Interspeech 2018
2-6 September 2018, Hyderabad

COSMO SylPhon: A Bayesian perceptuo-motor model to assess phonological learning

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

[1] Univ. Grenoble Alpes, CNRS, Grenoble INP*, GIPSA-lab, 38000 Grenoble, France
[2] Univ. Grenoble Alpes, CNRS, LPNC, 38000 Grenoble, France
[3] CNRS - Sorbonne Université - ISIR, Paris, France
* Institute of Engineering Univ. Grenoble Alpes

Jean-Luc.Schwartz@gipsa-lab.grenoble-inp.fr

Abstract

During speech development, babies learn to perceive and produce speech units, especially syllables and phonemes. However, the mechanisms underlying the acquisition of speech units still remain unclear. We propose a Bayesian model of speech communication, named "COSMO SylPhon", for studying the acquisition of both syllables and phonemes. In this model, speech development involves a sensory learning phase, mainly concerned with perception development, and a motor learning phase, mainly concerned with production development. We study how an agent can learn speech units during these two phases through an unsupervised learning process based on syllable stimuli. We show that the learning process enables to efficiently learn the distribution of syllabic stimuli provided in the environment. Importantly, we show that if agents are equipped with a bootstrap process inspired by the Frame-Content Theory of speech development, they learn to associate consonants to specific articulatory gestures, providing the basis for consonantal articulatory invariance.

Index Terms: speech development, Bayesian modeling, speech units, coarticulation, articulatory invariance

1. Introduction

It is generally considered that babies first acquire phonetic representations at the syllable level [1, 2] and that phonemes are acquired later [3]. However, the mechanisms underlying this developmental sequence still remain poorly understood.

One approach to study acquisition is computer modeling. Several models have already been proposed with different purposes. Some global models study speech development as a whole [4, 5], while other models focus on the understanding of some aspects of phonological development [6, 7, 8, 9].

However, to our knowledge, a model comparing the joint development of different kinds of speech units does not exist yet. In this paper, we propose a framework able to compare in the same model syllable and phoneme acquisition. It is called "COSMO SylPhon".

In the following, after presenting the COSMO SylPhon model in Section 2 and its detailed implementation in Section 3, we show with simulation results in Section 4 that it is indeed possible to learn acoustic distributions of syllabic inputs in an unsupervised manner, and that this results in efficient learning of vowel distributions. Still, the learning of consonants appears to involve specific motor development. These results highlight the primacy of syllable speech units and illustrate the possibility to use COSMO SylPhon to analyze both kinds of speech units in a sensory-motor framework.

2. COSMO SylPhon: a Bayesian model

COSMO SylPhon is a new extension of the COSMO model of speech communication. COSMO already was used to study the emergence of sound systems in human languages [10], the role of the motor system in speech perception [11, 12, 13] or the acquisition of idiosyncrasies [14].

COSMO SylPhon is a Bayesian model based on the Bayesian Programming methodology [15]. It consists of a main distribution including all variables of the model, called the joint distribution, which is decomposed into a product of distributions, each representing a piece of knowledge of the model (see Fig. 1, top, for a schematic representation).

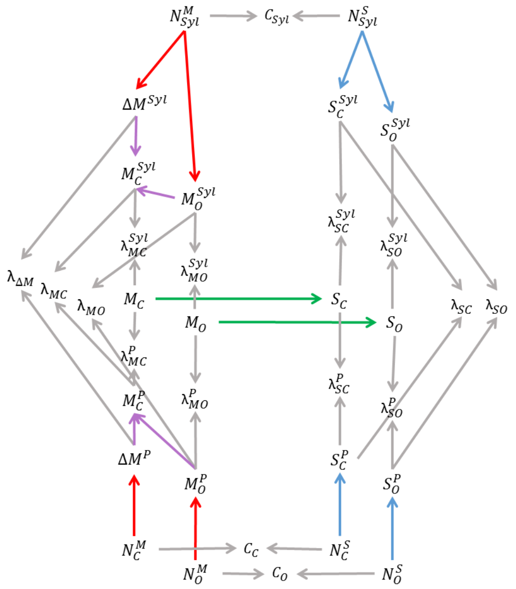
The joint probability space features four kinds of variables: motor variables M considered in this paper as articulatory parameters, sensory variables S restricted to the auditory modality (formants) in this paper, kernels K that are unlabelled categories that the agent learns to associate with speech units in an unsupervised manner, and, finally, coherence variables λ which link and ensure coherence between variables of the same nature.

