Un androïde doué de parole
Robotique Cognitive de la Parole et du Langage

J.L. Schwartz(1), P. Bessière(2), B. Davis(3), L.J. Boë(1)
J. Serkhane(1), C. Lalevée(1), A. Vilain(1), N. Vallée(1) , J. Diard(2), C. Matyear(3)

(1) ICP, Grenoble (2) Laplace-SHARP, Gravir, Grenoble (3) Université d'Austin, USA

Notre recherche dans la seconde année du projet *« Un androïde doué de parole - Robotique Cognitive de la Parole et du Langage »* s'est organisée essentiellement sur trois flancs. D'une part, les trois partenaires du projet ont consolidé les acquis de la première année autour des « premiers 'pas' d'un bébé robot parlant », avec des travaux sur l'exploration et l'imitation dans les premiers mois de la vie, qui ont été finalisés et sont en cours de publication, co-signées bien sûr, dans des congrès, colloques et revues internationales (Section 1). D'autre part, l'ICP et l'Université d'Austin ont entrepris une série de recherches sur « la maîtrise sensori-motrice d'un humain parlant : données sur l'enfant et sur l'adulte » (Section 2). Il s'agit ici de prendre un temps d'avance sur le bébé robot, en explorant comment son modèle humain règle le délicat problème de l'apprentissage progressif de la maîtrise de son conduit vocal, et ce à la fois par des données d'observation de bébés humains en situation, et également par une expérience originale – une première, en réalité – sur la mise en évidence des compétences d'un sujet adulte à déterminer les variables motrices adéquates à partir du son (donc du « timbre ») d'une voyelle. Enfin, l'équipe Laplace-Gravir poursuit son développement de la théorie de la « programmation bayésienne des robots » (Section 3).

Ces recherches, portant à la fois sur les données expérimentales sur l'humain et sur les outils de programmation du robot, préparent la dernière phase de notre projet, qui sera centrée sur la production des syllabes consonne-voyelle, et visera à déterminer comment un robot parvient à acquérir la maîtrise de la coarticulation (production en parallèle d'objectifs séquentiels), quelles représentations sensori-motrices il apprend, et aussi, peut-être, quels choix de production ferait un ensemble d'agents sensori-moteurs en interaction (ou comment la robotique de la parole pourrait, dans un paradigme de vie artificielle, nous faire progresser sur la structure du langage et ses tendances universelles) (Section 4).

1. Les premiers 'pas' d'un bébé robot parlant

Dans le cadre de ce projet qui réunit des chercheurs en phonologie, phonétique et sciences de la parole, psychologie expérimentale et robotique, nous nous posons le problème de l'apprentissage, à partir d'informations sensorielles (auditives, visuelles, proprioceptives, tactiles) et motrices, de structures phonologiques par un « androïde bébé », système sensori-moteur en croissance. Notre approche théorique passe par la définition d'un cadre de modélisation quantitative des mécanismes de mise en relation entre perception et action : la robotique de la parole. En effet, la robotique est le cadre naturel de modélisation et d'étude des relations entre variables sensorielles et motrices ; le noyau central en est la mise en œuvre de processus de navigation entre espaces moteurs et espaces sensoriels permettant de résoudre des problèmes complexes d'optimisation. Nous cherchons donc à programmer un *« androïde bébé »,* agent sensori-moteur doté d'actionneurs de la parole (un conduit vocal), de capteurs auditifs, visuels et orosensoriels et de mécanismes de contrôle et d'apprentissage, puis à le mettre en situation d'apprendre à parler, c'est-à-dire d'acquérir contrôles et représentations phonologiques à partir de signaux externes. Cet agent nous fournit un support théorique et un cadre de modélisation servant de système d'analyse quantitative des données de l'acquisition de la phonétique et de la phonologie.

Un parcours développemental

Deux mécanismes seraient requis à l'ontogenèse de la parole : une exploration des capacités sensori-motrices du conduit vocal, et l'imitation qui oriente les productions propres vers celles de la langue de l'environnement, le but étant d'être finalement compris. Les premiers essais en robotique du développement de la parole se basent sur l'hypothèse selon laquelle le nourrisson explorerait de manière exhaustive son espace articulatori-acoustique de réalisation, avant d'en sélectionner les zones « efficaces » pour communiquer avec les congénères de son environnement (Bailly, 1997). Ainsi, a-t-on supposé qu'au début l'enfant produit tous les sons possibles, dont ceux de toutes les langues du monde (en accord avec Jakobson, 1968). Or, la simple observation de sujets (réels) montre qu'il n'en est rien (Kent & Miolo, 1995) : les nourrissons, quelle que soit leur langue ambiante, ne produisent qu'un certain sous-ensemble des réalisations potentielles de l'appareil sensori-moteur dont ils héritent phylogénétiquement. Que sait-on de leurs capacités sensori-motrices précoces ? (a) <u>À la naissance</u>, ils sont capables d'imiter trois gestes, à partir de leur vision : la protrusion de la langue, celle des lèvres et l'abaissement de la mandibule (Meltzoff, 2000). Ces mouvements, recrutés dans la parole adulte, ne sont pas forcément liés à son développement, mais n'en sont pas moins disponibles avant les premières vocalisations. (b) <u>À quelques semaines</u>, les nourrissons vocalisent. De plus, ils sont à même d'orienter leurs productions vers des sons vocaliques fréquemment perçus (imitation vocale précoce) (Kuhl & Meltzoff, 1996), et de détecter la congruence entre le son d'une voyelle et l'image du visage en mouvement qui l'a produite (intégration multimodale) (Kuhl & Meltzoff, 1982). (c) <u>Vers 7 mois</u>, ils babillent : leur mandibule se lance dans des cycles d'élévation-abaissement, alors que leurs cordes vocales vibrent. C'est le Babillage Canonique (BC) (Koopmans-van Beinum & Van Der Stelt, 1986 ; MacNeilage & Davis, 1990). (d) <u>Par la suite</u>, les individus « apprennent » à contrôler, plus ou moins successivement, la rythmicité mandibulaire, les mouvements des articulateurs portés par ce cycle indépendamment les uns des autres, la forme globale de l'espace libre à l'intérieur de leur conduit, pour atteindre au final la maîtrise des sons de leur(s) langue(s) ambiante(s) (Vilain et al., 2000).

Notre objectif dans ce travail a été de faire parcourir à un bébé robot parlant les étapes parcourues par le bébé humain entre 0 et 7 mois, autour de deux paradigmes principaux : exploration des capacités sensori-motrices (apprentissage partie d'un modèle direct) et imitation vocale (apprentissage partie d'un modèle inverse).

Bébé robot et bébé humain

Le modèle GROWTH

Pour apprécier le jeu de configurations articulatoires dont les nourrissons disposent pour vocaliser et babiller, nous avons utilisé un modèle articulatoire statistique qui intègre la croissance du conduit vocal (Goldstein, 1980) : GROWTH (Boë, 1999). Son noyau computationnel (Maeda, 1990) est issu de l'analyse statistique de coupes sagittales cinéradiographiques d'un conduit vocal adulte en action, d'où 7 facteurs se dégagent. Ces descripteurs de la forme du conduit correspondent à des actions musculaires concrètes : ce sont les degrés de liberté d'un conduit vocal virtuel qui servent d'entrée au modèle. La sortie est une description de la forme du conduit ainsi que les formants du son résultant.

La programmation bayésienne des robots

Les compétences du système sensori-moteur sont modélisées par la distribution de probabilité conjointe des paramètres articulatoires et perceptifs qui le constituent. La décomposition de cette distribution fournit une structure statistique exprimant les relations d'interdépendance entre ses variables motrices et sensorielles. Ces relations perception-action se basent sur des considérations d'indépendance conditionnelle entre variables et s'expriment, essentiellement, dans le choix de la décomposition de la distribution conjointe. Le modèle s'ancre dans le « réel » via un ensemble de données d'apprentissages qui permettent d'instancier les distributions de probabilité figurant dans sa loi de décomposition (ou *structure de dépendance*). Le robot résultant est interrogé par l'intermédiaire de sa *description* effective. Les performances de l'androïde sont intrinsèquement liées à son *expérience* et à la *« finesse » de sa structure.*

Données de phonétique développementale

On dispose de deux jeux de données issues d'études en phonétique du développement. Le premier correspond à des pseudo-voyelles de pré-babillage produites par des sujets de 4 mois, au cours de tests sur l'imitation vocale précoce (Kuhl & Meltzoff, 1996), le second à des sons vocaliques produits par des bébés de 7 mois, en début de babillage canonique (Davis & MacNeilage, 1995). Dans chaque cas, les deux premiers formants et la description phonétique sont disponibles.

Exploration

Pour estimer les capacités articulatoires avant et au début du BC, nous avons développé 3 méthodes d'analyse : le cadrage acoustique, le cadrage articulatoire et le cadrage géométrique.

