INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Assessing idiosyncrasies in a Bayesian model of speech communication

Marie-Lou Barnaud[1,2,3,4], *Julien Diard*[3,4], *Pierre Bessière*[5], *Jean-Luc Schwartz*[1,2]

[1] Univ. Grenoble Alpes, Gipsa-lab, F-38000 Grenoble, France
[2] CNRS, Gipsa-lab, F-38000 Grenoble, France
[3] Univ. Grenoble Alpes, LPNC, F-38000 Grenoble, France
[4] CNRS, LPNC, F-38000 Grenoble, France
[5] SORBONNE Universités - UPMC - ISIR, Paris, France
marie-lou.barnaud@gipsa-lab.grenoble-inp.fr

Abstract

Although speakers of one specific language share the same phoneme representations, their productions can differ. We propose to investigate the development of these differences in production, called idiosyncrasies, by using a Bayesian model of communication. Supposing that idiosyncrasies appear during the development of the motor system, we present two versions of the motor learning phase, both based on the guidance of an agent master: "a repetition model" where agents try to imitate the *sounds* produced by the master and "a communication model" where agents try to replicate the *phonemes* produced by the master. Our experimental results show that only the "communication model" provides production idiosyncrasies, suggesting that idiosyncrasies are a natural output of a motor learning process based on a communicative goal.

Index Terms: speech development, motor learning, Bayesian modeling, idiosyncrasies

1. Introduction

Although speech acquisition is fast and efficient, the mechanisms underlying speech development are quite complex. If we only consider phonetic learning occurring during the first year of life, it can be decomposed in three steps [1, 2]. First, from birth, children learn to associate sounds with the phonemes of their native language. Then, from around seven months, a babbling phase occurs during which children learn to associate acoustic signals with motor gestures. Finally, around two months later, children begin to associate motor gestures with the phonemes of their native language.

These three learning steps, respectively called sensory, sensory-motor and motor learning in the following of this paper, are language specific. Indeed, the exposure to one particular language results in tuning the sensory and motor phonetic representations to this language (in the perceptual domain, this is called perceptual narrowing [3]). As a consequence, children speaking different languages have different phonetic representations.

Conversely, we may expect children speaking the same language to have similar phonetic repertoires. However, there is also intra-language variability, called idiosyncrasies. Typically, in speech production, when two agents produce the same phoneme, acoustic results may vary extensively [1, 4].

In this paper, we focus on the development of idiosyncrasies in speech production and aim at better understanding what component of the learning process could be at their origin. Since idiosyncrasies in production concern the relationship between motor gestures and phonemes, we assume that they appear during the motor learning phase. We compare two computational models of this phase of speech development, both based on an imitation algorithm during which a computational learning agent tries to reproduce speech utterances of a master agent. In the first model, named "repetition model", the agent tries to reproduce *sounds* uttered by the master. In the second one, named "communication model", the agent tries to replicate *phonemes* produced by the master.

Our two motor learning algorithms are embedded inside a Bayesian model of speech communication called COSMO (for "Communicating Objects using Sensory-Motor Operations"), that we have been developing in the past years. COSMO is in our view an efficient framework to study and simulate various aspects of speech communication, including the emergence of sounds systems in human languages [5, 6] or online speech perception [7, 8].

This paper is organized as follows: Section 2 presents the COSMO model and describes the two motor learning models. Section 3 compares results of experimental simulations with the two learning models, which are then discussed in Section 4.

2. COSMO, a Bayesian model of speech communication

2.1. Model description

Within a speech communication process between two agents, a speaker produces motor gestures, that result in acoustic signals perceived by a listener; this enables an exchange of linguistic information between the two agents. From this conceptual description of the communication process, the COSMO model relies on the assumption that communicative agents internalize in their brain all the involved motor, sensory and linguistic representations. In COSMO, these representations are modeled by probabilistic variables: M for motor gestures, S for sensory (acoustic) signals, O_S and O_L for the linguistic "objects" (in a general sense) of communication, O_S relating to the object for the speaker and O_L to the object for the listener, and C for the evaluation of communication success.