Each variable can be linked to one specific speech unit: syllables Syl or phonemes P. In the current model, we only consider Consonant-Vowel (CV) syllables. In this way, both syllables and phonemes are composed of consonants C (corresponding to Closed gestures) and vowels O (corresponding to Open gestures). For instance, regarding this notation, M_C^P corresponds to the Motor command for a consonant ("Closed") in the Phonemic part of the model and K_S^O corresponds to a vocalic ("Open") kernel linked to Sensory variables.

The decomposition of the joint distribution is shown in Fig. 1 (bottom). This decomposition has six family of distributions, nearly all present in the initial COSMO model. Prior distributions $P(K)$ and $P(M)$ are prior knowledge over kernels and motor parameters. Motor and sensory repertoires, $P(M \mid K)$ and $P(S \mid K)$ respectively correspond to motor and sensory knowledge related to kernels and required for their production and perception. The internal model $P(S \mid M)$ relates the motor and sensory spaces. Consonantal dependencies $P(M_C \mid \Delta M \; M_O)$ are distributions specific to COSMO SylPhon, which ensure adequate coarticulation between consonants and vowels. They express motor gestures for consonants M_C as a mixture of the motor gesture for the vowel M_O and a specific consonant gesture ΔM, that closes the vocal tract from M_O, and that is supposed independent of the vowel. Finally, coherence distributions associated to the λ variables, not displayed in Fig. 1 (bottom), link variables of the same nature by equality constraints (see [15, 16] for mathematical descriptions).

In the following, we only analyze learning and behavior of four kinds of distributions: prior distributions over kernels, sen-

10.21437/Interspeech.2018-73



$$
\begin{aligned}
&P(K_C^M\ K_C^S\ K_O^M\ K_O^S\ K_{Syl}^M\ K_{Syl}^S\ M_C\ M_O\ \Delta M^P\ M_C^P\ M_O^P\ \Delta M^{Syl}\ M_C^{Syl}\ M_O^{Syl}\\
&\ S_C^P\ S_O^P\ S_C^{Syl}\ S_O^{Syl}\ S_C\ S_O)\\
&\quad = \ P(K_C^M)\ P(K_C^S)\ P(K_O^M)\ P(K_O^S)\ P(K_{Syl}^M)\ P(K_{Syl}^S)\ P(M_C)\ P(M_O) && \text{(a)}\\
&\qquad P(\Delta M^P \mid K_C^M)\ P(M_O^P \mid K_O^M)\ P(\Delta M^{Syl}\ M_O^{Syl} \mid K_{Syl}^M) && \text{(b)}\\
&\qquad P(S_C^P \mid K_C^S)\ P(S_O^P \mid K_O^S) P(S_C^{Syl}\ S_O^{Syl} \mid K_{Syl}^S) && \text{(c)}\\
&\qquad P(S_C \mid M_C)\ P(S_O \mid M_O) && \text{(d)}\\
&\qquad P(M_C^P \mid \Delta M^P\ M_O^P)\ P(M_C^{Syl} \mid \Delta M^{Syl}\ M_O^{Syl}) && \text{(e)}
\end{aligned}
$$

Figure 1: *COSMO-SylPhon. Top: schematic architecture. Bottom: Joint probability distribution and its decomposition, with priors (a), syllable and phoneme motor repertoires (b), syllable and phoneme sensory repertoires (c), internal models relating sensory and motor variables (d) and consonantal dependencies (e). Coherence distributions associated to λ variables are not displayed for the sake of simplicity.*

sory and motor repertoires and internal models. We assume that prior distributions over motor gestures, consonantal dependencies and coherence distributions are initially defined.

