# Simulating Vocal Imitation in Infants, using a Growth Articulatory Model and Speech Robotics

**Serkhane, J. E.** [†], **Schwartz, J.L.** [†] **and Bessière, P.** [‡]
†Institut de la Communication Parlée - CNRS / INPG / Université Stendhal, Grenoble, France
‡ Laplace-Sharp, Inria Rhône-Alpes, Grenoble, France

E-mail: jihene@icp.inpg.fr, schwartz@icp.inpg.fr, Pierre.Bessiere@imag.fr

## ABSTRACT

In order to shed lights on the cognitive representations likely to underlie early vocal imitation, we tried to simulate Kuhl and Meltzoff's experiment (1996), using Bayesian robotics and a statistical model of the vocal tract that had been fitted to pre-babblers' actual vocalizations. It was shown that audition is compulsory to account for infants' early vocal imitation performance, inasmuch as the simulation of purely visual imitation failed to reproduce infants' score and pattern of imitation. Further, a small number of vocalizations (less than 100!) appeared to be enough for a learning process to provide scores at least as high as those of pre-babblers. Thus, early vocal imitation lies in the reach of a baby robot, with only a few assumptions about learning and imitation.

## 1. INTRODUCTION

The present study is part of a project that aims at modeling speech development through the construction of a virtual baby robot, viewed as a growing sensori-motor system which is able to learn and to interact, and that takes into account how infants progress from non-speech to the mastery of their ambient languages in line with the Frame/Content theory [5]. The held viewpoint is that phonetic development relies on two basic mechanisms: the exploration of the current sensori-motor abilities of the vocal tract and the imitation (overt simulation) of caretakers' language sounds. In a previous paper [7], the focus was on assessing infants' early articulatory skills to specify our robot first capacities, exploiting the *Variable Linear Articulatory Model* (VLAM) [1], which integrates the non-uniform growth of the tract, and sets of formant frequencies, produced before and at the beginning of canonical babbling by 4- and 7-month-olds, respectively.

This paper deals with the imitation issue. What about infants' pre-speech sensori-motor skills? (a) At birth, they are able to imitate three gestures from vision (facial imitation): tongue and lips protrusions, and mandible depression [6]. Although this ability is not obviously linked with speech development, infants are nonetheless likely to gain some sensori-motor experience from it. (b) At a few weeks old, infants vocalize. They tend to direct their productions towards vowels perceived in their environments (early vocal imitation) [3], and to match a vowel sound to the moving image of the face that utters it (multimodal matching) [2].

According to [3], vocal imitation "requires that infants recognize the *relationship* between articulatory movements and sound". However, as [8] pointed out, there is no clear consensus as to whether early vocal imitation, brought to light by Kuhl and Meltzoff [3], needs visual, auditory or both information, given that the subjects are exposed to *audiovisual* face-voice stimuli while newborns display *visual* imitation capabilities [6]. The present study tries to simulate [3]'s experiment, exploiting the VLAM that was fitted to prebabblers' inferred motor abilities [7] and Bayesian robotics [4] that provides our robot with a means to learn and to use the relationships between its tract movements and their perceived consequences.

The purpose was to gain some insights into the cognitive representations that might be involved in early vocal imitation and to test whether and how our robot is able to reproduce, at least, the actual infants' imitation performance when supplied with purely visual, purely auditory and audiovisual information.

## 2. HUMAN INFANTS AND BABY ROBOT

### 2.1. Early vocal imitation

In [3], 72 subjects, aged from 12 to 20 weeks old, were exposed to audiovisual adult face-voice stimuli corresponding to the vowels [i], [a] and [u]. Their subsequent vowel-like productions were, whenever possible, phonetically and acoustically described. The system of transcription was that of the set of English vowels but the transcribed items were merged into three main classes: the /a/-like, including [a æ ʌ], the /i/-like, with [i I ɛ], and the /u/-like for [ʊ u]. Table 1 provides the resulting confusion matrix. In sum, the pre-babblers produced vocalic sounds significantly more often categorized as being like the "target" after they had been exposed to this stimulus than otherwise, with about 59 % of total responses that are congruent (hereafter *%CR*) with an imitative behavior.