Cadrage acoustique

Tous les sons oraux que peut générer le modèle articulatoire s'inscrivent dans l'Espace Vocalique Maximal *(EVM)* (Boë et al., 1989). Il représente ce qu'un enfant du même âge pourrait produire s'il utilisait toutes les commandes de son système effecteur : dans le plan (F1,F2), c'est le triangle vocalique dont les sommets sont les voyelles [i a u]. Le cadrage acoustique consiste en la simple superposition des données formantiques réelles à l'EVM de GROWTH au même âge. Ainsi pouvons-nous tester si les vocalisations réelles appartiennent à cet EVM et *estimer le domaine de l'espace acoustique exploré par les enfants de 4 et 7 mois.*

Les vocalisations réelles s'inscrivent dans l'EVM correspondant : elles font partie de l'espace de réalisation de GROWTH au même âge. De plus, les données réelles ne recouvrent pas tout l'espace qu'elles pourraient occuper si elles étaient le produit d'un système moteur mature. En particulier, les vocalisations de 4 mois sont relativement centrales et mi-hautes : les productions les plus en avant, arrières et ouvertes ne semblent pas explorées. À 7 mois, les productions vocaliques exploitent plus la dimension haut-bas qu'au stade précédent (Fig. 1).

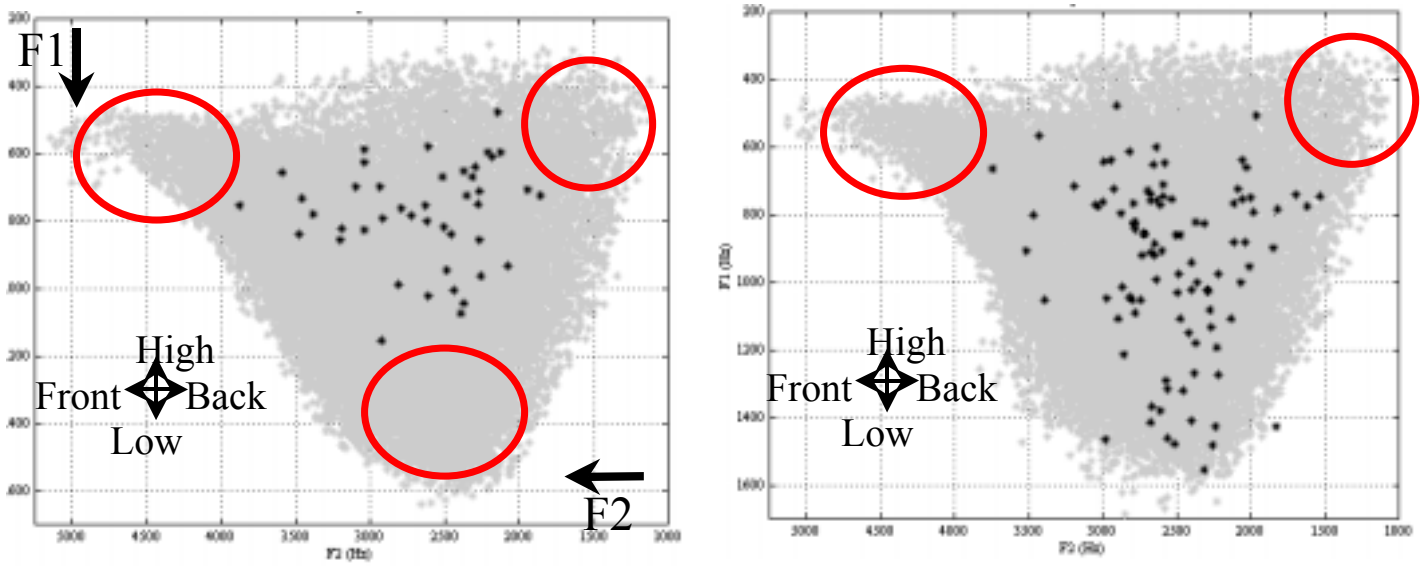


Figure 1 *– A 4 mois (à gauche), les voyelles sont centrales et mi-hautes ; à 7 mois (à droite), elles exploitent plus la dimension haut-bas.*

Cadrage articulatoire

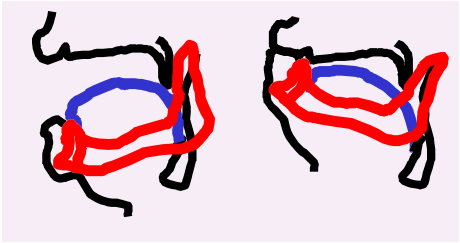
Certaines régions de l'EVM, généré en utilisant les 7 paramètres articulatoires, ne figurent pas dans les données réelles. Le cadrage articulatoire permet d'apprécier les capacités motrices des nourrissons en contraignant les variables motrices de GROWTH. Le but de cette méthode est d'estimer le jeu minimal de degrés de liberté requis pour reproduire les sons vocaliques observé. Selon un critère probabiliste, nous sélectionnons parmi plusieurs jeux possibles celui dont les réalisations s'ajustent le mieux aux données acoustiques réelles : on parlera de « sous-modèle gagnant ».

Il faut au moins trois paramètres articulatoires à 4 mois pour couvrir les données réelles. Le sous-modèle gagnant exploite les variables motrices de séparation des lèvres (*Lh*), du corps (*Tb*) et du dos (*Td*) de la langue, avec une exploration plus large de Td (cambrure-applatissement) que de Tb (protrusion-rétraction). D'après l'ensemble des modèles testés, le dos de la langue, Td, jouerait un rôle important dans l'explication des vocalisations réelles. Cela pourrait être mis en relation avec son activité lors de la *succion*. Notons que le mouvement mandibulaire ne semble pas recruté à ce stade de développement. A 7 mois, le sous-modèle qui s'ajuste le mieux aux données est le même qu'à 4 mois hormis l'ajout de la commande mandibulaire (*J*) et une exploitation plus large de Tb. La nécessité du paramètre mandibulaire est congruente avec le rôle fondamental de cet articulateur « pour » babiller.

Cadrage géométrique

Les deux modèles gagnants issus du cadrage articulatoire permettent d'estimer les configurations linguales, sachant les vocalisations observées à 4 et 7 mois. Le cadrage géométrique est une méthode d'inversion exhaustive : chaque vocalisation réelle mène à un ensemble de formes de conduit, produites par le modèle vainqueur et acoustiquement plausibles. La géométrie du conduit est décrite dans les systèmes suivants [BGP95] : (i) les coordonnées du point le plus haut de la langue (Xh, Yh) dans un référentiel fixe, (ii) l'aire (Ac) et la localisation (Xc) de la constriction, ainsi que l'aire aux lèvres (Al). A 4 mois, les formes moyennes de conduit correspondent à des plus hauts points de la langue plutôt centrés et regroupés. Les constrictions sont légèrement antérieures et plutôt larges : la constriction ne semble pas contrôlée. A 7 mois, les vocalisations correspondent à une exploration plus étendue des positions de langue, autant sur l'axe haut-bas qu'avant-arrière.

Il est alors possible d'appliquer des mouvements de mâchoire cycliques, « à la manière » des bébés en phase de babillage canonique à 7 mois. On obtient ainsi des séquences de pseudo voyelles et de pseudo-consonnes partageant le même lieu d'articulation, ce qui correspond aux « cooccurrences » démontrées par nos partenaires MacNeilage et Davis chez les enfants au même âge (Figure 2).



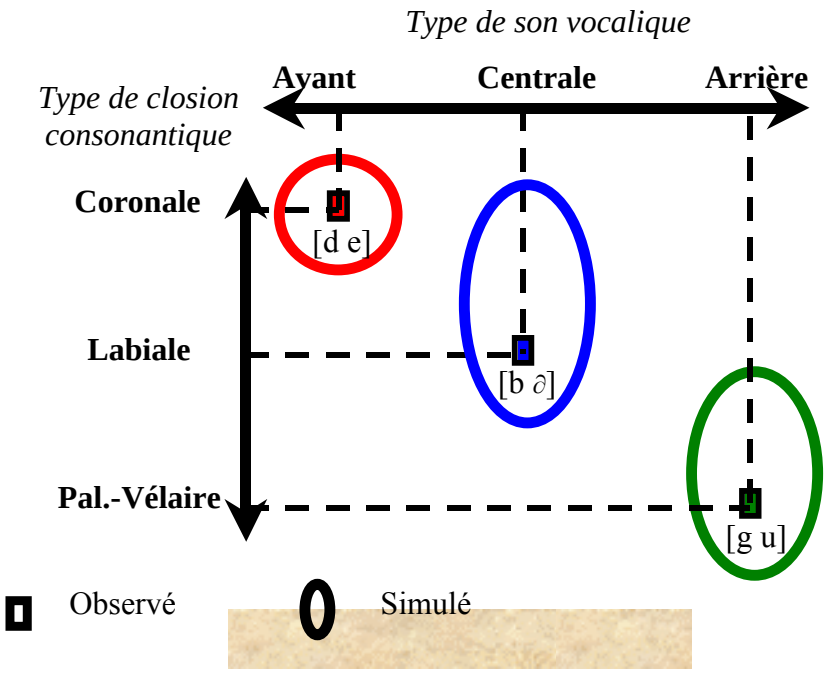


Figure 2 *– A 7 mois, un mouvement cyclique de la mâchoire (à gauche), produit sur le bébé robot des « co-occurrences » voyelles-consonnes conformes aux données (à droite)*

Imitation

Structure de dépendance statistique entre variables motrices et perceptives

Nous avons alors mis en place le modèle bayésien complet, en testant ses capacités à inférer une commande articulatoire à partir du son. Les variables motrices (M) sont les trois degrés de liberté articulatoire identifiés à 4 mois à l'étape précédente, du modèle, soit Lh, Tb, Td. Les variables perceptives combinent variables auditives (A) (les deux premiers formants F1 et F2, ou fréquences de résonance du conduit vocal, exprimés en Bark, échelle perceptive des fréquences) et visuelle (l'aire aux lèvres Al). Par définition les formants d'un son vocalique sont fonction de la forme du conduit vocal dont la coupe sagittale peut être décrite par trois variables : l'aire aux lèvres Al et les coordonnées du plus haut point de la langue, Xh et Yh. Nous avons choisi de prendre en compte ces variables géométriques (G) (qui peuvent être fournies par les capteurs tactiles du robot) car en formant un espace intermédiaire entre le moteur et l'auditif, elles permettent de réduire l'impact du problème many-to-one lors de l'inversion et de modulariser à terme une description qui comprendra plus de paramètres articulatoires.