3. Simulation principles and tools

3.1. Learning set

Let us first define the dataset of speech units that the agent must acquire during learning. It consists of twenty-one CV syllables composed of seven French vowels /i u e o ɛ ɔ a/ and three French plosive consonants /b d g/. We use them to learn both the phonemic and syllabic parts of the model.

Stimuli are generated and represented in terms of motor and sensory information, thanks to an articulatory model of the vocal tract called VLAM (Variable Linear Articulatory Model), able to transform articulatory parameters into formants [17]. VLAM is composed of seven articulatory parameters, controlling some of the organs of the vocal tract: one for the jaw, one for the larynx, three for the tongue and two for the lips.

Three articulatory parameters are considered sufficient to perform the selected vowels: one parameter for the open/closed movement of the lips (lip height, LH) and two for the movements of the tongue, that are tongue body (TB) for horizontal movements and tongue dorsum (TD) for vertical movements.

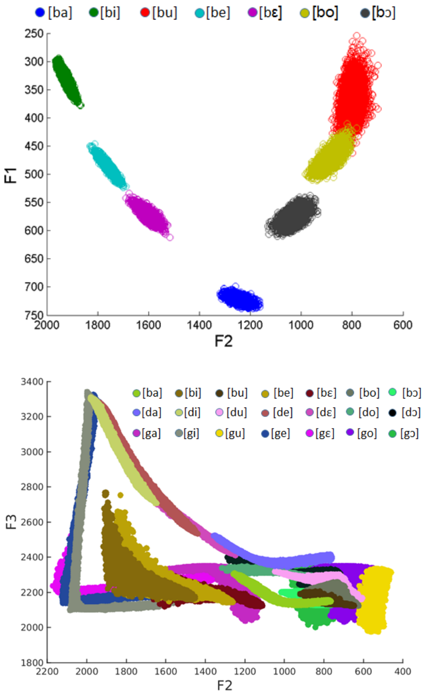


Figure 2: *Distribution of the acoustic input provided by the environment: Top, vocalic part in the (F1, F2) space, in Hz; Bottom, consonantal part in the (F2, F3) space, in Hz. One point represents formants for one consonant or one vowel in a CV sequence.*

Sensory values for vowels (open configurations) are represented in the (F1, F2) vowel space. Regarding plosive consonants, two parameters are used in addition to LH, TB and TD: one for the jaw (J) and one for the tongue apex (TA). Sensory values for plosives (closed configurations) are represented in the (F2, F3) dimensions [18].

Vowels are produced by drawing upon a Gaussian probability distribution in the (TB, TD, LH) space around prototypes for the 7 selected vowels. To realize the 21 corresponding syllables, the 3 consonants are produced by applying specific consonantal gestures on top of the vowel: a combination of jaw J and lips LH for consonant /b/, a combination of jaw J and tongue apex TA for consonant /d/ and a combination of jaw J and tongue dorsum TD for consonant /g/. Formants in the closed configurations are computed just before or after closure. The corresponding distribution of formants for open and closed configurations is shown in Fig. 2, with large variability for consonants due to coarticulation. This is the distribution of acoustic stimuli provided to the agent for learning.

3.2. Learning processes

In models of speech development, learning is generally based on supervised techniques (i.e., [4, 5, 9, 19]). However, babies appear able to learn statistical distributions of their phonetic environment in an unsupervised manner [20]. Some experiments on unsupervised learning of speech units have been performed, mainly focusing on learning the sensory repertoire associated to phonemic units [21, 22, 23, 24]. In this study, we investigate unsupervised learning for the acquisition of motor and sensory repertoires in the case of syllable and phonemic speech units.

