A computational model of perceptuo-motor processing in speech perception: learning to imitate and categorize synthetic CV syllables

Raphaël Laurent[123], *Jean-Luc Schwartz*[2], *Pierre Bessière*[4], *Julien Diard*[3]

[1]Laboratoire d'Informatique de Grenoble, UMR 5217 CNRS, Université de Grenoble, France
[2]GIPSA-Lab, Département Parole et Cognition, UMR 5216 CNRS, Université de Grenoble, France
[3]Laboratoire de Psychologie et NeuroCognition, UMR 5105 CNRS, Université de Grenoble, France
[4]LPPA, UMR 7152 CNRS, Collège de France, Paris, France

raphael.laurent@gipsa-lab.grenoble-inp.fr, Julien.Diard@upmf-grenoble.fr

Abstract

This paper presents *COSMO*, a Bayesian computational model, which is expressive enough to carry out syllable production, perception and imitation tasks using motor, auditory or perceptuo-motor information. An imitation algorithm enables to learn the articulatory-to-acoustic mapping and the link between syllables and corresponding articulatory gestures, from acoustic inputs only: synthetic CV syllables generated with a human vocal tract model. We compare purely auditory, purely motor and perceptuo-motor syllable categorization under various noise levels.

Index Terms: Speech perception, Bayesian modeling, computational model, sensorimotor fusion.

1. Introduction

It is now more or less accepted that motor information is available during speech perception [1, 2], but two crucial questions remain largely unanswered. (1) What is the nature of motor information; why and when is it useful for perception? (2) How can it be extracted by a listener; how does he/she learn the perceptuo-motor link?

Indeed, a few works have dealt with the fusion of auditory and motor information, either from the point of view of speech recognition (e.g. [3]) or to provide some support to motor or perceptuo-motor theories [4]. But their answers to question (1) are rather vague, as it is not made clear why motor information is useful. Question 2 is solved by providing motor information extracted from speech analysis instruments (electromyography [5], X ray microbeam data [6] or mixtures of laryngography, electroglottography or electropalatography, and electromagnetic articulography [4, 7]) but no real attempt is done to mimic the way a human might learn and exploit the perceptuo-motor link.

The present work is based on *COSMO*, a computational model we have proposed for studying perceptuo-motor interactions in speech perception and production. We propose a plausible developmental scenario for learning the perceptuo-motor link. Then, we use *COSMO* to implement audio [8], motor [9, 10] and perceptuo-motor [11] theories of speech perception, and to compare them on a task of categorization, under various levels of acoustic noise, of synthetic CV syllables generated on a human vocal tract model. We conclude on the feasibility and interest of fusing auditory and motor information in speech perception, particularly in adverse conditions.

2. Generating synthetic CV syllables on an articulatory model of the human vocal tract

2.1. VLAM

We use a realistic vocal tract model, *VLAM*, the Variable Linear Articulatory Model [12]. Seven articulatory parameters (*Jaw, Larynx, TongueBody, TongueDorsum, TongueApex, LipHeight, LipProtrusion*) describe the position of the jaw and larynx, and the shape of the tongue and lips. These parameters can be interpreted in terms of phonetic and muscular commands [13]. The areas of 28 sections of the vocal tract are estimated as linear combinations of these seven parameters, which then allow to compute the transfer function and the formants [14]. Hence, *VLAM* is a geometric model enabling to compute formants from articulatory parameters.

2.2. Generating CV syllables on *VLAM*

We consider the 9 Consonant-Vowel (CV) syllables obtained by combining the most frequent vowels and plosive consonants: /ba/, /bi/, /bu/, /ga/, /gi/, /gu/, /da/, /di/, /du/. Plosive-Vowel syllables are viewed as a pair of two articulatory states, one in which the vocal tract is closed (plosive), and the other one where it is stabilised in a more opened position (vowel). Thus, we assume that syllables are characterised by these two articulatory states only, neglecting the geometry and temporal aspects of the trajectory linking them.

Simplified vowels are described by three *VLAM* articulatory parameters (*TongueBody*, *TongueDorsum* and *LipHeight*), all other parameters being set to a neutral value (resting position). We define motor vowel prototypes for /a i u/, using average formant values for French vowels [15] as targets, and selecting values of the three *VLAM* parameters that best fit to the acoustic target. For each vowel category we generate a set of articulatory configurations by drawing according to a Gaussian probability distribution centered on the prototype value, with a given variance. Configurations with a vocal tract opening too small for a vowel are rejected.