La forme complète de la structure de dépendance de notre modèle bayésien est donnée par (voir rapport 2002, et Section 3) :

$$P(Lh \otimes Tb \otimes Td \otimes Xh \otimes Yh \otimes Al \otimes F1 \otimes F2)$$
$$= P(Xh) * P(Yh) * P(Al)$$
$$* P(Lh / Al) * P(Tb / Xh \otimes Yh) * P(Td / Xh \otimes Yh \otimes Tb)$$
$$* P(F1 / Xh \otimes Yh \otimes Al) * P(F2 / Xh \otimes Yh \otimes Al).$$

Apprentissage d'un modèle direct partiel

C'est cette structure de dépendance qu'il s'agit d'identifier par une phase d'apprentissage consistant en une exploration de vocalisations à 4 mois, puis d'utiliser en inversion par un processus classique d'inversion bayésienne où l'on cherche les variables motrices compatibles avec une sortie audio (F1, F2) donnée. Nous nous sommes alors demandés comment apprendre un tel modèle avec très peu de

données, en supposant que le bébé à 4 mois n'a que très peu vocalisé. En regard de cette question, il faut remarquer que les capacités d'imitation du nourrisson sont encore très frustres. Nous avons émis l'hypothèse que le degré de granularité de la discrétisation de nos variables pouvait être un paramètre clé dans cette recherche. Nous avons donc décidé d'explorer diverses situations, variant taille de la base d'apprentissage (BA), et discrétisation des variables géométriques, qui sont cruciales dans le processus d'apprentissage et d'inversion. On observe que, quand la taille de BA est maximale, la précision de l'imitation croît avec l'affinement de la discrétisation. Pour un profil de discrétisation donné et pour des tailles de BA supérieures à 50, l'erreur diminue avec l'augmentation de la taille de la BA, jusqu'à une certaine valeur plancher déterminée par le degré de discrétisation. Mis à part pour le profil de discrétisation 2_2_1, on note un pic d'erreur pour les tailles de BA petites. Ainsi, globalement, le système répond d'abord au hasard, puis il affine peu à peu ses réponses, (en passant éventuellement par une phase de sur-apprentissage de données non-significatives), et ce d'autant mieux, mais avec un besoin d'autant plus grand de données, que la discrétisation est fine. Le nombre de données nécessaires à un bon apprentissage (performances stabilisées à 10 % de l'erreur minimum accessible) est de l'ordre du triple du nombre total de valeurs géométriques après discrétisation.

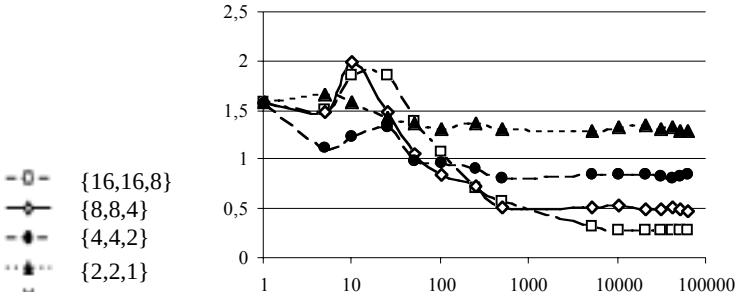


Figure 3 *– Apprentissage du modèle direct*
La taille de BA est en abscisse, la discrétisation selon Xh, Yh, Al en paramètre (par ex., 4_4_2 signifie 4 valeurs possibles pour Xh, 4 pour Yh, 2 pour Al). En ordonnée, erreur moyenne audio en Bark.

Performances d'imitation

Nous avons alors simulé l'expérience l'expérience de Kuhl et Meltzoff, en présentant au système, pour diverses valeurs de BA et de discrétisation, des entrées typiques de [i], [a] et [u] adultes, et en catégorisant les réponses émises par le processus d'inversion, respectivement comme des [i], des [a], et des [u], par un classifieur gaussien opérant sur les sorties acoustiques. Un grand nombre de tests ont été effectués, variant les conditions d'apprentissage, et contrastant des conditions d'imitation visuelle pure, auditive pure et audio-visuelle (Serkhane et al., 2003). Il en ressort que l'imitation ne peut être purement visuelle, car les matrices de erreurs du bébé robot sont dans ce cas incohérentes avec les erreurs du bébé humain. Ainsi, un premier résultat important de cette étude est que les simulations confirment l'existence d'un réel mécanisme **d'imitation audio-articulatoire** à l'œuvre chez les bébés de Kuhl et Meltzoff. D'autre part, les résultats (Figure 4) montrent que, pour des modèles « grossiers » opérant sur un petit nombre de valeurs géométriques, on peut obtenir des performances d'inversion assez bonnes : ainsi, 10 vocalisations en apprentissage suffisent au modèle grossier 2_2_1 pour fournir 60% de performance correcte, soit un score équivalent à celui des bébés de l'expérience ! Pour un modèle plus fin (4_4_2, 8_8_4), il faut monter à 50 vocalisations. Par contre, un modèle « trop fin » (16_16_8) ne peut opérer efficacement sur des données d'adultes qui échappent à l'espace d'apprentissage : on retrouve là les problèmes classiques de sur-apprentissage et de dilemme biais-variance. La conclusion forte que nous tirons de ces expériences, c'est la capacité de ce type de modèle à « imiter » aussi bien que les bébés humains, avec peu de données et un degré de discrétisation relativement grossier, en jouant sur trois facteurs : taille de la base d'apprentissage, degré de discrétisation et distribution des variances dans les données d'apprentissage.

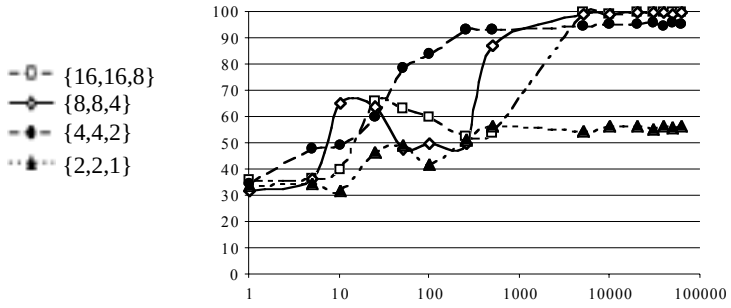


Figure 4 *– Performances d'imitation du « bébé robot »*
La taille de BA est en abscisse, le pourcentage d'imitation correcte est en ordonnée (voir texte).

Publications de l'équipe

(toutes ces publications sont cosignées par des membres d'au moins deux laboratoires partenaires du projet)

Schwartz, J.L., Boë, L.J., & Bessière, P. (submitted). Speech robotics, a computational framework for the study and modelling of speech acquisition and evolution. *Evolution of Communication*

Serkhane, J.E., Schwartz, J.L., Boë, L.J., Davis, B.L., & Matyear, C.L. (submitted). Characterization of infants' early articulatory exploration with an articulatory model integrating the non-linear growth of the vocal tract. *Journal of Phonetics*

Serkhane, J., Schwartz, J.L., Boë, L.J., Davis, B., Matyear, C. (2002). Motor specifications of a baby robot via the analysis of infants' vocalizations. Proc. *ICSLP'2002*, Denver, Colorado, 45-48.

Serkhane, J., Schwartz, J.L., Boë, L.J., Davis, B., Bessière, P., Mazer, E. (2002). Etude comparative de vocalisations de bébés humains et de bébés robots**.** Actes *XXIVèmes Journées d'Étude sur la Parole*, 149-152.

Serkhane, J., Schwartz, J.L., & Bessière, P. (2003). Simulating vocal imitation in infants, using a growth articulatory model and speech robotics. Proc. ICPhS, Barcelona.

Cette recherche est focalisée sur la thèse de Jihène Serkhane, co-encadrée par J.L. Schwartz (ICP), P. Bessière (Laplace) et L.J. Boë (J.L. Schwartz), et qui a également eu l'occasion de passer 15 jours à l'Université d'Austin en juillet, pour poursuivre le dialogue avec B. Davis et C. Matyear, et travailler sur la publication soumise au J. Phonetics.