3.2.1. *Global structure of the learning process*

In this unsupervised learning process, the agent receives auditory stimuli s resulting from CV syllables. This process involves a sensory learning phase during which the agent learns its sensory kernel priors and its sensory repertoires, and a motor learning phase during which the agent updates first its internal model and then its motor kernel priors and its motor repertoires.

During sensory learning, the agent computes which kernel best corresponds to stimulus s. At each learning step, it infers the best kernel k corresponding to the received stimulus s. Then, it updates the sensory kernel prior for the inferred kernel k and the sensory repertoires with the couple $< s, k >$.

During the first step of motor learning, the agent tries to imitate the sound received from the environment. It starts by inferring the best motor value m corresponding to the received stimulus s. Then, it produces this selected motor command m and perceives the corresponding auditory stimulus s'. It updates its internal model with the couple $< m, s' >$. This algorithm based on imitation is called "accommodation learning".

In the second step, the agent exploits the value m computed from input stimulus s to select the best motor kernel corresponding to m. As in sensory learning, the agent infers the best kernel k corresponding to the motor information m and updates its motor repertoire using the couple $< k, m >$.

These learning phases are performed sequentially for both syllabic and phonemic acquisition. To simplify analyses, syllabic and phonemic learning are considered independent (see details in [25]).

3.2.2. *Implementation*

Based on the learning data presented in Section 3.1 and applying the learning process described in Section 3.2.1, the agent learns all its probability distributions. We use a number of kernels larger than the number of corresponding speech units, with 10 consonantal kernels (for 3 consonantal categories), 50 vocalic kernels (for 7 vowels) and 60 syllabic kernels (for 21 syllables). Concerning distributions to be learned, kernel prior distributions follow Laplace laws of succession (that is, histograms with small residual probability values instead of 0 values). Motor and sensory repertoires as well as internal models are all represented by Gaussian probability distributions, with their classical parameters, that is, their mean vectors and covariance matrices. All variables are discrete.

An important aspect of the COSMO SylPhon implementation concerns consonantal acquisition. It is well known that different motor gestures can result in the same sensory stimulus (many-to-one mapping). Among all the potential motor gestures, just a few are really used (see optimal motor control models [26]). In this study, we question the way in which a baby acquires "optimal" motor gestures, that is, gestures displaying compact representations likely to provide articulatory invariants.

Inspired by the Frame/Content theory [27], we hypothesize that babies perform initial movements that influence learning, shape experience, and guide them to learn optimal motor gestures. We test this hypothesis with COSMO SylPhon in the consonantal case.

As a matter of fact, the learning set defined in Section 3.1 was based on supposedly well-formed consonants, in which the plosives /b d g/ were performed with $DeltaM$ consonantal gestures superimposed to open motor configurations M_O, and involving a single specific articulator per consonant, in addition to jaw movements: LH for /b/, TA for /d/ and TD for /g/. In a similar way, we implement a bootstrap mechanism at the beginning of the learning process, that favors the use of one specific motor parameter per consonant. Specifically, initial configurations for each Gaussian distribution of $DeltaM$ have mean vector values set to 0 in all dimensions, and have a large variance value for a single motor parameter (that is LH, TA or TD) in addition to J, and small variance values for all other parameters. This corresponds to a bootstrap configuration similar to what is described in the Frame-Content framework, with prototypical gestures at the onset of babbling in which jaw closure results in closing the vocal tract either at the region of the lips, at the dental region or at the palatal region. The question we ask in the following simulations is to know whether this initialization enables the agent to discover and replicate the adequate optimal consonantal gestures provided in the learning set.