### 2.2. Assessing prebabblers's motor skills

If pre-babblers' imitation abilities are to be simulated, the first step is to evaluate the set of articulatory configurations at their disposal to vocalize. This issue was tackled in a previous work [7] capitalizing on the *Variable Linear Articulatory Model* (VLAM) [1] which is a statistical articulatory-acoustic model that integrates the non-uniform growth of the vocal tract [9], thereby, taking into account the normalization phenomenon. Its computational core [10] stemmed from a statistical analysis of mid-sagittal sections of a speaking adult vocal tract, which led to 7 relevant axes. These tract shape descriptors turned out to be related to

ISBN 1-876346-48-5 © 2003 UAB

concrete muscular actions: they are degrees of freedom of a virtual vocal tract and serve as the VLAM inputs along with the selected age. The model output is a description of the tract shape (from the lips to the glottis), which includes its inter-lip area (Al), as well as the formants of the resultant sound. In sum, the VLAM simulates, with a *growing* tract, *adult* motor skills from which 4-mth-olds' skills were inferred. The foregoing were assessed by seeking the minimal set of the age-matched VLAM commands best able to recover the formant frequencies of 20-week-olds' actual vocalizations. These vocalic sounds were those plotted on figure 3 of [3]. Several articulatory sub-models having various subsets of VLAM motor parameters with diverse ranges of variation were comparatively assessed according to a probability criterion (see [7]): the “best” sub-model was the one which minimized the distance between the distribution of its acoustic outputs and that of the actual data in the first two formants (F1, F2) plan. Hereafter, *I4S* will refer to this Inferred 4 months Sub-model. The I4S set of motor parameters consists of the lower lip height (LH), the tongue body (TB) and dorsum (TD) VLAM commands with restricted ranges of variation.

## 3. SIMULATING EARLY VOCAL IMITATION

### 3.1. Testing for facial imitation

Newborns are able to imitate seen but unfelt specific facial gestures (performed by an adult) while they feel but do not see their own faces [6]. Could early vocal imitation be based on "hard-wired" purely visuo-motor imitation skills?

#### 3.1.1. Method

To test whether facial imitation could account for the performance of [3]'s experiment, the values of the [i a u] interlip areas (Al) in the 4 months old VLAM were used as targets: they were exhaustively inverted through the I4S as follows. A series of simulations was randomly generated within the I4S motor abilities. The configurations that had Al values falling within the neighborhood of each target were selected. The sounds corresponding to these simulations, having suitable lip- but arbitrary tongue-shapes, were computed and, then, categorized as [i], [a] or [u] according to their nearest targets in the (F1, F2) plan, in terms of Euclidean distance.

#### 3.1.2. Results

Globally, this modeling experiment yielded about 51% CR which is not much lower than the actual score (59%). This suggests that facial imitation could partially explain early vocal imitation. However, as there is no reason to suppose infants to perform a *perfect* visual imitation of Al values, the score of this experiment should be lower than 51%. Moreover, the confusion matrix of this test provided a typical pattern of visual confusion: many confusions occurred between [i] and [a] or [i] and [u], but none did between [a] and [u] whose interlip areas are very different. This is quite unlike the actual pattern (Table 1). In sum, visual information alone and, thus, pure visuo-motor imitation are not enough on their own to account for early vocal imitation.

### 3.2. Testing for auditory and audiovisual imitation

As they vocalize, the pre-babblers are likely to build a perceptuo-motor map of their vocal tract behaviors, i.e. a cognitive representation of the correspondence between the articulatory configurations they perform and the matched sensory feedbacks. Such acquired knowledge could underlie vocal imitation, since it allows to infer the motor configuration that can generate a sensory state equivalent to the perceived target.

#### 3.2.1. Sensori-motor relationships in Bayesian robotics

Bayesian robotics [4] was capitalized on to model audition- and audiovision-based imitations. In this framework, the robot learns a sensori-motor map of its vocal tract behavior corresponding to a probabilistic description of the observable links between its articulatory and its perceptual variables. Then, imitation conforms to inversion that is the conversion of a sensory state into a motor counterpart.

The chosen motor parameters were those of the I4S, i.e. the lower lip height (LH), the tongue body (TB) and dorsum (TD) commands while the auditory variables were the first two formant frequencies (F1, F2) expressed in Bark, that is, a scale of frequency perception. The formants of a vocalic sound are function of the tract shape which can be described by three geometric variables: the inter-lip area (Al) and the coordinates (Xh, Yh) of the tongue highest point in a fixed system of reference on the tract mid-sagittal section. The latter are potential outputs of the somesthetic system and Al can be either a somatosensory or a visual variable (depending on whether this piece of information is self- or other-generated). All the model variables are supposed to be discrete with mutually exclusive values.