2. Maîtrise sensori-motrice d'un humain parlant : données sur l'enfant et sur l'adulte

A. Le babillage comme base du développement de la parole

Afin d'émettre des hypothèses phylogénétiques et ontogénétiques sur le déroulement de l'acquisition de la parole, les chercheurs observent et étudient les mécanismes vocaux, moteurs, cognitifs et communicationnels qui entrent en compte dans l'émergence du langage chez l'enfant. Les nombreuses données et observations actuelles s'insèrent dans la lignée des idées de Piaget sur le rôle crucial des actions sensori-motrices dans le développement cognitif, ainsi que dans celles de Bruner et Vygotsky qui soutiennent l'importance des interactions sociales.

Les études actuelles s'attachent à donner une explication articulatoire à ce développement ontogénétique et à le décrire comme la maîtrise progressive des articulateurs. Peter MacNeilage a proposé, en 1998, la théorie *Cadres, puis Contenu* (*Frames, then Content*) de l'évolution de la production de parole qui cherche à donner une description de la structure de la parole humaine qui rende compte à la fois de son développement ontogénétique et phylogénétique. Le babillage canonique, qui marque chez l'enfant l'apparition vers sept mois de séquences proto-syllabiques rythmées, est considéré dans cette théorie comme l'émergence de la structure de base de la parole humaine, avec la maîtrise de l'oscillation de l'articulateur porteur, la mandibule. Cette structure sous-jacente à toute production de parole humaine est appelée *Cadre*, et sera complétée dans la suite du développement ontogénétique par un *Contenu* segmental dû au contrôle des articulateurs portés: la langue et les lèvres. Les simulations présentées dans la section précédente confortent l'idée que le cycle mandibulaire soit responsable en grande partie des productions du babillage.

Cependant, si le contrôle des oscillations mandibulaires peut bien être considéré comme la structure de base de la parole, ce cadre proto-syllabique ne sera pas adapté aux besoins des adultes humains, tant que le bébé ne sera pas capable de contrôler globalement son conduit vocal, de la glotte aux lèvres, afin d'obtenir un contraste acoustique efficace à l'intérieur de la syllabe. L'étude que nous proposons porte sur deux types de contrôles non physiologiquement liés à la mandibule, et qui permettent à l'enfant la maîtrise de cadres proto-syllabiques de plus en plus proches des normes adultes, donc des syllabes. D'une part celui du vélum, que l'on déplace pour ouvrir ou empêcher le passage de l'air du conduit vocal vers les fosses nasales, et qui permet de disposer d'un conduit vocal oral comme dans la grande majorité des productions adultes, pour produire des séquences consonne orale-voyelle acoustiquement bien saillantes. D'autre part, le contrôle fondamental qu'est celui de la vibration des cordes vocales. Le principe de base de la communication orale humaine est, en effet, la synchronisation entre mouvements glottiques (cordes vocales) et mouvements supra-glottiques (mandibule, lèvres, langue, velum). Le délai d'établissement du voisement (VOT) est un indice acoustique qui permet de rendre compte du contrôle de cette coordination oro-laryngée.

Méthodologie

La chronologie du développement de ces contrôles est ici étudiée à partir d'un suivi audio-visuel d'enfants français âgés de six à douze mois. Les enregistrements ont commencé lorsque les parents signalaient l'apparition du babillage canonique. Les sessions se sont déroulées dans l'environnement familial des sujets. Elles avaient lieu tous les quinze jours et duraient environ quarante-cinq minutes chacune. L'intérêt du corpus audio-visuel est que les productions des enfants lors du babillage sont peu

précises car peu contrôlées. Il est donc difficile, essentiellement au début du babillage, de reconnaître les sons produits, par le seul biais de la perception auditive. La vision est ainsi une aide importante pour la désambiguïsation.
Pour notre étude, nous avons utilisé les productions de deux enfants, un de sexe féminin et l'autre de sexe masculin, tous deux de langue maternelle française. Notre corpus comprend vingt sessions couvrant les productions des enfants de 6 mois et 25 jours à 12 mois et 4 jours pour l'un et de 7 mois et15 jours à 12 mois et 3 jours, c'est-à-dire du début du babillage aux premiers mots. Les stimuli ont été systématiquement numérisés, segmentés, puis annotés phonétiquement à partir de critères acoustiques et perceptifs, et ce par analyse croisée par deux expertes phonéticiennes.

Résultats principaux

Notre étude sur le développement du contrôle de la production de parole nous a permis d'observer des étapes de l'évolution du cadre proto-syllabique vers la syllabe.
D'abord, nous disposons désormais de deux corpus longitudinaux de productions d'enfants de sept à douze mois, complètement transcrits sous une forme qui permet de mener très facilement d'autres analyses. Les différentes observations des espaces vocaliques, des variations ainsi que celles des co-occurrences dans le babillage, nous ont montré, que les sujets semblaient faire progresser leurs productions en variant de plus en plus , au cours des mois, les closants et les vocants.

Contrôle du velum

Nous avons pu voir aussi que le contrôle d'un conduit vocal avec le velum relevé représentait une étape significative dans le développement de la parole, chronologiquement situé entre l'émergence des cyclicités mandibulaires et le développement des segments du corps de la parole à travers le contrôle des articulateurs portés tels que la langue et les lèvres. Ce contrôle représente un élément central dans la maîtrise progressive d'un contrôle global des cadres protosyllabiques, dont une autre étape est le développement de la coordination orolaryngée. Les données de notre corpus indiquent une diminution quantitative des productions nasales à partir du sixième mois (Figure 5). La proportion des nasales produites par l'enfant vers 11 mois correspond alors à celle de la langue maternelle, ce qui nous permet de supposer que l'activation du velum est contrôlée à ce stade.

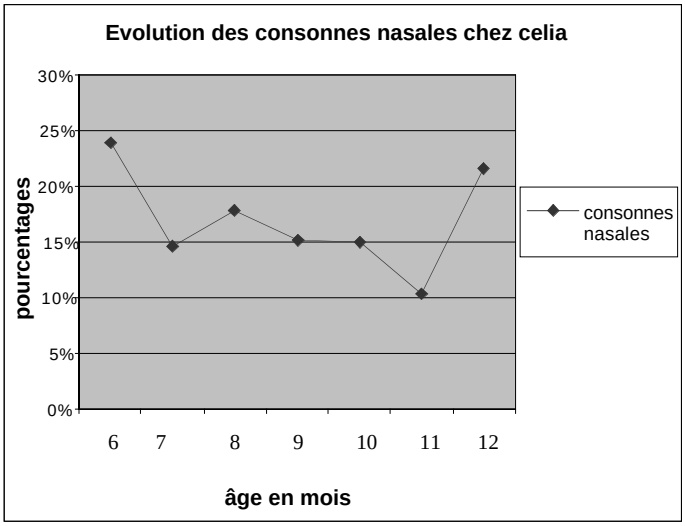


Figure 5 – : évolution des occlusives nasales de 6 à 12 mois pour Célia.
On observe entre six et onze mois une diminution de la proportion des consonnes nasales: de 25% à 10%, avec un minimum à onze mois à environ 10%. On peut expliquer le "pic" de nasales à douze mois par le fait que Célia semble acquérir ses premiers mots. Le contexte situationnel que l'on peut déduire des vidéos indique qu'elle utilise les mots "non" et "maman". C'est pourquoi elle produit un grand nombre de séquences CV: coronale nasale-voyelle nasale s'apparentant à [nɔ̃], et de CVCV bilabiale nasale/voyelle s'apparentant à [mamã].

Contrôle de la coordination source vocale – conduit vocal

En ce qui concerne la coordination orolaryngée (entre source vocale et conduit vocal), nous avons étudié un indice majeur de coordination , le « VOT ». Le VOT (Voice Onset Time) ou délai d'établissement du voisement est défini comme l'intervalle de temps entre le début des pulsations périodiques régulières glottiques et la détente de l'occlusion supra-glottique (Lisker et Abramson, 1964). Par convention, le VOT est négatif lorsque la mise en vibration des cordes vocales intervient avant la détente de l'occlusion (prévoisement) et il est positif lorsque cette mise en vibration commence après la détente. Dans la plupart des langues, la production du contraste entre plosives voisées (comme /b d g/) et non voisées (comme /p t k/) fait intervenir la relation temporelle entre le départ des vibrations laryngées et la détente de l'occlusion supraglottique. La localisation des catégories d'occlusives sur l'échelle VOT dépend de la langue (Serniclaes, 1987).

Nos données (Figure 6) montrent que chez nos deux sujets, il y a une nette séparation entre les valeurs négatives et les valeurs positives du VOT, ce qui indique, que les classes voisées et non-voisées sont bien distinctes chez les sujets. On peut voir une concentration des productions entre –60 ms et –20 ms pour les VOT négatifs et entre 15 ms et 50 ms pour les VOT positif chez nos deux sujets. Si on compare ces résultats à ceux de Serniclaes (1987), il semblerait que les enfants aient des valeurs de VOT un peu plus courtes que celles des adultes pour le VOT positif (≈15ms chez l'enfant pour ≈30ms chez l'adulte) et des valeurs bien plus courtes pour le VOT négatif (≈ -30ms chez l'enfant pour ≈ -130ms chez l'adulte).

On peut donc supposer que le contrôle du VOT est déjà présent à six mois chez les enfants, mais qu'il n'est pas encore contrôlé comme pour les normes adultes.