Based on the Bayesian Programming methodology [9, 10], the joint probability distribution $P(C\ O_S\ S\ M\ O_L)$ is decomposed as a product of five distributions: a prior on objects $P(O_S)$, a motor system $P(M \mid O_S)$, a sensory-motor system $P(S \mid M)$, an auditory recognition system $P(O_L \mid S)$ and a communication validation system $P(C \mid O_S\ O_L)$. These five distributions are the knowledge of our communicating agent.

http://dx.doi.org/10.21437/Interspeech.2016-396

In this study, we implement a "vowel version" of the COSMO model. It involves the use of an articulatory model of the vocal tract, VLAM (for "Variable Linear Articulatory Model") [11, 12, 13] in which orofacial articulators (jaw, larynx, tongue, lips) are controlled by 7 parameters : one for the jaw, one for the larynx, three for the tongue and two for the lips.

In our model, linguistic units O_S and O_L correspond to the seven vowels /i u e o ɛ ɔ a/, which are the seven preferred vowels in human languages [14]. The motor variable M only retains three parameters of VLAM sufficient for these vowels, that are lip height L_H, tongue body T_B, and tongue dorsum T_D, respectively monitoring vowel rounding, vowel height and vowel anterior/posterior configurations. The sensory variable S consists of formants $F1$ and $F2$ expressed in Barks [15]. We discretize $F1$ and $F2$ respectively into 59 and 73 values, while M contains 15*15*15 values. C is a boolean value, expressing that O_L and O_S are identical or different.

We define the probability distributions of the model. $P(O_S)$ is a uniform distribution: all vowels are equiprobable. $P(S \mid O_L)$, $P(S \mid M)$ and $P(M \mid O_S)$ are conditional Gaussian distributions. To express the lack of knowledge before learning, these distributions are initially set with means in the middle of their space and large variance, approximating uniform distributions. Learning consists in providing values for objects, sensory and motor variables (e.g. o, s and m) in ways that will be explained later. From these values, parameters of the Gaussian distributions $P(S \mid O_L)$, $P(S \mid M)$ and $P(M \mid O_S)$ are updated in a straightforward manner respectively using observed data $\langle s, o\rangle$, $\langle s, m\rangle$ and $\langle m, o\rangle$. Finally, $P(C \mid O_S\ O_L)$ is a "Bayesian switch" [16]: when $C = 1$, O_S and O_L are constrained to the same value.

We previously showed how respectively setting O_S or O_L or both O_S and O_L as the pivot of communication enabled to switch from a motor to an auditory to a sensory-motor theory of speech communication [8, 17]. In this paper, we keep the most general framework, that is a sensory-motor theory of speech production, so that O_L and O_S are always constrained (by $C = 1$) to be equal. Hence, to simplify notations, in the following, we note both O_L and O_S by a single O. This particularly concerns processes in the motor phase (see Eq. (3)) and processes used for the evaluation (see equations in Section 2.3).

2.2. Learning phases

Starting from scratch, we consider the three speech development stages previously introduced: a *sensory learning phase* associating sounds with phonemes, a *sensory-motor learning phase* associating motor gestures with sounds, and a *motor learning phase* associating motor gestures with phonemes. In agreement with other works [1, 2], we consider that these steps are consecutive and performed in interaction with a master agent.

2.2.1. Master agent

The master agent we use in this study disposes of a set of target motor commands for each vowel. These target sets have been defined so as to produce typical formant values for the seven considered vowels, based on data for French vowels [18]. For each vowel, the master agent draws values for M according to a Gaussian distribution around the motor target, with a given variance in the articulatory space. Motor commands are then translated into acoustic values thanks to VLAM. This provides a (non Gaussian) distribution $P(S|O_{mast})$ from which the master draws samples provided to the learning agent.