We adopt the view [16] that plosives are local perturbations (vocal tract closing gestures) of vowel configurations within CV syllables. /b/, /d/ and /g/ are stop consonants obtained, from a vowel position, by closing the vocal tract in different places (respectively bilabial, alveolar, and velar). Therefore we synthesize plosives by closing the vocal tract from a vowel position, using the *VLAM Jaw* parameter combined with one other articulator: *LipHeight* for /b/, *TongueApex* for /d/ and *TongueDorsum* for /g/. Hence, plosives are described by five pa-

rameters (*Jaw, TongueBody, TongueDorsum, TongueApex* and *LipHeight*). This choice to model a consonant as a perturbation added to a vowel means that consonants and vowels are linked by maximal coarticulation.

In acoustic space, vowels are characterized by the first two formants $(F1, F2)$ which VLAM computes from the articulatory parameters in the opened state. For plosives, as $F1$ is basically the same ($\approx 250Hz$) for all configurations, characterization is done by $(F2, F3)$. As *VLAM* is a geometrical model, $F2$ and $F3$ are computed when the closed state just starts opening towards the vowel.

Figure 1 displays the generated vowels and plosives, with formant values consistent with other works [17, 18].

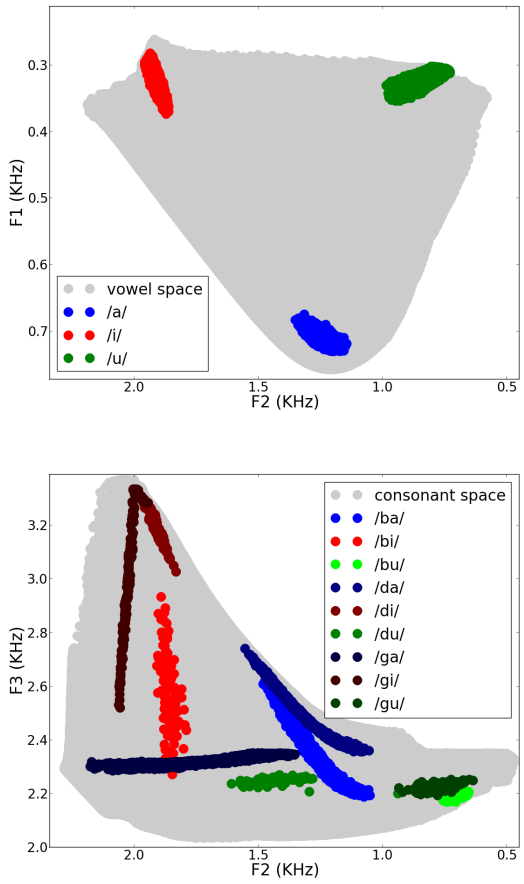


Figure 1: *Synthetic syllables in acoustic space. On top:* (F_2, F_1) *for vowels; below:* (F_2, F_3) *for plosives.*

3. Learning and processing CV sequences in *COSMO*

3.1. *COSMO* and its auditory, motor and perceptuo-motor instantiations

In previous works [19], we have developed a computational model within the framework of Bayesian Programming [20, 21, 22]. We baptized this model *COSMO* for "Communicating Objects using SensoriMotor Operations". It is grounded in the idea that a communicating agent, able to behave both as a speaker and as a listener, has an internal representation of the whole communication situation. The speaker, willing to communicate about the object O_S, performs a motor gesture M producing a sensory percept S enabling the listener to understand and recover an object O_L. Efficient communication C can be assessed by an external validation system (e.g. deixis). *COSMO* is based on a single mathematical object: the joint probability distribution (1) over the variables of interest, which we choose to decompose as follows (2).

$$P(C\ O_L\ S\ M\ O_S) \tag{1}$$

$$= \quad P(O_S)P(M \mid O_S)P(S \mid M)P(O_L \mid S)P(C \mid O_S\ O_L) \tag{2}$$

This model provides a framework unifying purely motor, purely auditive, and perceptuo-motor approaches to speech production and perception [23]. Here we focus on speech perception tasks, which amounts to computing probability distributions of the form $P(O \mid S)$, i.e. the probability distribution over possible messages, given a sensory input.

In a purely auditory approach, this gets instantiated as $P(O_L \mid S)$. Purely auditory perception therefore consists in following a direct association route between sensory inputs and possible messages.