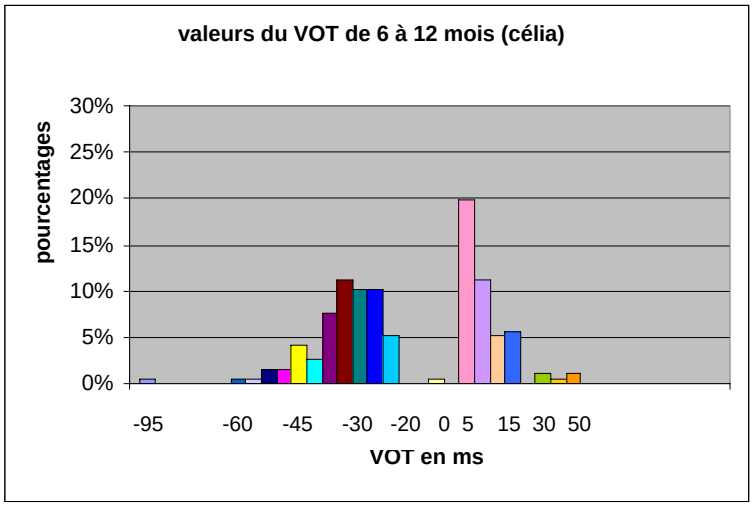


Figure 6 – Evolution du VOT de 6 à 12 mois pour Célia.

Conclusion

Les productions du babillage sont donc explicables, en grande partie, par la dominance du cadre mandibulaire comme le propose MacNeilage, mais ce contrôle des cadres se développe en parallèle avec d'autres éléments qui vont permettre à l'enfant de produire des syllabes qui respectent les contraintes de la communication adulte : le contrôle actif du velum et le contrôle de la coordination orolaryngée.

Ce travail a fait l'objet du stage de DEA Sciences du Langage de Claire Lalevée, sous la direction d'Anne Vilain (ICP), avec des discussions fréquentes avec Barbara Davis (Austin).

Lalevée, C. & Vilain, A., 2003. Development of speech frame control: a longitudinal study of the oral/nasal control. *Proceedings of the 15th International Congress of Phonetic Sciences*, Barcelona.

B. Les inférences sensori-motrices chez l'adulte

La perception de la parole est selon nous l'ensemble des processus perceptifs (auditifs, visuels, voire tactiles) permettant, au niveau segmental, de récupérer et de contraindre les cibles et les phases des gestes vocaux, fournissant par là même un ensemble de représentations pour le contrôle de ses propres actions (leur spécification), pour le suivi des actions de l'autre (leur récupération via leur mise en forme) et pour la morphogenèse des unités du langage (des contraintes et pressions périphériques sur l'émergence d'une phonologie). Le programme de recherche qui s'ensuit est celui de la détermination des traitements et représentations permettant le contrôle de l'action (spécification / récupération / mise en forme) et plus précisément la question que nous nous sommes posés ici est : comment un auditeur parvient-il à estimer les commandes articulatoires associées à un son donné ?
D'un point de vue perceptif, les travaux de Ladefoged (1967) montrent que des experts phonéticiens produisent des jugements auditifs cohérents sur les traits vocaliques de hauteur et d'antériorité, mais pas sur le trait d'arrondissement, ce dernier point étant confirmé par Lisker et Rossi (1992) montrant l'incapacité d'experts phonéticiens à estimer correctement par l'oreille l'arrondissement des voyelles hautes ou mi-hautes, particulièrement lorsqu'elles sont centrales, soulevant du même coup la problématique du degré de fiabilité de certains choix dans les descriptions phonologiques et phonétiques.
Dans ce contexte, nous avons imaginé un paradigme original visant à déterminer la capacité d'auditeurs experts ou naïfs à estimer deux contrôles majeurs pour la production des voyelles : la hauteur de la langue (dimension « verticale ») et son antériorité (dimension « horizontale »).

Estimation auditive de la dimension verticale de la position de la langue

Méthodologie

Sur le modèle articulatoire qui fournit la base du robot parlant, nous avons défini 38 prototypes, couvrant dans le triangle vocalique les différentes positions des voyelles des langues du monde (Vallée et al., 1995). Nous avons sélectionné 24 voyelles représentant 4 séries (antérieures non arrondies ou arrondies, centrales non arrondies, postérieures arrondies) et 6 degrés de hauteur par série, puis construit 40 paires de stimuli, contrastant deux degrés de hauteur consécutifs, au sein d'une série ou de deux séries « voisines » antérieures non arrondies vs. arrondies ou antérieures non arrondies *vs.* centrales non arrondies (Fig. 7). Nous avons alors présenté ces différentes paires, en ordre aléatoire et en variant aléatoirement l'ordre des deux stimuli au sein d'une paire, et demandé à des auditeurs de déterminer quel était pour chaque paire le stimulus associé à la position la plus haute de la langue. La question majeure que nous nous posions était de savoir si les sujets seraient capables d'effectuer cette tâche, c'est-à-dire de ne pas répondre au hasard. En début d'expérience, nous expliquions brièvement aux sujets naïfs que la paire [i-a] contraste deux positions de langue, tandis que la paire [i-y] ne contraste pas la langue mais les lèvres ; puis nous présentions une paire [i-ae] contrastant deux niveaux extrêmes de hauteur au sein d'une même série, et nous vérifiions que les sujets avaient bien compris la tâche, et qu'ils répondaient correctement.
Nous avons étudié trois types de sujets : 8 sujets phonéticiens avec au moins 7 années d'expérience (G1), 14 sujets étudiants ou chercheurs en phonétique-phonologie avec moins de 4 années d'expérience (G2), et 31 sujets absolument naïfs ne connaissant rien aux travaux sur la parole et la phonétique (G3).

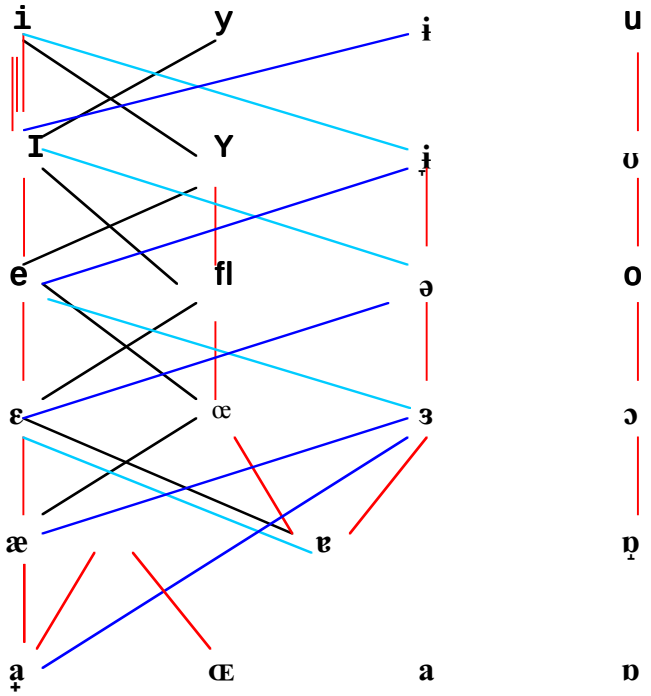


Figure 7 – Stimuli et paires impliquées dans l'expérience

Résultats

Les scores moyens de bonne réponse sont respectivement de 85% (G1), 71% (G2) et 64.5% (G3) avec des écarts types importants dans les groupes G2 (12%) et G3 (9%). Ceci montre que l'expérience compte beaucoup dans cette tâche (les scores moyens des 3 groupes sont significativement différents) mais aussi, ce qui est essentiel, qu'il est possible d'effectuer la tâche avec des scores significativement supérieurs au hasard (ce qui est le cas pour les 3 groupes) *même en l'absence de toute expérience « phonétique » consciente*, et ce bien que les stimuli soient acoustiquement proches, nous le verrons.

La seconde question que nous nous posions était de savoir s'il était plus facile d'effectuer la tâche pour deux stimuli d'une même série que pour deux stimuli différant non seulement d'un degré de hauteur, mais aussi par l'arrondissement ou l'antériorité. Il apparaît qu'il n'y a pas de différence significative entre ces deux conditions, et que l'intervention d'un contraste supplémentaire non pertinent ne gène pas dans l'exécution de la tâche. Nous avons également vérifié que la tâche n'était pas significativement plus difficile pour des séries « inconnues » des sujets naïfs comme la série centrale, que pour des séries partiellement connues comme les antérieures ou les postérieures arrondies.

Enfin, la dernière question était celle des corrélats acoustiques. Il apparaît que les variations du premier formant F1 (qui est classiquement le principal corrélat de l'aperture, voir Jakobson, Fant et Hallé, 1952) autant que celles du deuxième formant F2 contribuent à la réalisation de la tâche. Nous avons calculé pour le groupe G3 des sujets naïfs les coefficients de corrélation entre écarts de formants sur une échelle perceptive semi-logarithmique de Barks[1], et nombre de réponses correctes. Nous avons d'abord mis en évidence un effet néfaste des trop grands écarts en F2 entre les deux stimuli d'une paire, qui font chuter les scores, et les ramènent souvent au niveau du hasard. Si l'on écarte les paires caractérisées par un écart en F2 supérieur à 1 Bark en valeur absolue, on obtient une bonne corrélation entre écart en F1 et nombre de réponses correctes (Fig.8), avec un coefficient de corrélation de 0.65, tandis que F2 n'apparaît plus comme un corrélat acceptable. On observe qu'une distance de l'ordre de 1.2 Barks en F1 suffit à produire un pourcentage de réponses correctes de l'ordre de 50% au-dessus du hasard. Or une telle distance acoustique est faible, inférieure en tout état de cause à celle qui sépare les voyelles hautes et mi-hautes ou mi-hautes et mi-basses en français (de l'ordre de 1.5 Bark).