2.2.2. Sensory and sensory-motor phases

During sensory learning, the agent learns its probability distribution $P(S \mid O_L)$. This learning phase is straightforward: the master produces a linguistic object o resulting in an acoustic signal s, and we assume that the learning agent is able to access both s from its auditory system and o from a given parallel communication stream, e.g. deixis [6]. The learning agent then directly updates its distribution $P([S = s] \mid [O_L = o])$ thanks to the $\langle o, s\rangle$ couple.

During sensory-motor learning, the agent learns its probability distribution $P(S \mid M)$. This phase is a little more complex: as the master agent cannot directly inform the learning agent about the motor gestures it produces, the learning agent needs to infer them. We suppose that inference is based on an imitation process. As in the sensory phase, the master produces a linguistic object o resulting in an acoustic signal s. Then, the learning agent tries to imitate the master by inferring a motor gesture m thanks to the distribution $P(M \mid [S = s])$. The selection of a given motor command m results in the production of a sound s' (computed thanks to VLAM). Of course, s' has no reason to be equal to the target sound s provided by the master. However, the agent exploits this $\langle s', m\rangle$ pair to update its sensory-motor system $P([S = s'] \mid [M = m])$.

2.2.3. Motor phase

Once the sensory-motor learning phase is completed, the motor learning phase begins. During this phase, the learning agent updates its distribution $P(M \mid O_S)$. Although it uses an imitation process similar to the sensory-motor phase, the inference process is different.

We consider two versions of this inference process. In the first version, called "repetition model", the agent attempts to reproduce the exact sound produced by the master for a given object. For this aim, inference is based on the distribution $P(M \mid [O = o]\ [S = s])$, which means: select a motor gesture likely to be associated to the phoneme o and to result in the sound s. In the second version, called "communication model", the agent tries to select a motor gesture likely to ensure communication and hence to realize a vowel o similar to the one produced by the master. For this aim, inference is based on the distribution $P(M \mid [O = o]\ [C = 1])$.

More formally, both distributions $P(M \mid [O = o]\ [S = s])$ and $P(M \mid [O = o]\ [C = 1])$ are computed in the COSMO model using Bayesian inference, which yields:

$$P(M \mid [O = o]\ [S = s]) \propto P(M \mid [O_S = o])P([S = s] \mid M)\ , \quad (1)$$

$$P(M \mid [O = o]\ [C = 1]) \propto P(M \mid [O_S = o])\textstyle\sum_S (P(S \mid M)P([O_L = o] \mid S))\ , \quad (2)$$

where $P(M \mid O_S)$, $P(S \mid M)$, and $P(O_L \mid S)$ are probability distributions of the learning agent. In both versions, the inferred motor gesture m is used to update parameters of the motor system $P([M = m] \mid [O_S = o])$.

2.2.4. Summary of the complete learning sequence

For each motor learning model, we performed 12 simulations in which each learning phase lasted 300,000 steps. Due to random sampling, simulations differed in couples $\langle s, o\rangle$ given by the master and in motor gestures m (and resulting s') selected at each learning step. This enabled to test whether different simulations would result in different final stages at the end of the