In a purely motor approach, the speech perception task gets instantiated as shown on equation (3):

$$P(O_S \mid S) \propto \sum_M P(M \mid O_S)P(S \mid M) \tag{3}$$

The motor purely motor categorization term $P(O_S \mid S)$ is computed by combining an articulatory decoder $P(M \mid O_S)$ with a forward model $P(S \mid M)$ of the articulatory-to-acoustic transform. This is commonly referred to as "analysis by synthesis".

Finally, in a perceptuo-motor approach, information from the perceptual association route and the motor simulation route is combined. Having C=1 ensures the coherence of O_S and O_L, thanks to which perceptuo-motor perception (4) is expressed as a Bayesian fusion of the motor and auditory answers (5):

$$P(O_S \mid S\ C{=}1) = P(O_L \mid S\ C{=}1) \tag{4}$$

$$\propto \quad P(O_L \mid S) \sum_M P(M \mid O_S)P(S \mid M) \tag{5}$$

3.2. *COSMO* implementation for CV syllables

We extend the *COSMO* model to CV syllables processing. The objects, O_S from the speaker point of view and O_L in a listener perspective, refer to the syllables we consider: /ba/, /bi/, /bu/, /ga/, /gi/, /gu/, /da/, /di/, /du/. Since we model a syllable as a vowel state and a consonant state, variable S separates into S_V and S_C, and variable M into M_V and M_C.

In this extended version of *COSMO* (Figure 2), the motor system (in red), the auditory system (in blue), and the perceptuo-motor system (in green) are linked together by coherence variables λ, which are a mathematical tool we use to force duplicate variables to have the same values at all time during probabilistic inference [24]. This allows to integrate constraints coming from the different submodels into the global model. Likewise, the specification of C in an inference task allows to combine motor and auditory cues.

The motor system describes a state of knowledge of the link between the phonetic objects O_S and articulatory gestures. It involves M'_V, the articulatory configuration of the vowel (*TongueBody*, *TongueDorsum* and *LipHeight* in *VLAM*), G'_C the

articulator used to make a plosive consonant (*LipHeight* for /b/, *TongueDorsum* for /g/, and *TongueApex* for /d/), and Δ'_C the variation of this articulator and of the *Jaw* necessary to achieve a consonant from M'_V. The term $P(\Delta'_{MC} \mid M'_V\ G'_C)$ shows that the consonant is conditioned by the vowel, according to the "perturbation model" described in section 2.2.

The sensorimotor system describes the knowledge the agent has of the articulatory-to-acoustic mapping, i.e. of the mapping between articulatory gestures M_V (vowel) and M_C (consonant), and formant values S_V and S_C. The term $P(M_C \mid M_V)$ encodes a support for consonants achievable from each vowel, according to the perturbation hypothesis.

The auditory system describes the knowledge the agent has of the link between phonetic objects O_L and sensory variables: S'_V (F_1 and F_2 for the vowel) and S'_C (F_2 and F_3 for the consonant).

This work is done under the assumption of an upstream normalization [25]: the formant values the agent can produce (S_V, S_C) and perceive (S'_V, S'_C) are expressed in the exact same acoustic space.

3.3. A developmental scenario for learning model parameters

Some probability distributions of the model are not learned. The priors $P(O_S)$, $P(O_L)$ and $P(M_V)$ are set as uniform probability distributions. The biological constraints $P(M_C \mid M_V)$ describing what consonants are achievable from what vowels are hardcoded into the model. Finally, probability distributions over coherence variables, $P(\lambda_{SV} \mid S_V\ S'_V)$, $P(\lambda_{SC} \mid S_C\ S'_C)$, $P(\lambda_{MV} \mid M'_V\ M_V)$, $P(\lambda_{MC} \mid M'_V\ \Delta'_{MC}\ M_C)$ and $P(C \mid O_S\ O_L)$ are set as Dirac probability distributions, with value 1 when both variables on the right hand side have the same value.

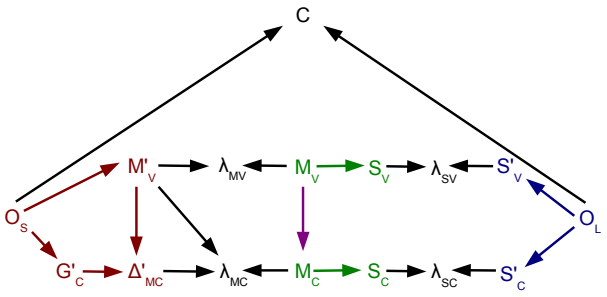
While for infants learning motor, auditory and perceptuo-motor knowledge certainly overlaps, we made the simplifying assumption to separate three learning stages. Our model follows them in an order consistent with a real developmental sequence [26]. 1) Learning auditory categorization. 2) Learning the articulatory-to-acoustic mapping through babbling and imitation. 3) Learning motor categorization.