[1] Bark=7 ArgSh(Hz/650), formule de Schroeder et al., 1979

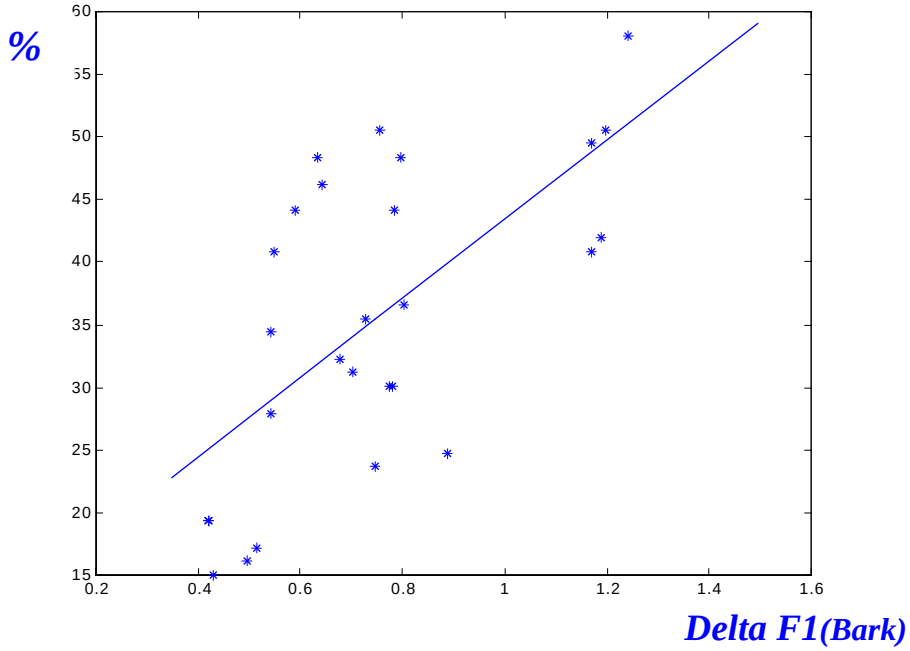


Figure 8 – Relation entre paramètre physique caractéristique de la paire et réponse des sujets
Ecart en F1 en Bark entre les deux stimuli de la paire (abscisse) et pourcentage de bonnes réponses corrigé du hasard (0% pour une réponse au hasard, 100% pour toutes les réponses correctes pour tous les sujets) (ordonnée)

Estimation auditive de la dimension horizontale de la position de la langue

Méthodologie

Pour tester la dimension avant-arrière de la langue, nous avons sélectionné 18 voyelles sur 6 séries ((antérieures, centrales, postérieures) * (non arrondies et arrondies)) et trois degrés de hauteur sur chaque série (haute, mi-haute, mi-basse). Puis nous avons construit 92 paires de voyelles différant de 1 ou 2 degrés sur l'échelle avant-arrière, avec 3 répétitions de chaque paire dans un ordre aléatoire (276 paires au total). Avec le même protocole que précédemment, la tâche était de déterminer auditivement quelle voyelle d'une paire donnée était la plus antérieure. L'expérience a été faite avec 23 sujets français naïfs, extraits du groupe G3 de l'expérience précédente.

Résultats

Les résultats sont cette fois beaucoup moins bons : en moyenne, 15% de réponses correctes en données corrigées du hasard contre 29% sur la hauteur pour les sujets naïfs. Ce résultat est cohérent avec des données qui montrent qu'il est plus facile d'estimer proprioceptivement la hauteur de sa langue, que sa position avant-arrière (Lindblom & Lubker, 1985).

Un point particulièrement intéressant est que les données perceptives sont cette fois plutôt corrélées avec F2 et non F1 (Figure 9). Ainsi, il semble que des sujets, même naïfs, soient effectivement capables d'estimer sélectivement la position de la langue sur une grille à deux dimensions, en se focalisant sur le premier formant pour extraire la dimension verticale, et sur le second formant pour extraire la dimension horizontale. Ce résultat remarquable montre qu'il semble exister une représentation sensorielle de la configuration de la langue dans le conduit vocal, accessible à partir des stimuli auditifs, avec une possible organisation somatotopique de cette représentation.

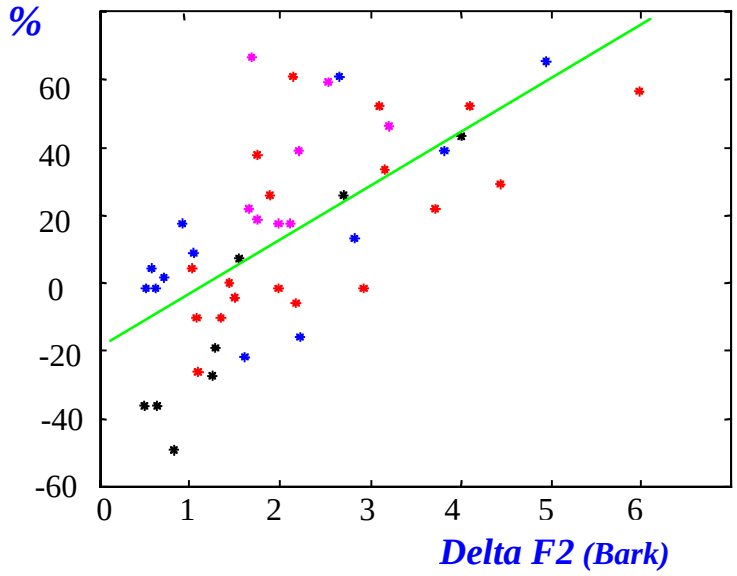


Figure 9 – Idem que Figure 8, pour la dimension horizontale de la langue

Conclusion

Selon notre hypothèse, la perception des gestes vocaux permet de retrouver des caractéristiques du contrôle de ces gestes. Le principe même est évident : sans perception auditive, pas d'acquisition correcte du contrôle de la production de la parole chez les sourds ; mais aussi, des retards pour les aveugles à l'acquisition de contrastes visibles mais peu audibles comme [m] vs. [n] (Mills, 1987); et l'incapacité à contrôler correctement son conduit vocal pour un sujet atteint d'un déficit somesthésique congénital (MacNeilage et al., 1967). Mais l'étude expérimentale fine des capacités de « percevoir les gestes de l'autre » est délicate, et peu de paradigmes sont disponibles. À travers les résultats de notre expérience, il apparaît que la capacité d'auditeurs, même naïfs, à estimer la hauteur de la langue à l'intérieur du conduit vocal est excellente et que ces mêmes auditeurs sont capables de surmonter l'existence de « sosies vocaliques » (Boë et al., 2000), c'est-à-dire des géométries différentes du conduit vocal qui fournissent en sortie le même résultat acoustique. Evidemment, il faut noter que cette dimension d'aperture est connue comme étant celle qui se soumet le plus efficacement à l'analyse perceptive, auditive mais aussi proprioceptive (Lindblom & Lubker, 1985). C'est aussi cette dimension qui est la plus recrutée par les systèmes pour contraster leurs phonèmes vocaliques (Vallée, 1994).

Publications

Vallée, N., Schwartz, J.L., Uomini, T., Boë, L.J., Kandel, S. (2002). Une étude expérimentale de l'identification perceptive du trait vocalique d'aperture. *4èmes Journées Internationales du GDR « Phonologie »*, Grenoble.

Vallée, N., Schwartz, J.L., Kandel, S. (2003). Can we recover vowel gestures from speech sounds ? *International Congress of Phonetic Sciences*, Barcelona, 3-9 August.

3. Programmation bayésienne des robots

Présentation générale

La contribution de l'équipe LAPLACE-SHARP du laboratoire GRAVIR au projet ROBEA « un androïde doué de parole » porte sur deux sujets :

1. Le savoir-faire en programmation bayésienne des robots qui a conduit avec l'expertise linguistique amenée par l'ICP à la formulation probabiliste du modèle de contrôle du système articulatoire du bébé robot.
2. La fourniture d'une librairie de calcul bayésien permettant l'exécution efficace de ces calculs.

Le projet ROBEA « un androïde doué de parole » a contribué à l'avancement des travaux de l'équipe LAPACE-SHARP en lui permettant :

1. d'apporter un nouvel exemple de programmation bayésienne des robots dans un domaine très différent de ceux de la robotique mobile, de la robotique d'assemblage et des véhicules automatiques que nous traitions jusqu'alors.

Et avec l'aide de nombreux autres projets :

2. de terminer les développements du moteur d'inférence probabiliste et de la librairie Pro-BT® destinée à programmer des applications.
3. et de créer, avec le soutien du CNRS, de l'INRIA, de l'UJF et de l'INPG, une entreprise ProBAYES® destinée à la commercialisation d'applications bayésiennes et de la librairie Pro-BT®.

