Etude comparative de vocalisations de bébés humains et de bébés robots

J. Serkhane [(1)], J.L. Schwartz[(1)], L.J. Boë[(1)], B. Davis[(2)], P. Bessière[(3)], E. Mazer[(3)]

(1) Institut de la Communication Parlée - CNRS / INPG / Université Stendhal
INPG, 46 Av. Félix Viallet, 38031 Grenoble Cedex 1 Tél.: ++33 (0)4 76 57 45 33 - Fax: ++33 (0)476 57 47 10
(2) Dept of Psychology, University. of Texas, Austin, TX, 78712, USA
(3) Laplace-Sharp, Inria Rhône-Alpes, Zirst - 655 avenue de l'Europe, 38330 Montbonnot

ABSTRACT

In order to assess infants' motor skills during speech development, we used a statistical model of the vocal tract that integrates growth of the effector system. This model allowed us to infer, from actual vocalizations, the likeliest explored acoustic regions, articulatory degrees of freedom and vocal tract shapes, and to test MacNeilage and Davis' cooccurrence hypothesis. Our results will feed the building of a virtual robot, modelling speech development.

INTRODUCTION

Cette étude est préliminaire à l'élaboration d'un robot virtuel, pour une modélisation du développement de la parole qui s'inspire du chemin que le petit humain semble emprunter. 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 vers les sons 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 [Bai97]. 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 Jak68). Or, la simple observation de sujets (réels) montre qu'il n'en est rien [KeM95] : 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. De plus, au niveau computationnel, la modélisation d'une exploration exhaustive complique l'apprentissage du lien sensori-moteur [Bes00].

Postulons que « la vérité sort de la bouche des enfants » : ils n'explorent pas tout l'univers des possibles articulatori-acoustiques de leur conduit vocal pour en maîtriser le comportement. Que sait-on de leurs capacités sensori-motrices précoces? (a) À la naissance, 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 [Mel00]. 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) À quelques semaines, les nourrissons **vocalisent**. De plus, ils sont à même d'orienter leurs productions vers des sons vocaliques fréquemment perçus (imitation vocale précoce) [KuM96], et de détecter la congruence entre le son d'une voyelle et l'image du visage en mouvement qui l'a produite (intégration multimodale) [KuM82]. (c) Vers 7 mois, ils **babillent** : leur mandibule se lance dans des cycles d'élévation-abaissement, alors que leurs cordes vocales vibrent. C'est le Babillage Canonique (BC) [KoV86, MnD90]. (d) Par la suite, 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) [Vil00].

Notre projet est d'implémenter ces étapes dans un robot virtuel, doué d'un mécanisme d'apprentissage bayésien. Dans la présente étude, nous essayons d'**estimer les capacités motrices** disponibles aux phases b et c, à partir de la description acoustique de vocalisations produites par de vrais enfants à ces stades de développement.

1. BÉBÉ ROBOT ET BÉBÉ HUMAIN

1. 1. 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 [Gol80] : GROWTH [BoM98]. Son noyau computationnel [Mae90] est issu de l'analyse statistique de coupes sagittales cinéradiographiques d'un conduit vocal adulte en action, d'où 7 facteurs principaux 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.

1. 2. 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. Nous les avons récupérées d'une publication de Kuhl et Meltzoff [KuM96]. Matyear et Davis nous ont fourni le second jeu de données, relevé pour étudier les productions pseudo-syllabiques de babillage. Nous avons sélectionné les sons vocaliques produits par leurs sujets de 7 mois, en début de babillage canonique. Dans chaque cas, les deux premiers formants et la description phonétique sont disponibles.

2. ANALYSE DES DONNÉES

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. Une dernière expérience veut tester l'hypothèse selon laquelle, au début du BC, les sons vocalique et consonantique d'une même pseudo-syllabe ont le même lieu d'articulation (hypothèses de cooccurrences, DaM95).