3.3.1. Learning auditory categorization

The auditory system, linking perceptual stimuli and corresponding syllables, is learned by association. $P(S'_V\ S'_C \mid O_L)$ consists of 9 4-dimensional Gaussian probability distributions (one for each O_L) on the formant space (F_{1V}, F_{2V}, F_{2C}, F_{3C}). They are learned in a supervised manner using all *(formant values; syllable)* pairs taken from the data presented Figure 1.

3.3.2. Learning the sensory-motor link through babbling and imitation

The sensorimotor system is learned without supervision, following an imitation scenario. Given a syllable acoustic target, and using its current state of knowledge, the model carries out an imitation task, by infering a motor gesture (M_V, M_C) likely to reach the target. This gesture is sent to *VLAM*, which here plays the role of an external vocal tract simulator. *VLAM* outputs the formants (S_V, S_C) corresponding to motor command (M_V, M_C), and the model updates its knowledge with the observation that the chosen motor commands produce a given set of formants. This knowledge is stored in $P(S_V \mid M_V)$ and $P(S_C \mid M_C)$, which are Gaussian probability distributions,



$$
\begin{aligned}
&P(O_S G'_C\ M'_V\ \Delta'_{MC}\ \lambda_{MV}\ \lambda_{MC}\ M_V\ M_C\ S_V\ S_C\ \lambda_{SV}\ \lambda_{SC}\ S'_V\ S'_C\ O_L\ C)\\
&\quad = P(O_S) \times P(M'_V \mid O_S) \times P(G'_C \mid O_S) \times P(\Delta'_{MC} \mid M'_V\ G'_C) \times\\
&\qquad P(\lambda_{MV} \mid M'_V\ M_V) \times P(\lambda_{MC} \mid M'_V\ \Delta'_{MC}\ M_C) \times\\
&\qquad P(M_V) \times P(S_V \mid M_V) \times P(M_C \mid M_V) \times P(S_C \mid M_C) \times\\
&\qquad P(\lambda_{SV} \mid S_V\ S'_V) \times P(\lambda_{SC} \mid S_C\ S'_C) \times\\
&\qquad P(O_L) \times P(S'_V\ S'_C \mid O_L) \times\\
&\qquad P(C \mid O_S\ O_L)\ .
\end{aligned}
$$

Figure 2: *Dependency structure of COSMO, described as a graphical model (on top), and by its joint probability distribution (below).*

evolving through the learning process.

The syllable targets to imitate are taken from the data presented on Figure 1. At a given step, whether the target has been reached or not is not taken into account. Each new observation guarantees that the motor inversion, i.e. the process driving the choice of motor gestures allowing to imitate auditory inputs, becomes more and more accurate. This imitation scenario we propose can be viewed as target-oriented babbling.

3.3.3. Learning motor categorization

The motor system is learned in a supervised way, i.e. syllable labels are given. But while in other works [3, 4] the articulatory data is provided, here we provide the model with labeled acoustic data. We use the same *(formant values; syllable)* pairs used to learn auditory categorization in stage 1), and we use the perceptuo-motor link learned in stage 2) to retrieve motor information. Given an acoustic target and the corresponding syllable, the model infers a motor gesture allowing to reach the target, by inversing the articulatory-to-acoustic mapping, and by using the state of knowledge at that time of the correspondence between syllables and motor gestures. The chosen motor gesture is then used to update parameters of the following probability distributions: the Gaussian probability distribution $P(M'_V \mid O_S)$, the histogram $P(G'_C \mid O_S)$, and the Gaussian probability distribution $P(\Delta'_{MC} \mid M'_V\ G'_C)$.

3.4. Processing CV syllables in *COSMO*

3.4.1. Evaluation

At this point, we can compare motor categorization, auditory categorization and their perceptuo-motor fusion. We use the data already used for the learning processes, and add to acoustic inputs (S_V, S_C) various levels of white noise on a perceptive scale using Barks [27]. Then we assess recognition scores $P(O \mid S_V\ S_C)$ from these degraded noisy inputs. For each

noise level, and for each model, we compute the correct recognition rate as the average over all stimuli of the probability of it being attributed to the right category according to this model. The objective is to quantify how well the motor, auditory and perceptuo-motor versions of the model generalize, by measuring their robustness to noise.