Nous définissons un programme bayésien comme un moyen de spécifier une famille de distributions de probabilités. Notre objectif est de montrer qu'il est possible, en utilisant de telles spécifications, de définir des tâches robotiques, de réaliser de l'apprentissage et enfin de commander effectivement un robot.
La Figure 1 présente les éléments qui constituent un programme bayésien :
– Un *programme* est construit à partir d'une *description* et d'une *question*.
– Une *description* est définie par des *connaissances préalables* et les valeurs des *paramètres* libres.
– Les *connaissances préalables* décrivent un ensemble de *variables* pertinentes, une *décomposition* et un ensemble de *formes*.
– Les *formes* sont soit *paramétriques* soit des *questions* posées à une autre description.
– Les valeurs des paramètres sont soit fixées a priori soit obtenues par un mécanisme d'identification à partir d'un ensemble de *données expérimentales*.

$$\left[\begin{array}{l}
\texttt{Programme :} \\
\left[\begin{array}{l}
\texttt{Description } \omega\texttt{ :} \\
\left[\begin{array}{l}
\texttt{Connaissances préalables } \pi\texttt{ :} \\
\quad \texttt{Variables} \\
\quad \texttt{Décomposition} \\
\quad \left[\begin{array}{l}
\texttt{Formes :} \\
\quad \texttt{Formes paramétriques} \\
\quad \texttt{Question à une autre description } \omega'
\end{array}\right]
\end{array}\right] \\
\\
\left[\begin{array}{l}
\texttt{Paramètres } \lambda\texttt{ :} \\
\quad \texttt{Données expérimentales } \delta \\
\quad \texttt{Paramètres a priori } \lambda_0
\end{array}\right]
\end{array}\right] \\
\\
\left[\begin{array}{l}
\texttt{Question :} \\
\quad \mathrm{P}(\texttt{Search} \mid \texttt{Known} \otimes \omega) = \frac{1}{\Sigma} \times \sum_{\texttt{Hidden}} \mathrm{P}(\texttt{Search} \otimes \texttt{Known} \otimes \texttt{Hidden} \mid \omega)
\end{array}\right]
\end{array}\right]$$

Figure 10. Structure d'un programme bayésien

Pour le bébé robot le programme bayésien est le suivant :

1. Description
    a. Connaissances préalables
        i. Variables
        Les 3 variables motrices Lh, Tb et Td identifiées grâce à l'étude faite sur les cadrages.
        Les 3 variables géométriques Al, Xh et Yh
        Les 2 variables auditives F1 et F2
        ii. Décomposition
        La décomposition retenue est :
        P(Lh Tb Td Xh Yh Al F1 F2)
        = P(Xh) x P(Yh) x P(Al)
        x P(Lh | Al) x P(Tb | Xh Yh) x P(Td | Xh Yh Tb)
        x P(F1 | Xh Yh Al) x P(F2 | Xh Yh Al)
        Cette décomposition traduit le rôle central qui a été choisi pour les variables géométriques. En effet, on considère qu'à partir des trois variables géométriques on est, d'une part, capable de retrouver les paramètres articulatoires et, d'autre part, capable de synthétiser les formants. Cette hypothèse raisonnable conduit à des simplifications de modélisation très importantes.
        iii. Formes
        Les 3 premières distributions apparaissant dans la décomposition sont uniformes, les 5 suivantes sont des gaussiennes.
    b. Paramètres
        L'apprentissage des paramètres libres des formes précédentes est effectué en engendrant suivant des lois uniformes les valeurs des variables articulatoires puis en utilisant le modèle GROWTH pour calculer les valeurs des variables géométriques et auditives correspondantes. La base de données ainsi constituée est alors utilisée pour identifier les valeurs des moyennes et des écarts-types des gaussiennes.

2. Question

   La qualité du modèle et de l'apprentissage est finalement évaluée en inversant le modèle c'est-à-dire en répondant à la question probabiliste : P(Lh Tb Td | F1 F2). On cherche les variables articulatoires permettant d'atteindre une cible auditive donnée.

Notion de cartes bayésiennes

Dans le développement d'un cadre global de programmation bayésienne d'un robot, une question importante concerne la forme que peut prendre la notion de « carte sensori-motrice » dans ce contexte. Qu'est-ce qu'une carte ? Quelle est son utilité ? Qu'est-ce qu'un lieu, un comportement ? Qu'est-ce que naviguer, se localiser et prédire, pour un robot mobile devant accomplir une tâche donnée ? Ces questions n'ont pas de réponses uniques ou évidentes à ce jour, et restent centrales pour de nombreux domaines de recherches. La robotique, par exemple, souhaite y répondre en vue de la synthèse de systèmes sensori-moteurs performants. Les sciences cognitives placent ces questions comme essentielles à la compréhension des êtres vivants, de leurs compétences, et au-delà, de leurs intelligences.

Les travaux réalisés dans le cadre de la thèse de Julien Diard (2003) se situent à la croisée de ces disciplines. En partant de l'étude des méthodes probabilistes classiques (Localisation Markovienne, PDMPOs, MMCs, etc.), puis de certaines approches dites « bio-inspirées » (Berthoz, Franz, Kuipers), nous avons analysé les avantages et inconvénients respectifs de ces approches en les replaçant dans un cadre général de programmation des robots basé sur l'inférence bayésienne. Nous avons ensuite proposé un formalisme original de modélisation probabiliste de l'interaction entre un robot et son environnement : la carte bayésienne. Dans ce cadre, définir une carte revient à spécifier une distribution de probabilités particulière. Certaines des questions évoquées ci-dessus se ramènent alors à la résolution de problèmes d'inférence probabiliste. Nous avons défini des opérateurs d'assemblage de cartes bayésiennes, replaçant ainsi les notions de « hiérarchie de cartes » et de développement incrémental comme éléments centraux de notre approche, en accord avec les données biologiques. En appuyant l'ensemble de notre travail sur le formalisme bayésien, nous profitons d'une part d'une capacité de traitement unifié des incertitudes, et d'autre part, de fondations mathématiques claires et rigoureuses.

Dans ce formalisme, nous avons pu étudier et définir certaines notions de navigation pour les robots mobiles. Une carte est représentée par une distribution de probabilités : c'est notre façon de décrire une interaction. La définition de cette carte correspond à la phase déclarative de notre approche. Naviguer, se localiser, prédire, et générer une loi de contrôle sont en revanche des compétences, donc du domaine procédural. Il s'agit alors de restituer les connaissances contenues dans la carte, par le biais de questions probabilistes. Notre formalisme s'enrichit de la définition de deux opérateurs permettant d'assembler des cartes pour en créer de nouvelles : l'opérateur de superposition, et celui d'abstraction. Le premier permet d'enrichir le vocabulaire pour décrire l'environnement du robot, en fusionnant des modèles disponibles. La carte obtenue dispose de plus de lieux que les cartes initiales. Cela donne au robot une plus grande finesse de représentation de l'environnement. Le second opérateur permet de construire des hiérarchies de cartes. Grâce à cet opérateur, nous obtenons une carte abstraite qui résume de manière considérable les cartes sous-jacentes. Cette réduction d'information permet d'obtenir une carte portant sur un petit espace, donc facilement manipulable (spécification, apprentissage, planification), mais néanmoins

fortement ancré dans le niveau d'abstraction inférieur (reconnaissance de modèles). Ce dernier point souligne la sémantique originale que nous associons à nos cartes : ce sont des modèles de la dynamique de l'interaction entre un robot et son environnement.

Perspectives et applications de la robotique bayésienne

La librairie Pro-BT® permet de traduire simplement ce modèle probabiliste en un programme C++. Le moteur d'inférence associé permet d'effectuer automatiquement les calculs nécessaires pour répondre à la question posée. Ce moteur d'inférence inclut des algorithmes originaux lui donnant une efficacité très grande. De plus, la librairie et le moteur d'inférence ont été développés de manière à pouvoir être exécutés sur des machines parallèles SMMP à 2, 4 ou 8 processeurs ou sur des fermes de processeurs.

La robotique n'est qu'une des nombreuses applications possibles de la programmation bayésienne. Ce projet a donc contribué avec de nombreux autres à la création de la société ProBAYES® qui commercialise Pro-BT® et développe des applications de la programmation bayésienne. À titre d'exemple, ProBAYES® développe actuellement des applications dans le domaine médical (Aide à la pose de prothèse du genou), le domaine financier (aide à l'investissement), le domaine industriel (aide au télédiagnostic), le domaine de distribution (aide à la gestion de stock), le domaine de la gestion de données (aide à la détection de doublons dans les bases de données) et le domaine du commerce sur internet (extraction automatique d'information dans des pages web).

Publications

Diar, J. (2003). *La carte bayésienne, un modèle probabiliste hiérarchique pour la navigation en robotique mobile*. Thèse de doctorat INPG (Institut National Polytechnique de Grenoble).

Diard, J., Bessière, P. & Mazer, E. (2003). A survey of probabilistic models using the Bayesian programming methodology as a unifying framework. *International Conference on Computational Intelligence, Robotics and Autonomous Systems* (IEEE-CIRAS), Singapour.