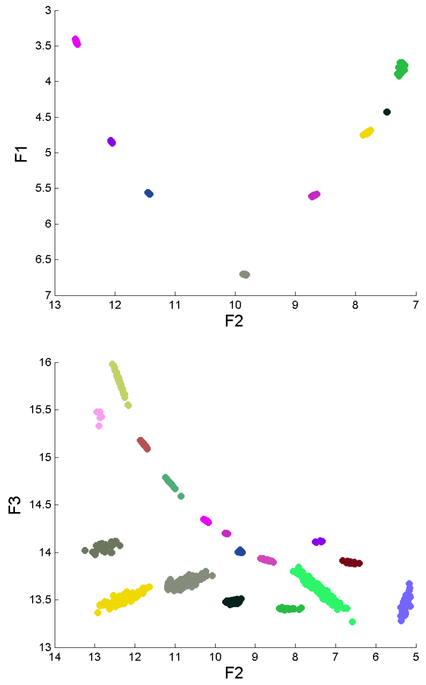


Figure 3: *Most probable kernels recognized for each sensory value in the phonemic portion of the model: Top, vocalic part in the (F1, F2) space, in Barks [28]; Bottom, consonantal part in the (F2, F3) space, in Barks.*

4. Results

4.1. Analysis of sensory learning

We first investigate the sensory portion of the model, that is, distributions of the model acquired during the sensory learning phase. We examine these distributions in both the syllabic and phonemic parts of the model.

To do this, we first analyze convergence of the sensory distributions $P(S)$ for syllables, vowels and consonants to those of the learning set generated with VLAM. We compute the symmetric Kullback-Leibler divergence between the sensory distributions $P(S)$ of the agent and those of the learning set, and verify that it decreases towards 0 during learning. Therefore, all learned distributions converge towards those of the learning set, which means that the agent correctly learns the distributions of the stimuli in the environment. We report in Fig. 3 the sensory repertoires obtained at the end of the learning phase in the phonemic part of the model, by displaying the most probable kernel corresponding to each sensory value provided by the Master. The distributions for vowels show that in the sensory portion of the model, the agent acquires seven main vocalic kernels, each corresponding to one of the seven vowels. Similarly, analysis of the syllabic repertoire (not displayed on the figure) shows that each significant syllabic kernel is linked to one specific syllable, although there are several kernels for each speech unit [25]. However, we observe on the figure that the sensory representations of consonants are less well-defined, since each principal kernel is related to several potential consonants and no clear representation of consonant place of articulation emerges.

4.2. Analysis of consonantal learning in the motor branch

We now examine articulatory gestures learned for consonants, and their relation with the bootstrap mechanism. Bootstrapping should ensure that, in the consonantal space ΔM, motor distributions are spread along specific dimensions for each consonant place of articulation: LH (for /b/), TA (for /d/) or TD (for /g/).

We show in Fig. 4 three 2-D plots, displaying Gaussian distributions related to kernels with significant priors (> 0.01), in the form of contour plots. We observe that each Gaussian is indeed elongated along one dimension, with a large variance along this dimension and small variances in the other dimensions.

Conversely, we notice that this is not found in kernels that have prior probability close to 0 (< 0.01), that are characterized by large variance values in several motor dimensions. This confirms that bootstrapping is indeed necessary for ensuring “optimal” consonantal gestures associated to specific articulatory configurations.

5. Discussion and conclusion

Our experimental results suggest interesting developmental behaviors. First, in each learning phase, we show convergence of the agent’s sensory distributions towards the sensory distribution of the environment. However, it appears that only vowels and syllables seem to be correctly acquired during the sensory learning phase, whereas consonantal representations appear to require motor learning. Since the sensory phase is known to be faster than the motor phase [12], it is consistent with the fact that vowels are acquired before consonants, and suggests that syllables are acquired before phonemes.

Secondly, bootstrapping enables to learn “optimal” consonants with invariant articulatory characteristics. Hence, preference for some specific, stereotypical gestures at the beginning of the learning process could help babies to select “optimal” gestures in the course of speech development.

Importantly, it appears that all units (and particularly consonants and syllables) are represented by several Gaussian distributions, associated with several kernels. In other words, learning has not identified a one-to-one correspondence between speech units and kernels. Interestingly, the presence of several Gaussian distributions to represent a given speech unit could correspond to allophonic variations, that are sub-categories corresponding to the same phoneme but not necessarily used in the same context. It could be the case that these are regularized or grouped together by some other learning process later on, such as, maybe, explicit learning of a written speech code.