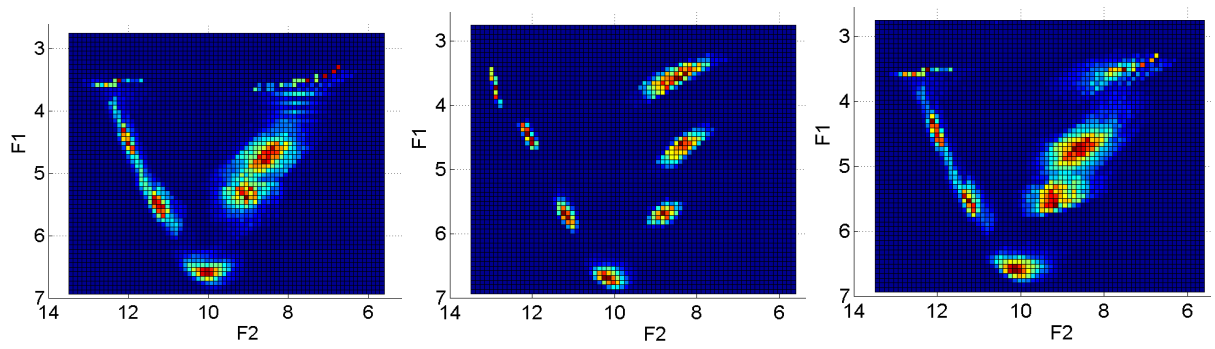

Figure 1: Vowel distributions $P(S \mid O)$. Plots use the classical view of the acoustic space, with F_1 on the y-axis, F_2 on the x-axis, both reversed. Axes values are in Barks. High probabilities are in red, low probabilities in blue. Each region with a color scale from green-yellow to red represents a vowel. **Left:** Distributions $P(S \mid O_{mast})$ of the master; **Middle:** Learned distributions $P(S \mid O_{ag})$ in the "communication model"; **Right:** Learned distributions $P(S \mid O_{ag})$ in the "repetition model".

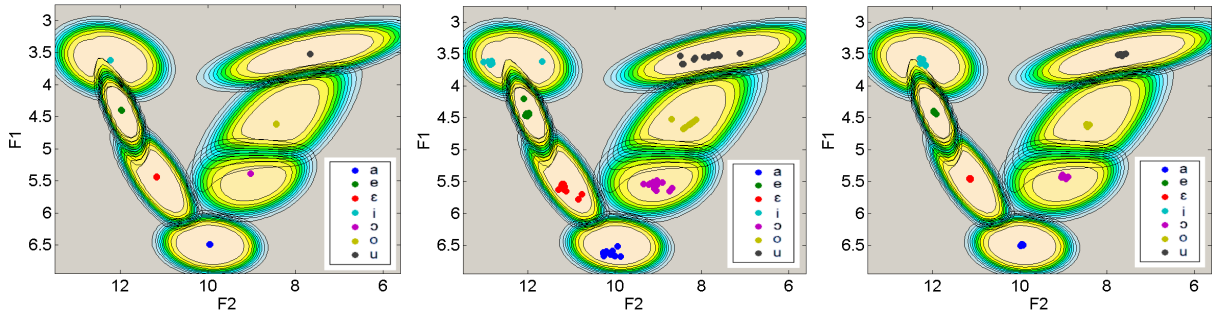

Figure 2: Acoustic space is represented as in Figure 1, with values in Barks. Ellipses in the three plots correspond to the categorization regions of the master (distribution $P(O_{mast} \mid S)$). Points respectively correspond to the means of: **Left:** the master distributions $P(S \mid O_{mast})$; **Middle:** the distributions $P(S \mid O_{ag})$ in the "communication model" for the 12 simulated agents; **Right:** the distributions $P(S \mid O_{ag})$ in the "repetition model" for the 12 simulated agents.

whole learning process, which could possibly provide idiosyncrasies.

2.3. Model evaluation

At the end of the whole learning process, models are evaluated in two ways, assessing both communication performance and possible motor and sensory idiosyncrasies. To assess communication performance, the learning agent tries to communicate an object O_{ag} to the master agent, by producing motor commands resulting in sounds from which the master infers O_{mast}. We compute the confusion matrix $P(O_{mast} \mid O_{ag})$:

$$P(O_{mast} \mid O_{ag}) = \sum_S \left(P(O_{mast} \mid S) P(S \mid O_{ag})\right) \ , \quad (3)$$

where $P(O_{mast} \mid S)$ is the perceptual categorization system of the master, while $P(S \mid O_{ag})$, the sensory result of the productions of the learning agent, is computed by:

$$P(S \mid O_{ag}) \propto \sum_M \left(P(S \mid M) P(M \mid O_{ag})\right) \ . \quad (4)$$

Here, $P(S \mid M)$ is the real motor-to-acoustic transformation provided by VLAM, and $P(M \mid O_{ag})$ is the production process of the learning agent.

