruit.
Quand vous êtes prêt(e) dites à votre interlocuteur d'appuyer sur le bouton.
Dès que la grille est finie, dites à votre interlocuteur de rappuyer sur le bouton.
A vous de jouer…

(c) Exemple de consigne : Condition bruitée

FIGURE A.3 – Consignes de l'expérience d'intéraction du Chapitre 4

Résumé

Le travail présenté dans cette thèse vise à étudier la coordination entre gestes manuels et parole lors de la production d'énoncés multimodaux. Cette coordination est étudiée plus précisément dans le cadre de la désignation, réalisable à la fois dans la modalité manuelle et en parole. Les études présentées sont menées dans un environnement contrôlé de laboratoire afin d'obtenir des mesures précises et reproductibles. Un travail particulier de mise en place des protocoles a néanmoins permis de maintenir des tâche assez naturelles afin de ne pas induire des productions trop artificielles.

Les deux premières études s'intéressent à la production conjointe de gestes manuels et de parole contenant de la focalisation. Plusieurs types de gestes sont comparés (geste de pointage, geste de battement et geste d'appui sur un bouton). Il est montré que la production de focalisation attire le geste manuel quel que soit son type mais que l'attraction est plus « précise » et fine pour le pointage. Par ailleurs, l'apex du geste de pointage semble être cooccurent à une cible articulatoire plutôt qu'acoustique. La seconde étude manipule le lien de désignation le geste de pointage et la parole. Elle montre, en exhibant deux stratégies adoptées par les participants, la complexité des mécanismes mis en jeu dans cette coordination.

Finalement, une troisième étude s'intéresse à la coordination dans une tâche interactive et collaborative plus naturelle. Les résultats montrent une cooccurrence de la partie du geste qui montre avec l'information qui lui est complémentaire en parole. La perturbation de l'interaction par un bruit ambiant modifie les productions : la parole subit un effet Lombard classique et la production de gestes semble adaptater de la durée de la partie du geste qui montre à l'allongement de la parole.

Ce mémoire propose enfin une exploration des procédés d'annotation multimodaux mis en place pour l'annotation de tâches semi-contrôlées mais applicables à des cas plus généraux. Le manuscrit se conclut par une mise en perspective des résultats pour l'amélioration de certains modèles de production conjointe gestes manuels/parole et fournit quelques pistes utilisables dans le domaine des agents conversationnels ainsi que pour la détection de pathologies.

Mots-clefs

multimodalité, gestes manuels, parole, désignation, pointage, battement, Optotrak, video, interaction, coordination

---

Abstract

The work presented in this manuscript aims at describing the coordination between manual gestures and speech during the production of multimodal utterances. Since it is possible to designate using either manual gestures or speech (or both), the coordination is studied within this framework. All the experiments presented have been conducted in a lab-environment in order to obtain precise and reproducible data. Nevertheless, substantial thinking about experimental setups have been done so as to end up with rather natural tasks and try to avoid "fake" multimodal productions.

The first two studies shed light on the coordination when producing manual gestures and narrow-focused speech. Various types of manual gestures have been taken into account (pointing, beats, button-press). It is shown that focus production attracts manual gesture production whichever its type. Moreover, pointing gesture is more precisely and more consistently attracted than the other types of gestures. Interestingly, manual gestures seems to be coordinated with articulatory targets rather than acoustic ones. The second study modifies the "designation" link that unites pointing gesture and speech. It shows the great complexity of mechanisms underlying coordination by raising two strategies of coordination used by participants.

The last experiment uses a more natural interactive and collaborative task. Results show a cooccurrence in production between the part of manual gesture that shows and its complementary information found in speech. Moreover, hindering the interaction using noise has an influence on multimodal productions: a classical Lombard effect is observed for speech and the part of gesture that shows is lengthened by a duration close to the speech lengtening.

Finally, this manuscript presents an overview of the annotation process that was set up to label the aformentionned tasks. This annotation process can be generalized and thus constitutes an important part of the work. The conclusion put into perpective the results found in order to enhance existing models of manual gestures/speech coproduction and suggests applications in fields such as embodied conversational agents or pathology detection/rehabilitation.

Keywords

multimodality, manual gestures, speech, designation, pointing, beats, Optotrak, video, interaction, coordination