2. 1. Cadrage acoustique

Méthode. Tous les sons oraux que peut générer le modèle articulatoire s'inscrivent dans l'Espace Vocalique Maximal (**EVM**) [BPG89]. 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.*

Résultats. Les vocalisations réelles s'inscrivent dans l'EVM correspondant (figures 1 et 2) : 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 contrôle moteur mature. En particulier, les vocalisations de 4 mois (figure 1) sont relativement centrales et mi-hautes : les productions les plus en avant, arrières et ouvertes ne semblent pas explorées. À 7 mois (figure 2), les productions vocaliques exploitent plus la dimension haut-bas qu'au stade précédent.

2.2. Cadrage articulatoire

Méthode. 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. Autrement dit, le but de cette méthode est *d'estimer le jeu minimal de degrés de liberté requis pour reproduire les sons vocaliques observés.* Nous avons conçu plusieurs sous-modèles articulatoires à partir d'un nombre restreint, et varié, de paramètres moteurs de GROWTH. Un *sous-modèle* se caractérise donc par le nombre de paramètres articulatoires qui le constituent, leur nature, ainsi que le domaine de variation de leurs valeurs. La capacité de chaque sous-modèle à reproduire les sons vocaliques observés, à chaque stade de développement, est évaluée par sa *probabilité, sachant les vocalisations réelles*[1] : P(Mi/f1f2), où Mi est le i^{ième} sous-modèle, caractérisé par l'ensemble des données acoustiques qu'il génère, et f1f2, les formants des données réelles. Le sous-modèle « *gagnant* » est celui qui s'ajuste le mieux aux données réelles : il maximise le critère de probabilité conditionnelle.

Résultats. 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, 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.

2. 3. Cadrage géométrique

Méthode. Les 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, produit par le modèle vainqueur et acoustiquement plausible. 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). Un son vocalique donné renvoie à la moyenne et la variance des valeurs de chaque descripteur géométrique pour les configurations issues de l'inversion. Les phénomènes compensatoires impliquent que les variances obtenues sont souvent très élevées, en particulier pour des configurations proches de celle de repos. Par soucis de lisibilité, ne sont donc représentées que les ellipses de dispersion obtenues par l'inversion, via GROWTH, de 4 « prototypes » ajoutés au pool de données réelles : [i a u] sont choisis à une position équivalente à celle de l'adulte sur l'acoustique de l'EVM, et [´] correspond au neutre articulatoire de GROWTH. Les résultats d'inversion de [i a u ´] servent ainsi de point de repère.

Résultats. A 4 mois (figure 3), 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 (figure 4), les vocalisations correspondent à une exploration plus étendue des positions de langue, autant sur l'axe haut-bas qu'avant-arrière.

Une histoire de /u/. En théorie, notre conduit vocal peut produire trois types de /u/, aux trois premiers formants identiques [BAB00] : vélopalatal, vélopharyngal et pharyngal dont la constriction est, respectivement, palatale, dans le haut-pharynx et pharyngale. In natura, les locuteurs (adultes) de toutes les langues testées ne produisent que le [u] *vélopalatal* [Woo79]. Le [u] vélopharyngal n'est observable que chez de rares sujets, dans des expériences de perturbation (lip-tube, [Sav95]). Le [u] vélopharyngal restant n'a jamais été observé.
D'après AbB96, le [u] palatal serait la première stratégie de production de [u] adoptée sur le chemin de développement de la parole : sa dominance, quasi totale, au stade adulte se justifierait par sa cartographie sensori-motrice précoce. Nous avons cherché à savoir si l'analyse des données acoustiques à notre disposition est en faveur ou non de cette hypothèse, en faisant l'inversion exhaustive de la vocalisation la plus proche de [u] à 4

mois. Il en résulte que les formes de conduit vocal aptes à produire cet [u]-juvénile présentent une articulation : (a) palatale à pharyngale, si le modèle utilisé pour l'inversion est GROWTH (Témoin, figure 5), (b) **uniquement palatale** si l'inversion se base sur le modèle simulant le mieux le comportement moteur de ce stade de développement (Modèle gagnant, figure 6). Ceci conforte l'hypothèse d'AbB96.