3. Results

3.1. Communication performance

We computed the confusion matrix $P(O_{mast} \mid O_{ag})$ (Eq. (3)), at the end of the learning process for each of the 12 simulations for each motor learning model. A global communication performance index was provided by the mean proportion of correct answers for all phonemes, that is the average value of the diagonal of the confusion matrix. The average over the 12 simulations provides 99.1 % of correct recognition in the "communication model" and 98.4 % in the "repetition model". Those two values are quite close and both indicate high performance, illustrating that both motor learning models are able to correctly learn the phoneme repertoires of their master.

To further analyze our results, let us first display the distribution $P(S \mid O_{mast})$ of the master. Figure 1 (left) provides the classical distribution of reference acoustic data [18], where each vowel covers a unique portion of the acoustic space, though with some small overlap at their boundaries.

We also display the distributions $P(S \mid O_{ag})$ for a typical simulation of one learning agent (see Eq. (4)) at the end of learning. The middle and right plots of Figure 1 respectively show an instance of $P(S \mid O_{ag})$ in the "communication" and the "repetition" models. We notice that in both cases, vowels are well defined and distinguishable. However, we notice that while the "repetition" model on the right reproduces the master

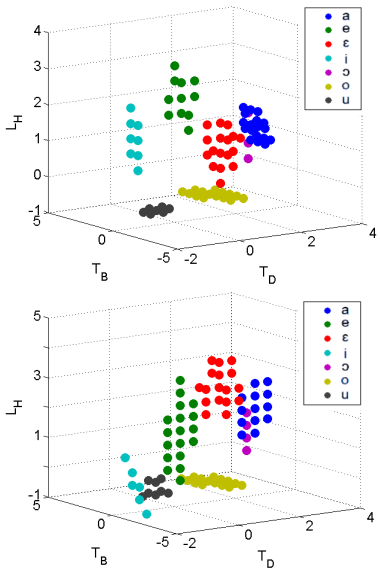

Figure 3: Comparison of the motor distribution $P(M \mid O_S)$ for two simulations of learning agents in the "repetition model": tongue body (T_B) on the x-axis, tongue dorsum (T_D) on the y-axis and lip height (L_H) on the z-axis. Axes values are based on VLAM values. Points in the same color correspond to the same vowel.

distribution accurately, the "communication" model in the central plot provides a distribution clearly different from the master, characterized by both different means and smaller variances.

3.2. Idiosyncrasies

On Figure 2 (left), we display both $P(O_{mast} \mid S)$, i.e. the categorization regions of the master, and the means of $P(S \mid O_{mast})$, i.e. the sensory prototypes of phonemes (see Section 2.2.1)). As expected, prototypes of the master are well centered in each categorization region. This describes the way the sensory space is structured by the distribution of vowels in the master space, acting as a reference for the learning agent.

From this basis, the other plots of Figure 2 show how the 12 simulations of the "communication model" (middle) and the "repetition model" (right) compare to the stimuli provided by the master at the end of the learning stage. These displays were obtained by computing the means of $P(S \mid [O = o])$ (see Eq. (4)) for each vowel o in each of the 12 simulations, for the "communication model" and "repetition model". The 12 corresponding means are shown as colored dots, keeping the master categorization regions as reference.

We observe that idiosyncrasies appear only in the "communication model". Indeed, only in this case do the 12 mean values of $P(S \mid O_{ag})$ vary between simulations. Importantly, despite these idiosyncrasies, the means of each vowel are still in their respective categorization regions, supporting the idea that idiosyncrasies do not alter perceptual categorization, and thus do not alter communication efficiency – as indeed shown by the measured communication performance in the previous section.

In the "repetition model", in contrast, there are no idiosyncrasies: vowel means are not variable from one simulation to the other, and are concentrated around the means of stimuli provided by the master distributions $P(S \mid O_{mast})$. Indeed, it can be mathematically shown that, in this learning algorithm, $P(S \mid O_{ag})$ progressively converges towards $P(S \mid O_{mast})$.