The computational core of a Baseyian robot is the joint probability whose decomposition states the set of hypotheses about the statistical relationships between its variables so as to represent their links. (Xh, Yh, Al) were used as pivots of the decomposition. Indeed, as they form an intermediate space between the auditory and the articulatory spaces, they help reduce the impact of the many-to-one problem on inversion when they function as independent variables in the joint probability decomposition which is defined in the present implementation by:

$$
\begin{aligned}
&P(LH \otimes TB \otimes TD \otimes Xh \otimes Yh \otimes Al \otimes F1 \otimes F2) \qquad (1)\\
&= P(Xh) * P(Yh) * P(Al)\\
&\quad * P(LH/Al) * P(TB/Xh \otimes Yh) * P(TD/Xh \otimes Yh \otimes TB)\\
&\quad * P(F1/Xh \otimes Yh \otimes Al) * P(F2/Xh \otimes Yh \otimes Al)
\end{aligned}
$$

In Equation (1), Xh, Yh and Al are considered to have uniform distributions. All other factors are supposed to obey conditional Gaussian laws whose means and variances must be tuned in a learning phase.

#### 3.2.2. Learning the statistical relationships

To become an actual (and useful) description of the robot's sensori-motor behavior, the distributions composing the above probabilistic structure need to be learnt from a set of experimental data that corresponds, here, to a random exploration of the I4S articulatori-geometrico-acoustic skills. The robot's "proficiency" in inversion, that is, in exploiting its map via Bayesian inference to draw motor values likely to make it reach a given target-state of its perceptual variables, will mainly depend on the learning database size (DBS) and the degree of discretization of the geometric parameters (GDD). Indeed, as Xh, Yh and Al are the linchpin of the description, the GDD partly determines the accuracy of the distributions the robot learns: it sets the minimal gap required to distinguish two items in the discretized geometric domain and the size of the learning space, i.e. the number of articulatory and auditory distributions that have to be learnt for the description to represent the whole range of the I4S abilities. However, there is a trade-off between the GDD and the DBS because a given geometric "box" must include enough configurations for the matched motor and auditory distributions to be learnt.

To evaluate which description could best account for the performance reported in [3], 4 GDD x 15 DBSs were tested. The DBS ranged from 1 to 60,000 items. The GDD, ranked in descending order, were {16, 16, 8}, {8, 8, 4}, {4, 4, 2} and {2, 2, 1} for the number of {Xh, Yh, Al} classes, which yielded 2048, 256, 32 and 4 boxes in the geometric space, respectively. In a first step, the GDD/DBS trade-off was assessed by studying the ability of the model to invert vocalizations of its exploration domain.

Figure 1 illustrates the results for the auditory inversion of 1000 items randomly chosen within the I4S abilities. At maximal DBS, the error decreases, as the GDD increases, and reaches lower than 0.5 Bk values (roughly, formant *jnd*) for the highest two GDD. Moreover, for a given GDD, the error tends to decrease, along with the DBS rise, until a limit that is the lowest this GDD can make the robot perform. However, all the GDD, but the smallest, yield erratic scores as long as the DBS is below a certain value from which the GDD-matched *under-learning phase* ends. Under-learning is due to the relatively small number of actually learnt geometric boxes whose affiliated motor configurations are invariably chosen by the robot regardless of their irrelevance given the target. Indeed, the smallest DBS that is required to have an error at most 10% from the GDD-matched lowest was found to be three times the size of the geometric space (in boxes). In other words, the more boxes there are in the geometric space (the larger the GDD is), *the more precise* its variables are, but *the larger the DBS* must be for the robot's map to be representative of its sensori-motor skills.