3.4.2. Simulations

A comparison of the different model correct recognition rates over various noise levels is shown Figure 3.

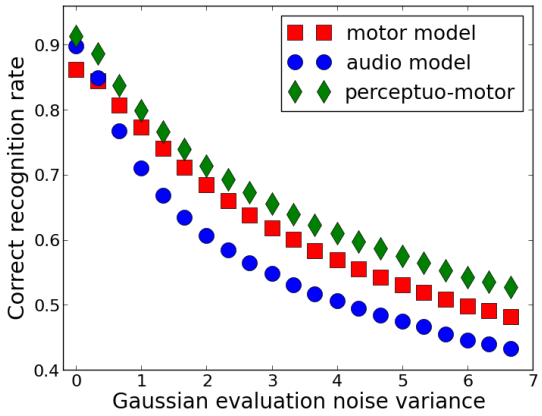


Figure 3: *Motor, audio and perceptuo-motor model robustness to noise in a syllable classification task.*

When there is no noise, the audio model is more accurate than the motor model. But as soon as there is some noise, the motor model, more robust, performs better than the audio model. Motor knowledge brings robustness to noise: this is due to the summation, in the inference, over all possible motor gestures (see section 3.1). The perceptuo-motor model performs always better than both audio and motor models. Motor knowledge complements audio knowledge, allowing a higher performing sensorimotor fusion. The overall good performance of the motor model validates our learning-by-imitation algorithms.

4. Conclusion

These results are to be interpreted in a more general framework, where it has been shown on very simple cases [19] that auditory representations are likely to be more accurate than articulatory ones, particularly when the articulatory-to-acoustic mapping is highly nonlinear, while motor representations are more robust in adverse conditions. This paper generalizes these results to CV syllable categorization and shows that motor knowledge brings robustness to noise and, complementing auditory knowledge, allows a better performing sensori-motor fusion. This work extends and complements a recent series of modelling attempts to integrate perception and action in a coherent computational framework [28, 29].

The target-oriented babbling algorithm learning the sensory-motor link allows the model to acquire and encode motor information from acoustic inputs only. A *VLAM* version including vocal tract growth during learning [30], combined with appropriate auditory normalization, could enable to simulate developmental changes related to growth of the articulatory system [31].

In the current version of the model, O-M and O-S mappings are learned independently, in a supervised way, given the number of classes. These hypotheses are unrealistic, and further work will focus on exploring different learning paradigms, where motor and auditory systems are co-constructed in parallel (as proposed in the Perception-for-Action-Control Theory (PACT) [11]).

5. References

[1] M. J. Pickering and S. Garrod, "An integrated theory of language production and comprehension," *Behavioral and Brain Sciences*, (in press).

[2] F. Pulvermüller and L. Fadiga, "Active perception: sensorimotor circuits as a cortical basis for language," *Nature Reviews. Neuroscience*, vol. 11, no. 5, pp. 351–360, 2010.

[3] L. Deng and J. Ma, "Spontaneous speech recognition using a statistical coarticulatory model for the vocal-tract-resonance dynamics," *The Journal of the Acoustical Society of America*, vol. 108, no. 6, pp. 3036–3048, 2000.

[4] C. Castellini, L. Badino, G. Metta, G. Sandini, M. Tavella, M. Grimaldi, and L. Fadiga, "The use of phonetic motor invariants can improve automatic phoneme discrimination," *PLoS ONE*, vol. 6, no. 9, p. e24055, 2011.

[5] S.-C. Jou, L. Maier-Hein, T. Schultz, and A. Waibel, "Articulatory feature classification using surface electromyography," in *Proceedings of the 2006 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2006)*, vol. 1, 2006, pp. I–I.

[6] J. Sun, L. Deng, and X. Jing, "Data-driven model construction for continuous speech recognition using overlapping articulatory features," in *Proceedings of the sixth International Conference on Spoken Language Processing (ICSLP 2000)*, vol. 1, 2000, pp. 437–440.

[7] A. Wrench and K. Richmond, "Continuous speech recognition using articulatory data," in *Proceedings of the sixth International Conference on Spoken Language Processing (ICSLP 2000)*, vol. 4, 2000, pp. 145–148.

[8] R. L. Diehl, A. J. Lotto, and L. L. Holt, "Speech perception," *Annual Review of Psychology*, vol. 55, no. 1, pp. 149–179, 2004.

[9] A. M. Liberman and I. G. Mattingly, "The motor theory of speech perception revised," *Cognition*, vol. 21, no. 1, pp. 1–36, 1985.