Notice that, even if there are no sensory idiosyncrasies, the "many-to-one" relation from motor to sensory spaces may generate motor idiosyncrasies, since a given sensory percept can result from various different motor gestures. As a matter of fact, we display in Figure 3 distributions $P(M \mid O_S)$ in the motor space for two simulations of the "repetition model". Motor distributions are clearly different. Detailed analyses of simulation results confirm that such motor idiosyncrasies appear in both the "communication" and "repetition" models, even though sensory idiosyncrasies appear only in the first case.

4. Discussion

In this paper, we compared two versions of the motor learning stage in speech development, to investigate idiosyncratic learning in speech production: a "communication model" and a "repetition model". For this aim, we implemented a sequence of learning steps proposed by specialists of speech development [1, 2] into the COSMO model. Our first experimental result is that, in the scope of the phonetic material considered in this paper and involving a small set of oral vowels, COSMO is able to correctly produce learned phonemes whatever the version used.

The second and main result of this study is that idiosyncrasies are only obtained in the "communication model" of motor learning. Since idiosyncratic behaviors are a commonly observed phenomenon, we infer that speech development likely involves some motor learning process guided by a communicative goal, during which children would try to replicate perceived phonemes rather than perceived sounds. Such learning process based on a communicative goal could actually take a wide variety of forms, including communication scenarios based on inverse imitation games (see, e.g. [19]).

The sequence of learning stages within speech development that we considered in the present study could be embedded within a more general scenario based on hierarchical learning, with a first stage guided by sensory representations (our sensory and sensory-motor phases), followed by a second, higher-level stage guided by phonetic representations (our motor phase).

Our model has several limitations. Just to mention one, we only considered learning interaction with a single master, which is unrealistic for child speech development. Simulations with several masters are likely to provide idiosyncrasies also in the "repetition model". However, such idiosyncrasies would be centered on the average of the different masters' productions, and iteration of this process over generations would likely gradually reduce the spread of idiosyncrasies. It is not sure whether that would reflect realistic idiosyncrasies.

Whatever the obvious limitations of this initial study, we believe that the proposed strategy – based on the comparison of different computational architectures within a single computational framework – is promising, in order to assess the role of specific components of the general speech communication model we are aiming at here. The specific component tested here, that is, the existence of a learning process based on efficient communication, will be used in the future developments of COSMO. We are presently working on a more complex implementation of the model with more elaborated linguistic units like syllables.

5. Acknowledgements

Research supported by a grant from the ERC (FP7/2007-2013 Grant Agreement no. 339152, "Speech Unit(e)s").

6. References

[1] P. K. Kuhl, "Early language acquisition: Cracking the speech code," *Nature Reviews Neuroscience*, vol. 5, no. 11, pp. 831–843, Nov 2004. [Online]. Available: http://dx.doi.org/10.1038/nrn1533

[2] P. K. Kuhl, B. T. Conboy, S. Coffey-Corina, D. Padden, M. Rivera-Gaxiola, and T. Nelson, "Phonetic learning as a pathway to language: New data and native language magnet theory expanded (NLM-e)," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 363, no. 1493, pp. 979–1000, Mar 2007. [Online]. Available: http://dx.doi.org/10.1098/rstb.2007.2154

[3] J. F. Werker and R. C. Tees, "Influences on infant speech processing: Toward a new synthesis," *Annual review of psychology*, vol. 50, no. 1, pp. 509–535, 1999. [Online]. Available: http://dx.doi.org/10.1146/annurev.psych.50.1.509

[4] L. Ménard, J.-L. Schwartz, and J. Aubin, "Invariance and variability in the production of the height feature in French vowels," *Speech communication*, vol. 50, no. 1, pp. 14–28, 2008.