2. 4. Test des hypothèses de cooccurrences

Méthode. D'après MacNeilage et Davis [Mac98, MnD90] les pseudo-syllabes du BC sont le fruit de l'oscillation mandibulaire, appliquée à des configurations de langue (*pre-settings*). Pour tester leurs trois hypothèses de cooccurrences [DaM95], nous avons simulé une élévation de la mandibule à partir des formes de conduit (pre-settings) issues de l'inversion exhaustive, par le modèle-gagnant à 7 mois, d'une vocalisation de BC par classe de transcription[2]. Chaque configuration linguale renvoie donc à un lieu de closion (méthode inspirée de Vil99).

Résultats. Les réalisations vocaliques antérieures et centrales sont le plus fréquemment associées à une articulation coronale, ce qui est conforme à la morphologie propre au modèle (cf. Vil99). De plus, nos simulations montrent qu'un recul de la langue se traduit par un recul à la fois de l'articulation du vocant et de celle du closant (co-occurrences, mises en évidence par DaM95). Nos résultats donnent une dimension statistique à ceux de Vil99, qui utilisent des modèles adultes du même type mais de morphologie différente, et pour pre-settings des prototypes articulatoires. L'ensemble de ces résultats est congruent avec l'idée selon laquelle si le cadre syllabique se développe avant le contenu segmental [MnD90], la position (horizontale) de la langue ne change pas au cours de la réalisation de la pseudo-syllabe.

CONCLUSION

Cette étude fait le lien entre les productions d'un modèle articulatoire en croissance et celles de nourrissons. Nos résultats soulignent que *le développement de la parole ne part pas d'une exploration exhaustive*. Une stratégie du type « exploration totale de ses capacités puis sélection du sous-ensemble des productions efficaces, pour interagir avec le groupe et minimiser le coût énergétique des manœuvres articulatoires dans l'atteinte des sons-cibles » ne serait pas adaptative, vu le temps et l'énergie dépensés en une exploration laborieuse qui se solderait par une régression, en fin de parcours.

Les fruits de cette étude sont à prendre avec précaution : ils souffrent des biais propres à la modélisation, notamment ceux hérités du modèle dont ils émanent. Toutefois, GROWTH est le seul modèle articulatoire intégrant le *processus de croissance*. De plus, malgré ces limites de modélisation, nos résultats n'en sont pas moins plausibles. Les paramètres de séparation des lèvres et du corps de la langue, issus du cadrage géométrique des données de pré-BC, se retrouvent dans les capacités motrices néonatales. De même, le paramètre du dos de la langue (Td), qui semble recruté dans les productions de 4 mois, le serait aussi dans le mode d'ingestion de ce stade développemental. Le paramètre mâchoire (J) n'apparaît que dans le cadrage articulatoire des données de babillage.

Notre programme de modélisation robotique du développement de la parole, qui dote son avatar d'une morphologie en croissance et de capacités motrices réalistes, peut désormais voir le jour.

Remerciements

Un grand merci à B. Davis et C. Matyear pour avoir mis à notre disposition leurs données expérimentales.
Ce programme est subventionné par les projets CNRS-STIC-Robea, CNRS-SHS-OHLL et Eurocores-OMLL.

BIBLIOGRAPHIE

[AbB96] Abry, C., Badin, P. (1996), "Speech Mapping …", *Proc. 4th Speech Prod. Sem*., Autrans, 175-184.

[BAB00] Boë, L.J., et al. (2000), "Les sosies vocaliques… ", in *JEPXXIII*, 257-260.

[Bai97] Bailly, G. (1997), "Learning to speak.…" *Speech Com.*, 22, 251-267.

[Bes00] Bessière, P. (2000) ; Vers une théorie probabiliste des systèmes sensori-moteurs ; HDR, Université Joseph Fourier, Grenoble, France.

[BPG89] Boë, L.-J., et al. (1989) : "Maximal Vowel Space" in *Eurospeech89*, 2, 281-284

[BoM98] Boë, L.-J. & Maeda, S. (1998) : "Modélisation de la croissance du conduit vocal", *Jour. d'Et. Ling.*, Nantes, 98-105.

[BGP95] Boë, L.-J et al. (1995) : "Vers une unification des espaces vocaliques"; in C. Sorin et al. (eds.) *Levels in Speech Communication: Relations and Interactions* (pp. 63-71). Elsevier B.V.