Diard J., Bessière P. & Mazer E. (2003). Hierarchies of probabilistic models of space for mobile robots : the bayesian map and the abstraction operator. *IJCAI 2003, Workshop on Reasoning with Uncertainty in Robotics*, Acapulco, Mexico.

Lebeltel, O., Bessière, P., Diard, J. & Mazer, E. (2003). Bayesian Robot Programming. *Autonomous Robots*, in press.

4. Conclusion et Perspectives

Notre projet part d'une idée simple, et en fait assez classique : dans les problèmes d'apprentissage quels qu'ils soient, et notamment dans les problèmes d'apprentissage sensori-moteurs, il peut être très profitable de contraindre les mécanismes d'apprentissage. Ceci se fait en général en introduisant des contraintes sensori-motrices explicites ou implicites : ici, ces contraintes sont celles des articulateurs et capteurs de notre robot parlant. Mais une seconde manière de contraindre le problème consiste à en réduire la complexité, en réduisant le domaine de navigation, ce qui permet de simplifier, voire de régulariser des problèmes ardus tels que l'inversion, le « perceptual aliasing », la « global connectivity », tous problèmes rendus compliqués par la multiplicité des correspondances possibles entre actionneurs et capteurs. Or, la réalité biologique fournit une variante beaucoup plus riche, en proposant un « dépliement progressif » de l'espace de navigation, dépliement qui permet de construire progressivement des représentations complexes à parti de représentations simples.

L'originalité de notre travail réside dans la mise en œuvre systématique de comparaison des « paroles » de notre bébé androïde à celles de bébés humains, ce qui permet d'exploiter les connaissances croissantes que nous avons sur le « dépliement ontogénétique » dans l'acquisition du contrôle orofacial pour le langage. Grâce aux travaux de l'équipe Laplace rappelés dans la Section 3, nous disposons maintenant d'un cadre de modélisation, de simulation et d'expérimentation solide pour faire apprendre à notre bébé robot des comportements vocaux de complexité croissante. Les résultats des simulations de la Section 1 montrent comment le fait de contraindre l'exploration sensori-motrice à 4 mois simplifie grandement la mise en œuvre de mécanismes d'inversion basiques, qui permettent de reproduire les performances d'imitation de bébés humains, comportements précurseurs de compétences plus fines telles que celles étudiées dans la Section 2.

L'enjeu majeur de la dernière année de notre projet, c'est le passage au dynamique, et principalement l'acquisition du contrôle des articulateurs portés (la langue, les lèvres, le velum) sur la ligne de base de la porteuse naturelle qu'est la mâchoire. Ce travail se fera en deux étapes. D'une part, nous sommes en train d'analyser, dans nos corpus de productions enfantins, le décours temporel qui permet, entre 7 et 16 mois, de maîtriser la coordination motrice entre plosive et voyelle. Ensuite, il nous faudra passer à l'étape de simulation, qui devra comporter l'introduction, dans notre bébé bayésien, de séquences temporelles de commandes. Nos premières réflexions nous ont fourni quelques éléments de départs, qu'il faudra affiner avant de passer aux simulations. Celles-ci devraient nous permettre de mieux comprendre comment peut s'acquérir un contrôle complexe comme celui de la coarticulation consonne-voyelle, mais aussi d'estimer si l'information perceptive suffit pour acquérir ce contrôle, et, corrélativement, s'il existe des formes perceptives meilleures que d'autres, car plus faciles à apprendre : c'est là le point d'entrée dans les expériences d'émergence de systèmes de communication dans des sociétés de robots, que nous avons mentionnées en introduction de ce rapport.

Enfin, un enjeu important de ce travail consistera à se doter d'une grille de lecture de ce que le bébé robot apprend au cours de son parcours d'exploration/imitation. Les travaux de thèse de Julien Diard au sein de l'équipe Laplace ont conduit, nous l'avons vu, à des résultats très intéressants et originaux sur la notion de « carte bayésienne ». Une perspective de notre travail sera précisément d'utiliser les outils de la simulation robotique pour tenter de caractériser les cartes sensori-motrices acquises par nos bébés robots, et ainsi, de « lire », par analogie, les cartes sensori-motrices cachées dans le cerveau de nos petits modèles vivants que sont les bébés humains en phase d'acquisition du langage.

Bibliographie

Bailly, G. (1997). Learning to speak. Sensori-motor control of speech movements. *Speech Communication, 22*, 251-267.

Boë, L.-J. (1999). Modelling the growth of the vocal tract vowel spaces of newly-born infants and adultes. Proc. *ICPhS99* San Francisco, pp. 2501-2504.

Boë, L.J., Abry, C., Beautemps, D., Schwartz, J.L., & Laboissière, R. (2000). Les sosies vocaliques – Inversion et focalisation. *XXIIIèmes Journées d'Étude sur la Parole*, Aussois, 257-260.

Boë, L.-J., Perrier, P., Guérin, B. & Schwartz, J.-L. (1989). Maximal Vowel Space. Proc. *Eurospeech89*, 2, 281-284.

Davis, B. & MacNeilage, P. F. (1995). The articulatory basis of babbling. *Am. Speech-Lang.-Hearing Ass., 38*, 1199-1211.

Goldstein, U.G. (1980). *An articulatory model for the vocal tract of the growing children*. Thesis of Doctor of Science, MIT, Cambridge, Massachusetts.

Jakobson, R. (1968). *Child language, aphasia, and phonological universals.* The Hague : Mouton.

Kent, R.D. & Miolo, G. (1995). Phonetic Abilities in the First Year of Life. In Fletcher, P. & MacWinney (Eds.) *The Handbook of Child Language*, (pp.303-334). Blackwel Publishers.

Koopmans-Van Beinum, F., & Van Der Stelt, J. (1986). Early speech development in childs acquiring Dutch. mastering general basic éléments. In Gillis, S. & de Hower, A. (Eds.) *The Acquisition of Dutch* (pp.101-425). Amsterdam/Philadelphia: Benjamins.

Kuhl, P. K., & Meltzoff, A. N. (1996). Infant vocalizations in response to speech: vocal imitation and developmental change. *JASA, 100*, 2425-2438.

Ladefoged, P. (1967). The nature of vowel quality. In Ladefoged, P. *Three Areas of Experimental Phonetics* (pp. 50-142). London: Oxford University Press.

Lindblom, B., & Lubker, J. (1985). The Speech Homunculus: A Problem of Phonetic Linguistics. In V.A. Fromkin, (Ed.) *Phonetic Linguistics* (pp. 169-192). Orlando, Florida : Academic Press.

Lisker, L., & Abramson, A.S. (1964). A cross-language study of voicing in initial stops, Acoustical measurements. *Word, 20*, 384-422.

Lisker, L. & Rossi, M. (1992). Auditory and visual cueing of the [± rounded] feature of vowels. *Language and Speech, 35*, 391-417.

MacNeilage, P. F. & Davis, B. (1990). Acquisition of Speech Production, Frames then Content. In M. Jeannerod (Ed.) *Attention and Performance, XIII Motor Representation and Control*, (pp.453-476).

MacNeilage, P.F., Rootes, T.P., & Chase, R.A. (1967). Speech production and perception in a patient with severe impairment of somesthesic perception and motor control. *JSHR, 10*, 449-467.

Maeda, S. (1989). Compensatory Articulation during Speech: Evidence from the Analysis and Synthesis of Vocal-Tract Shapes using an Articulatory Model. In W.J. Hardcastle & A. Marchal (eds.) *Speech Production and Modelling* (pp. 131-149). Kluwer : Academic Publishers.

Meltzoff, A. N. (2000). Newborn imitation. In. Min, D. et Blater, A. al. (Eds) *Infant development, the essential readings* (pp 165-181). Blackwell. Publishers

Mills, A .E. (1987). The development of phonology in the blind child. In B. Dodd and R. Campbell (Eds) *Hearing by eye: The psychology of lipreading* (pp. 145-161). London: Lawrence Erlbaum Associates.

Schroeder, M.R., Atal, B.S., & Hall, J.L. (1979). Objective measure of certain speech signal degradations based on masking properties of human auditory perception. In *Frontiers of Speech Communication Research* (B. Lindblom & S. Ohman, eds.), pp. 217-229. London: Academic Press.

Serkhane, J., Schwartz, J.L., & Bessière, P. (2003). Simulating vocal imitation in infants, using a growth articulatory model and speech robotics. *Proc. ICPhS*, Barcelona.

Serniclaes, W. (1987). *Etude expérimentale de la perception du trait de voisement des occlusives du Français*. Thèse Sciences Psychologiques, Institut de phonétique, Université Libre de Bruxelles.

Vallée, N. (1994). *Systèmes vocaliques : de la typologie aux prédictions.* PhD Stendhal University: Grenoble, France.

Vallée, N. , Boë, L.J., & Payan, Y. (1995). Vowel prototypes for UPSID's phonèmes. *Proc. XIIIth International Congress of Phonetic Sciences*, Stockholm, vol. 1, pp. 424-427.

Vilain, A., Abry, C., & Badin, P. (2000). Coproduction strategies in French vcvs. *Proceedings of the 5th Seminar on Speech Production: Models and data, Munich*, Germany, pp. 81-84.