[5] C. Moulin-Frier, "Rôle des relations perception-action dans la communication parlée et l'émergence des systèmes phonologiques: étude, modélisation computationnelle et simulations," Ph.D. dissertation, Grenoble, Jun. 2011. [Online]. Available: https://tel.archives-ouvertes.fr/tel-00625453

[6] C. Moulin-Frier, J. Diard, J.-L. Schwartz, and P. Bessière, "COSMO ("Communicating about Objects using Sensory-Motor Operations"): A Bayesian modeling framework for studying speech communication and the emergence of phonological systems," *Journal of Phonetics*, vol. 53, pp. 5–41, 2015.

[7] C. Moulin-Frier, R. Laurent, P. Bessière, J.-L. Schwartz, and J. Diard, "Adverse conditions improve distinguishability of auditory, motor, and perceptuo-motor theories of speech perception: An exploratory Bayesian modelling study," *Language and Cognitive Processes*, vol. 27, no. 7-8, pp. 1240–1263, Sep 2012. [Online]. Available: http://dx.doi.org/10.1080/01690965.2011.645313

[8] R. Laurent, J.-L. Schwartz, P. Bessière, and J. Diard, "A computational model of perceptuo-motor processing in speech perception: Learning to imitate and categorize synthetic CV syllables," in *Proceedings of Interspeech 2013*, F. Bimbot, Ed. Lyon, France: International Speech Communication Association (ISCA), Aug 2013, pp. 2796–2800. [Online]. Available: http://hal.archives-ouvertes.fr/hal-00827885

[9] O. Lebeltel, P. Bessière, J. Diard, and E. Mazer, "Bayesian robot programming," *Autonomous Robots*, vol. 16, no. 1, pp. 49–79, 2004.

[10] P. Bessière, E. Mazer, J. M. Ahuactzin, and K. Mekhnacha, *Bayesian Programming*. Boca Raton, Florida: CRC Press, 2013.

[11] S. Maeda, "Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model," in *Speech production and speech modelling*. Springer, 1990, pp. 131–149.

[12] L.-J. Boë and S. Maeda, "Modélisation de la croissance du conduit vocal," in *Journées d'Études Linguistiques, La voyelle dans tous ses états*, 1998, pp. 98–105.

[13] L. Ménard, J.-L. Schwartz, L.-J. Boë, S. Kandel, and N. Vallée, "Auditory normalization of French vowels synthesized by an articulatory model simulating growth from birth to adulthood," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, p. 1892, 2002. [Online]. Available: http://dx.doi.org/10.1121/1.1459467

[14] J.-L. Schwartz, L.-J. Boë, N. Vallée, and C. Abry, "The dispersion-focalization theory of vowel systems," *Journal of Phonetics*, vol. 25, no. 3, pp. 255–286, 1997.

[15] M. R. Schroeder, B. Atal, and J. Hall, "Objective measure of certain speech signal degradations based on masking properties of human auditory perception," in *Frontiers of speech communication research*. Academic Press, London, 1979, pp. 217–229.

[16] E. Gilet, J. Diard, and P. Bessière, "Bayesian action–perception computational model: Interaction of production and recognition of cursive letters," *PLoS ONE*, vol. 6, no. 6, p. e20387, Jun 2011. [Online]. Available: http://dx.doi.org/10.1371/journal.pone.0020387

[17] M.-L. Barnaud, J. Diard, P. Bessière, and J.-L. Schwartz, "COSMO, a Bayesian computational model of speech communication: Assessing the role of sensory vs. motor knowledge in speech perception," in *Development and Learning and Epigenetic Robotics (ICDL-EpiRob), 2015 Joint IEEE International Conference on*. IEEE, 2015, pp. 248–249.

[18] C. Meunier, "Phonétique acoustique," in *Les dysarthries*, P. Auzou, Ed. Solal, 2007, pp. 164–173. [Online]. Available: https://halv3-preprod.archives-ouvertes.fr/hal-00250272

[19] P. Messum and I. S. Howard, "Creating the cognitive form of phonological units: The speech sound correspondence problem in infancy could be solved by mirrored vocal interactions rather than by imitation," *Journal of Phonetics*, vol. 53, pp. 125–140, 2015.