### 3.2.3. Implementing A and AV imitation

After a model, defined by a given GDD, had been learnt with a given DBS, it was evaluated by imitation tests. In auditory (hereafter *A*) imitation, the perceptual target was the (F1, F2) pair of a vowel, while in the audiovisual (*AV*) imitation it was its (F1, F2, Al) values. Two target sets were focused on: they are referred to as “external” and “internal” [i a u] items, respectively. The former corresponded to the 4 months old VLAM [i a u] vowels, the latter were their closest simulations within the I4S capacity. Thus, both target sets fitted the 4 months old vocal tract size so as to stand for “normalized” [i a u] vowels, but the first one consisted of [i a u] items that are outside the acoustic and the inferred articulatory regions the actual infants explore, whereas the second one did of simulations whose sounds are at the three corners of the I4S (F1, F2) acoustic space. For each target, 300 motor configurations were drawn from the *P(LH⊗TB⊗TD/PerceptualTarget)* distribution. The formants matching each articulatory pattern were worked out and the sound was categorized as [i], [a] or [u] according to its nearest target in the (F1, F2) plan, in terms of Euclidean distance. In other words, the effect of whether the targets belong to the robot articulatory-acoustic abilities was investigated, as [3]'s infants imitate vowels that ought to be out of their motor abilities.

### 3.2.4. A and AV imitation results

The %CRs as functions of the GDD and the DBS in the AV inversion of the internal and external [i a u] targets are displayed in Figures 2 and 3, respectively. The A inversion scores (not shown here) were globally lower than the AV ones. Further, the following trends appear.

*GDD/DBS Trade-off and under-learning*

As could be expected, whatever the case, the same GDD/DBS trade-off as in Figure 1 is found. Further, all the GDD, but the smallest, require the robot to have learnt a greater number of data to get over the under-learning phase with external than with internal targets, be the inversion A or AV.

*External vs. internal targets: the risk of over-learning*

For a given DBS, the external targets tend to yield lower scores than the internal ones: this is understandable considering that the former are outside the I4S vocalization space whereas the latter are not. Strikingly, in the A case, the highest GDD (2048 geometric boxes) never reached 100% CR for external targets, even with the maximal DBS (60,000 items)! This is ascribable to the *over-learning problem*. Indeed, even when the description is completely representative of the robot sensori-motor abilities (e.g. with a maximal DBS), as all the distributions of the motor variables have small variances, i.e. are very accurate, while none of them matches the target, the robot tends to draw articulatory configurations regardless of their irrelevance given the perceptual goal. In other words, the GDD *has to* contain a rather small number of (large) boxes for the robot to be able to imitate vocalic sounds that are out of its sensori-motor abilities. However, the over-learning problem is overcome if the visual information is also provided (Figure 3): since the VLAM [i a u] interlip areas belong to the I4S ones, the robot is enabled to select configurations that produce the nearest sounds to the target.

*Early vocal imitation does not need much learning*

Altogether, it is noteworthy that the robot needs *neither a high GDD nor a large DBS*, in order to perform as good as, and even better than, the actual infants. Further, the required DBS is generally lower in the AV than in the A condition. For instance, in the case of external targets, which are out of the robot motor abilities so as to best parallel [3]'s actual experiment, the score goes beyond 60% CR with a GDD of 32 boxes and DBSs of 50 and 25 data, in A and AV inversion, respectively.

## 4. CONCLUSION

Two main conclusions can be drawn from this work. First, audition ought to be compulsory to account for infants' early vocal imitation performance with 4-month-olds' inferred articulatori-acoustic abilities, inasmuch as the simulation of purely visual imitation failed to reproduce the score and the pattern of response reported in [3]. Second, a few vocalizations (less than 100!) are necessary for a robotic learning process to provide imitation scores at least as high as the pre-babblers'.

Further, the A and AV imitation experiments revealed a trade-off between the somesthetic acuity of the tract shape representation (GDD) and the amount of information (DBS) to learn in order to build a sensori-motor map that is representative enough of the robot skills. Moreover, our results show that the GDD *has to* be inaccurate enough for the robot to be able to imitate vocalic sounds that are out of its articulatori-acoustic abilities. This is of the utmost interest as, in fact, the infants must acquire, by imitation, the speech sounds of their ambient languages although they are not endowed from birth with the matched motor skills. Last but not least, this investigation supports the view that the formation of the cognitive representation likely to underlie early vocal imitation would require the infants to map less configurations with audiovisual speech perception than without vision. This gives some evidence that the latter can facilitate phonetic development and is congruent with the slight differences in speech development between sighted and blind children.

To sum up, this pioneer work confirms that infants complement their very early visuo-facial imitation map by that of auditory-to-articulatory relationships, and shows that a few data are required to reproduce realistic imitation scores if the tract shape acuity is rough enough.