In conclusion, the COSMO SylPhon model is able to learn phonetic representations from syllabic inputs, providing some computational basis for the emergence of phonemes. The model suggests potential complementary properties for vowels, that are well learned in the sensory branch, and consonants, that require the development of the motor branch for adequate representation of their articulatory properties. Of course, this model, which extends a previous series of developments of COSMO [12, 13, 14] towards the processing of sequences and temporal information, is just a very preliminary attempt. It deserves extensive work involving natural stimuli, and relating the model architecture to the literature on sequence processing in the human brain.

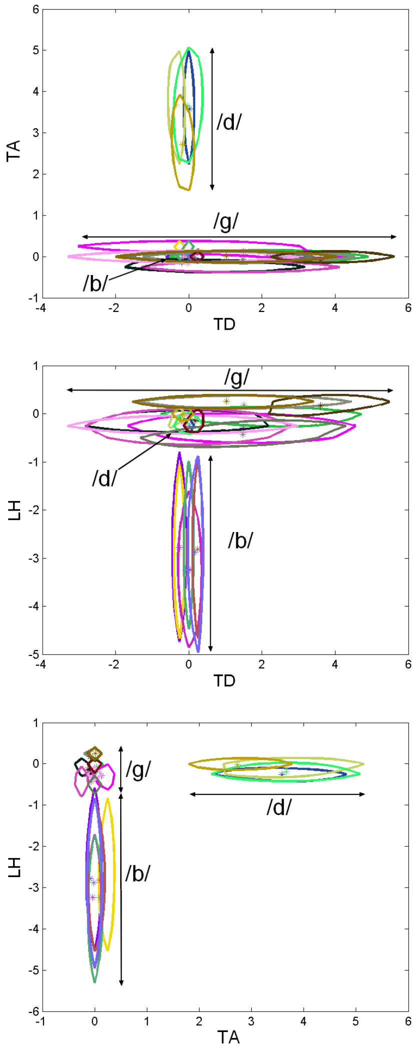


Figure 4: *Illustration of consonantal Gaussian distributions for significant kernels (with prior probability $>$ 0.01), in the dimensions TD, LH and TA of the motor space ΔM. Each ellipse with one color represents the same Gaussian distribution in the three sub-plots.*

6. Acknowledgements

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

7. References

[1] P. Hallé and A. Cristia, "Global and detailed speech representations in early language acquisition," in *Speech production and perception: Planning and dynamics*, S. Fuchs, M. Weirich, D. Pape, and P. Perrier, Eds. Frankfurt am Main: Peter Lang, 2012, pp. 11–38.

[2] O. Rsnen, G. Doyle, and M. C. Frank, "Pre-linguistic segmentation of speech into syllable-like units," *Cognition*, vol. 171, pp. 130 – 150, 2018.

[3] P. K. Kuhl, "Early language acquisition: cracking the speech code," *Nature Reviews Neuroscience*, vol. 5, no. 11, pp. 831–843, Nov 2004.

[4] J. A. Tourville and F. H. Guenther, "The DIVA model: A neural theory of speech acquisition and production," *Language and Cognitive Processes*, vol. 26, no. 7, pp. 952–981, 2011.

[5] B. J. Kröger, J. Kannampuzha, and E. Kaufmann, "Associative learning and self-organization as basic principles for simulating speech acquisition, speech production, and speech perception," *EPJ Nonlinear Biomedical Physics*, vol. 2, no. 1, pp. 1–28, 2014.

[6] S. Peperkamp and E. Dupoux, "Learning the mapping from surface to underlying representations in an artificial language," *Laboratory Phonology*, vol. 9, pp. 315–338, 2007.