[10] C. A. Fowler, "An event approach to the study of speech perception from a direct-realist perspective," *Journal of Phonetics*, vol. 14, no. 1, pp. 3–28, 1986.

[11] J.-L. Schwartz, A. Basirat, L. Ménard, and M. Sato, "The Perception-for-Action-Control Theory (PACT): A perceptuo-motor theory of speech perception," *Journal of Neurolinguistics*, vol. 25, no. 5, pp. 336–354, 2012.

[12] 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 modeling*, W. Hardcastle and A. Marchal, Eds. Kluwer Academic, 1990, pp. 131–149.

[13] S. Maeda and K. Honda, "From EMG to formant patterns: the implication of vowel spaces," *Phonetica*, vol. 51, pp. 17–29, 1994.

[14] P. Badin and G. Fant, "Notes on Vocal Tract Computation," in *Quarterly Progress and Status Report, Dept for Speech, Music and Hearing, KTH, Stockholm*, 1984, pp. 53–108.

[15] C. Meunier, "Phonétique acoustique," in *Les dysarthries*, P. Auzou, Ed. Solal, 2007, pp. 164–173.

[16] S. E. G. Öhman, "Coarticulation in vcv utterances: Spectrographic measurements," *The Journal of the Acoustical Society of America*, vol. 39, no. 1, pp. 151–168, 1966.

[17] J.-L. Schwartz, L.-J. Boë, P. Badin, and T. 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, pp. 20–36, 2012.

[18] H. M. Sussman, D. Fruchter, J. Hilbert, and J. Sirosh, "Linear correlates in the speech signal: the orderly output constraint," *The Behavioral and Brain Sciences*, vol. 21, no. 2, pp. 241–259; discussion 260–299, 1998.

[19] C. Moulin-Frier, R. Laurent, P. Bessière, J.-L. Schwartz, and J. Diard, "Adverse conditions improve distinguishability of auditory, motor and percep-tuo-motor theories of speech perception: an exploratory Bayesian modeling study," *Language and Cognitive Processes*, vol. 27, no. 7-8 Special Issue: Speech Recognition in Adverse Conditions, pp. 1240–1263, 2012.

[20] P. Bessière, C. Laugier, and R. Siegwart, Eds., *Probabilistic Reasoning and Decision Making in Sensory-Motor Systems*, ser. Springer Tracts in Advanced Robotics. Berlin: Springer-Verlag, 2008, vol. 46.

[21] F. Colas, J. Diard, and P. Bessière, "Common bayesian models for common cognitive issues," *Acta Biotheoretica*, vol. 58, pp. 191–216, 2010.

[22] O. Lebeltel, P. Bessière, J. Diard, and E. Mazer, "Bayesian robot programming," *Autonomous Robots*, vol. 16, p. 49–79, 2004.

[23] C. Moulin-Frier, J.-L. Schwartz, J. Diard, and P. Bessière, "A unified theoretical bayesian model of speech communication," in *Advances in Applied Digital Human Modeling*, ser. Advances in Human Factors and Ergonomics Series, V. Duffy, Ed. Boca Raton, Florida: CRC Press, Taylor & Francis Group, 2010, pp. 457–466.

[24] 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.

[25] 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, p. 1892, 2002.

[26] P. K. Kuhl, "Early language acquisition: cracking the speech code," *Nature reviews neuroscience*, vol. 5, no. 11, pp. 831–843, 2004.

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

[28] R. K. Moore, "Presence: A human-inspired architecture for speech-based human-machine interaction," *IEEE Transactions on Computers*, vol. 56, no. 9, pp. 1176–1188, 2007.

[29] B. J. Kröger, J. Kannampuzha, and C. Neuschaefer-Rube, "Towards a neurocomputational model of speech production and perception," *Speech Communication*, vol. 51, no. 9, pp. 793 – 809, 2009.

[30] L.-J. Boë and S. Maeda, "Modélisation de la croissance du conduit vocal. Espace vocalique des nouveaux-nés et des adultes. Conséquences pour l'ontogenèse et la phylogenèse," *Journées d'Études Linguistiques, La voyelle dans tous ses états*, pp. 98–105, 1998.

[31] D. E. Callan, R. D. Kent, F. H. Guenther, and H. K. Vorperian, "An auditory-feedback-based neural network model of speech production that is robust to developmental changes in the size and shape of the articulatory system," *Journal of Speech, Language and Hearing Research*, vol. 43, no. 3, p. 721, 2000.