**Acknowledgements** – We are grateful to Louis-Jean Boë. This study was supported by the following research grants: European ESF Eurocores OMLL, and French MESR Cognitique ABISPA, CNRS STIC Robea and CNRS SHS OHLL

## REFERENCES

[1] Boë, L. J. (1999), "Modelling the growth of the vocal tract vowel spaces of newly-born infants and adults" in *ICPhS99*, San Francisco, 2501-2504 .

[2] Kuhl, P. K., & Meltzoff, A. N. (1982), “The bimodal perception of speech in infancy“, Science218, 1138-1141.

[3] Kuhl, P. K. & Meltzoff, A. N. (1996), “Infant vocalizations in response to speech: Vocal imitation and developmental change”, JASA100, 2425-2438.

[4] Lebeltel, O., et al. (2000), "Bayesian Robot Programming," J. Artificial Intelligence (Submitted).

[5] MacNeilage, P. F. (1998), “The Frame/Content Theory of Evolution of Speech Production,” *BBS21* (4), 499-511.

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

[7] Serkhane, J., Schwartz, J.L, Boë, L.J., Davis, B. & Matyear C. (2002), "Motor specifications of a baby robot via the analysis of infant's vocalizations," *ICSLP2002*, pp.

[8] Studdert-Kennedy, M. (1993). "Some theoretical implications of cross-modal research in speech perception," in Developmental Neurocognition …, ed. de Boysson-Bardies, et al. (pp. 461-466). Kluwer Academic, Dordrecht, The Netherlands.

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

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

*Table 1:* The confusion matrix of early vocal imitation reported in [3]. In columns, the targets. In lines, the phonetic classes of the infants' vowel-like productions

| | i | a | u | Total |
|---|---|---|---|---|
| i-like | 22 | 11 | 4 | 37 |
| a-like | 25 | 66 | 14 | 105 |
| u-like | 20 | 18 | 44 | 82 |
| Total | 67 | 95 | 62 | 224 |

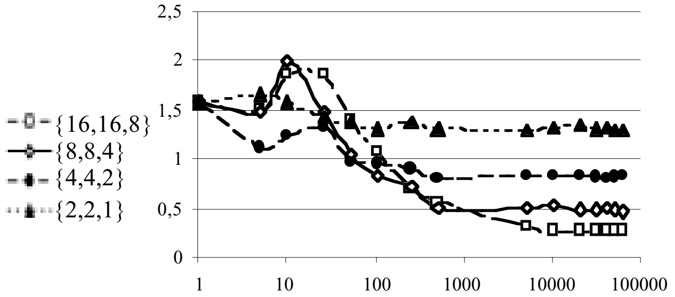

*Figure 1*: Assessing the GDD/DBS trade-off. Mean formant error at the output of the inversion process (in Bk) as a function of the DBS (GDD as parameter).

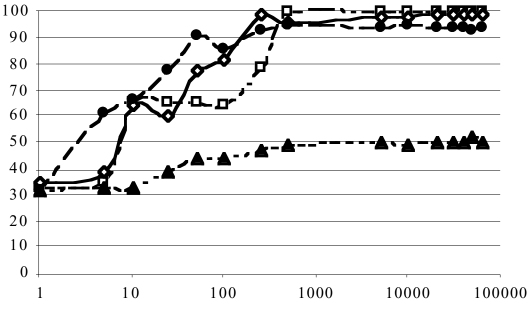

*Figure 2*: %CR for the AV inversion of the “internal” [i a u] vowels, as a function of the DBS (GDD as parameter).

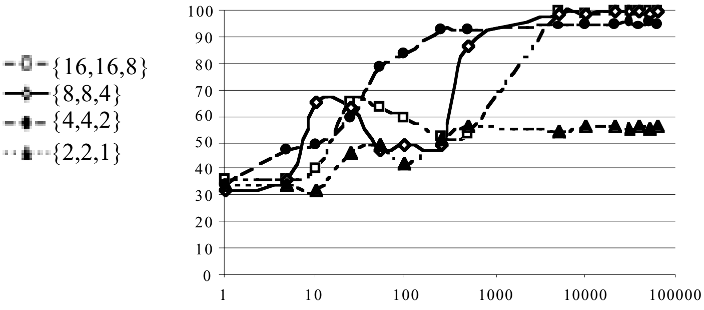

*Figure 3*: %CR for the AV inversion of the “external” [i a u] vowels, as a function of the DBS (GDD as parameter).