[7] C. Moulin-Frier, S. M. Nguyen, and P.-Y. Oudeyer, "Self-organization of early vocal development in infants and machines: the role of intrinsic motivation," *Frontiers in Psychology*, vol. 4, no. 1006, pp. 1–20, 2013.

[8] N. H. Feldman, T. L. Griffiths, S. Goldwater, and J. L. Morgan, "A role for the developing lexicon in phonetic category acquisition," *Psychological Review*, vol. 120, no. 4, pp. 751–778, 2013.

[9] A. K. Philippsen, R. F. Reinhart, and B. Wrede, "Learning how to speak: Imitation-based refinement of syllable production in an articulatory-acoustic model," in *The 4th Joint IEEE International Conference on Developmental and Learning and on Epigenetic Robotics (ICDL-Epirob 2017)*, 2014, pp. 195–200.

[10] 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.

[11] 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, 2012.

[12] R. Laurent, M.-L. Barnaud, J.-L. Schwartz, P. Bessière, and J. Diard, "The complementary roles of auditory and motor information evaluated in a Bayesian perceptuo-motor model of speech perception," *Psychological Review*, vol. 124, no. 5, pp. 572–602, 2017.

[13] M.-L. Barnaud, P. Bessière, J. Diard, and J.-L. Schwartz, "Reanalyzing neurocognitive data on the role of the motor system in speech perception within COSMO, a Bayesian perceptuo-motor model of speech communication," *Brain and Language*, 2017.

[14] M.-L. Barnaud, J. Diard, P. Bessière, and J.-L. Schwartz, "Assessing Idiosyncrasies in a Bayesian Model of Speech Communication," in *Proceedings of Interspeech 2016*, 2016, pp. 2080–2084.

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

[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, 2011.

[17] 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*, W. J. Hardcastle and A. Marchal, Eds. Berlin: Kluwer Academic Publishers, Springer, 1990, pp. 131–149.

[18] J.-L. Schwartz, L.-J. Boë, P. Badin, and T. R. Sawallis, "Grounding stop place systems in the perceptuo-motor substance of speech: On the universality of the labial–coronal–velar stop series," *Journal of Phonetics*, vol. 40, no. 1, pp. 20–36, 2012.

[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.

[20] J. Maye, D. J. Weiss, and R. N. Aslin, "Statistical phonetic learning in infants: Facilitation and feature generalization," *Developmental Science*, vol. 11, no. 1, pp. 122–134, 2008.

[21] B. de Boer and P. K. Kuhl, "Investigating the role of infant-directed speech with a computer model," *Acoustics Research Letters Online*, vol. 4, no. 4, pp. 129–134, 2003.

[22] B. McMurray, R. N. Aslin, and J. C. Toscano, "Statistical learning of phonetic categories: insights from a computational approach," *Developmental Science*, vol. 12, no. 3, pp. 369–378, 2009.

[23] G. K. Vallabha, J. L. McClelland, F. Pons, J. F. Werker, and S. Amano, "Unsupervised learning of vowel categories from infant-directed speech," *Proceedings of the National Academy of Sciences*, vol. 104, no. 33, pp. 13 273–13 278, 2007.

[24] B. Varadarajan, S. Khudanpur, and E. Dupoux, "Unsupervised learning of acoustic sub-word units," in *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies*. Association for Computational Linguistics, 2008, pp. 165–168.

[25] M.-L. Barnaud, "Modélisation bayésienne du développement conjoint de la perception, l'action et la phonologie," Ph.D. dissertation, Université Grenoble-Alpes, 2018.

[26] E. Todorov, "Optimality principles in sensorimotor control," *Nature neuroscience*, vol. 7, no. 9, p. 907, 2004.

[27] P. F. MacNeilage, B. L. Davis, and C. L. Matyear, "Babbling and first words: Phonetic similarities and differences," *Speech Communication*, vol. 22, no. 2-3, pp. 269–277, 1997.

[28] M. R. Schroeder, B. S. 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*, B. Lindblom and S. Öhman, Eds. London: Academic Press, 1979, pp. 217–229.