[DaM95] Davis, B. & MacNeilage, P. F. (1995), "The articulatory basis of babbling", *Am. Speech-Lang.-Hearing Ass.* (38), 1199-1211.

[Gol80] Goldstein, U.G. (1980), "*An articulatory model for the vocal tract of the growing children"*. Thesis of Doctor of Science, MIT, Cambridge, Massachusetts.

[Jak68] Jakobson, R. (1968), "Child language, aphasia, and phonological universals", The Hague : Mouton.

[KeM95] Kent, R.D. & Miolo, G. (1995) : "Phonetic Abilities in the First Year of Life" in *The Handbook of Child Language*, Fletcher, P. & MacWinney(Eds.),(pp.303-334)Blackwel

[KoV86] Koopmans-Van Beinum, F. et Van Der Stelt, J. (1986), "Early stages in the development of speech movements", in B. Lindblom, B. et Zetterstrom, R. (eds,), *Precursors of Early Speech* (pp. 37-49), New York : Stockton Press.

[KuM82] Kuhl, P. K. et Meltzoff, A. N. (1982), "The bimodal perception of speech in infancy", *Science218*, 1138-1141.

[KuM96] Kuhl, P. K. et Meltzoff, A. N. (1996), " Infant vocalizations in response to speech…", *JASA100*, 2425-2438.

[Mae90] Maeda, S. (1990), "Compensatory articulation during speech…", in W.J. Hardcastle & A. Marchal (eds.) *Speech Production and Modellling* (pp. 131-149), Kluwer.

[Mac98] MacNeilage, P. F. (1998) : "The Frame/Content Theory of Evolution of Speech Production" in *BBS21* (4), 499-511.

[Mel00] Meltzoff, A. N. (2000),"Newborn imitation", in. Min, D. et Blater, A. al. (eds) *Infant development, the essential readings* (pp 165-181), Blackwell.

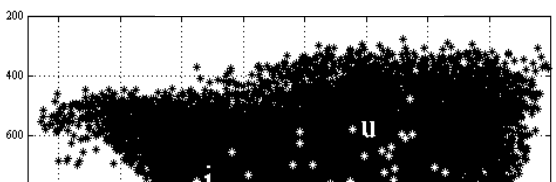
[MnD90] MacNeilage, P. F. & Davis, B. (1990) : "Acquisition of Speech Production, Frames then Content" in M. Jannerod (ed), *Attention and Performance, XIII Motor Representation and Control*, (pp.453-476).

[Sav95] Savariaux, C. (1995), "*Etude de l'espace de contrôle distal en production de la…*", Thèse INP-Grenoble.

[Vil99] Vilain A. et al. (1999) : "From idiosyncratic pure frames to variegated babbling…", ICPhS'99, San Francisco, USA.

[Vil00] Vilain, A. et al. (2000), "Coproduction strategies in French vcvs…", Proceedings of the 5th Seminar on Speech Production: Models and data, Munich, Germany, pp. 81-84.

[Woo79] Wood, S. (1979), "A radiographic analysis of constriction locations for vowels", in *J. Phon.*, 7, 25-43.



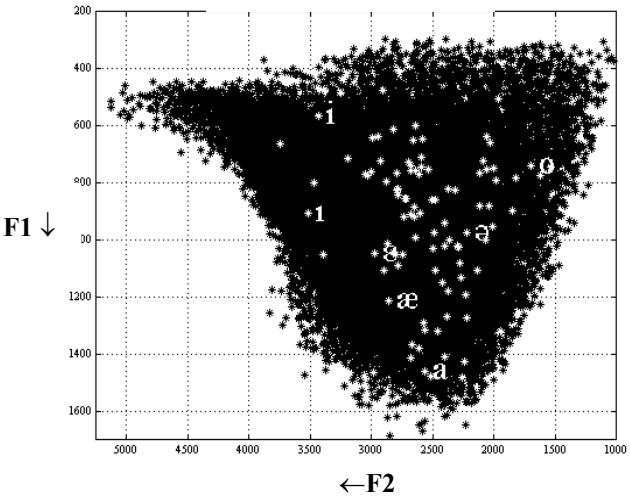


Figure 2 : Cadrage acoustique des vocalisation de 7 mois (points blancs) par l'EVM du même âge (points noirs).

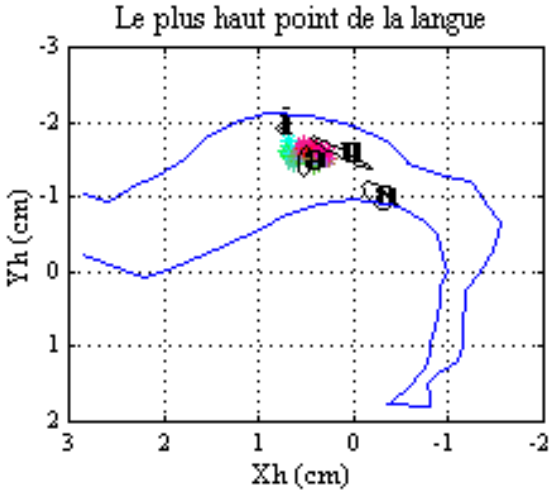


Figure 3: Représentation graphique de la moyenne de (Xh, Yh) résultant du cadrage géométrique des vocalisations de 4 mois par le modèle gagnant Lh-Tb-Td.

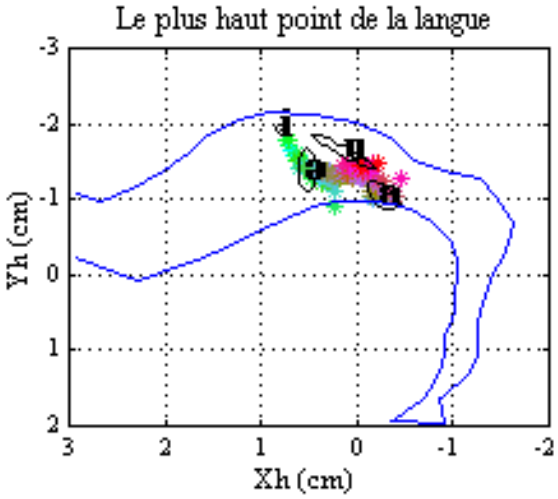


Figure 3: Représentation graphique de la moyenne de (Xh, Yh) résultant du cadrage géométrique des vocalisations de 7 mois par le modèle gagnant Lh-J-Tb-Td.

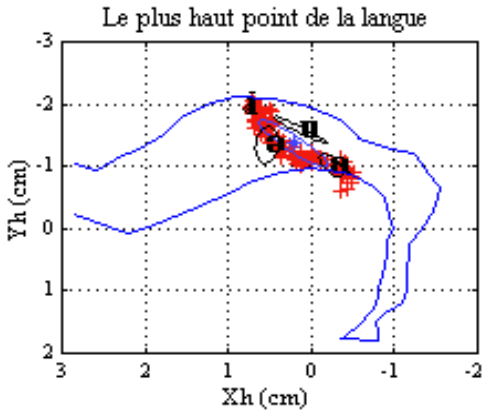


Figure 4: Représentation graphique de toutes les valeurs de (Xh, Yh) issues de l'inversion exhaustive, via GROWTH, du /u/-juvénile.

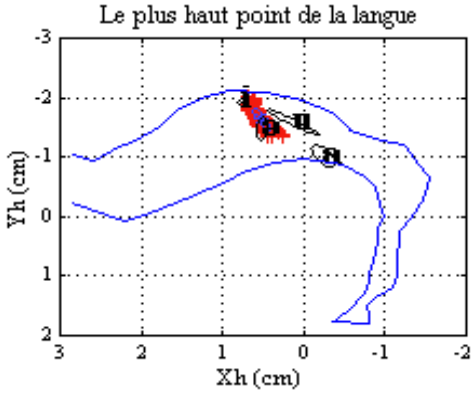


Figure 5: Représentation graphique de toutes les valeurs de (Xh, Yh) issues de l'inversion exhaustive, via le modèle gagnant, du /u/-juvénile.