Towards Biological Plausibility of Vocal Learning Models: a Short Review

Silvia Pagliarini
Inria Bordeaux Sud-Ouest, Talence, France.
LaBRI, UMR 5800, CNRS, Bordeaux INP,
IMN, UMR 5293, CNRS,
Université de Bordeaux, France.
silvia.pagliarini@inria.fr

Arthur Leblois*
IMN, UMR 5293, CNRS,
Université de Bordeaux, France.
arthur.leblois@u-bordeaux.fr

Xavier Hinaut*
Inria Bordeaux Sud-Ouest, Talence, France.
LaBRI, UMR 5800, CNRS, Bordeaux INP,
IMN, UMR 5293, CNRS,
Université de Bordeaux, France.
xavier.hinaut@inria.fr

***Abstract*—Sensorimotor learning represents a challenging problem for artificial and natural systems.**

Several computational models try to explain the neural mechanisms at play in the brain to implement such learning. These models have several common components: a motor control model, a sensory system and a learning architecture. Our challenge is to build a biologically plausible model for song learning in birds including neuro-anatomical and developmental constraints. In this review, we thus focus on a specific type of sensorimotor learning referred to as imitative vocal learning and examplified by song learning in birds or human complex vocalizations. We aim to compare the various approaches used in existent sensorimotor models relevant for our purpose and to place them in a common framework.

I. Imitative vocal learning

Humans and animals such as songbirds show imitative vocal learning. Imitation involves the production of the motor command corresponding to a given experienced sensory stimulus. For instance, a baby imitating a word or a bird imitating a syllable of its tutor. Imitative vocal learning is characterized by several phases [1]: (i) a sensory phase enables juveniles to build a neural representations of adult vocalizations, which will guide later vocal production [2]; (ii) in the sensorimotor phase, the juvenile then adapts its vocal output to imitate previously heard sounds.

During this process, the brain must harness sensory feedback to adaptively modify performance in reference to the object of imitation [3]. Interestingly, some neurons, called mirror neurons, called mirror neurons, shows a similar response during the perception and the production of a vocal or motor gesture. Convergence of sensory and motor responses in individual neurons points to a possible mechanism through which auditory and motor signals may be linked to enable vocal learning. More precisely, it suggests the presence of internal models, such as invers or forward model, built through experience and used for vocal production in the brains of humans and songbirds [4], [5], [6].

Here we review existing models linked to vocal imitation. Specifically, we are interested in how internal models are learned and motor control is defined. Using an idea of Oudeyer [7], we can identify a motor space (corresponding to the articulatory parameters, e.g. the tongue height), a sensory space (corresponding to the physical parameters, such as the frequency of a sound) and the perceptual space (corresponding to the information sent to the brain when perceiving a stimulus, such as a minor number of formants in speech). The perceptual map may be connected with the motor map via an artificial neural network to drive production. Alternatively, vocal production is driven by an internal goal ([8]).

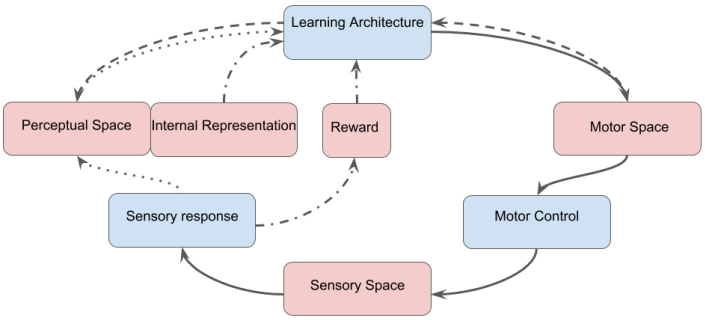

Fig. 1: **Sensorimotor model components** A model for sensorimotor learning includes an input layer, that lies in the perceptual space or conveys the internal representation of a goal. The representation of the vocal output in the motor space is connected to its representation in the sensory space via the motor control apparatus. This sensory output evokes a sensory response. This process is represented by the plane line in the figure. In inverse models, the sensory response is linked to the perceptual input of the model (dotted line). Alternatively, in reinforcement learning models an internal representation of the goal is used as input while the sensory response drives a reward signal (dashed-dot line). Finally, dashed lines show the connection in the case of using a forward model.

Figure 1 summarizes how these three spaces are connected through a learning architecture, a motor control and a sensory system in a sensorimotor model. The properties of the sensory system, shaping the perceptual representation, is beyond the

* Corresponding authors that co-supervised the study.

TABLE I: Model summary

	Topic	Motor space	Motor control	Sensory space	Sensory response	Perceptual space / Internal representation	Learning architecture
Warlaumont 2016 [9]	Humans	Muscle activity time series (900ms)	Praat (lip and jaw)	Sound	Auditory salience.	N.A.	Auditory salience, Random RNN via Izhikevich's DA-modulated STDP
Philippsen 2014 [10]	Humans	26 param. ($[-1,1]^{26}$)	VocalTractLab (Birkholz [11])	39 formants ($[-1,1]^{39}$)	Nonlinear function	Acoustic representation	ESN, non-dynamic ELM
Oudeyer 2005 [7]	Humans (multi-agent exp.)	3 param. ($[0,1]^3$)	de Boer model [12]	4 formants (4D space)	Nonlinear sensory response	2D space	Correlation rule
Moulin-Frier 2014 [13]	Humans	7 param.	VLAM [14]	2 formants ($[-1,1]^2$)	NNA	Ep. Mem.	Different exploration strategies
Moulin-Frier 2014 [15]	Humans	13 param.	DIVA	2 formants ($[-1,1]^2$), intensity ($[0,1]$)	Mean value	6D-vector (formants and intensity)	Active-goal exploration strategy
Cohen 2018 [8]	Humans	Gesture (3 joint ang.) or word (2 syll.)	N.A.	N.A. (identical to motor space)	Caregiver choice	Internal goal (desired object)	PerAc [16] with reward kernels
Teramoto 2017 [17]	Marmoset	3 params. ($[0,1.1]^2$, const.)	Resonant vocal tract	Sound	N.A.	N.A.	Weight change influenced by development rate
Doya 1998 [18]	Songbird	4 params.	Source-filter model with amplifier	Sound waveform	N.A.	N.A.	Stochastic controller and a critic
Our model [19]	Songbird	3 params. $[0.5,1.5]^3$	N.A.	N.A. (identical to motor space)	Nonlinear auditory response	$[0,1]^3$	Normalized Hebbian rule

Ep.Mem: Episodic memory of previous experience; DMP: Dynamical Movement Primitives; VLAM: Vocal Linear Articulatory Model; DIVA:Directions Into Velocities of Articulators; GMM: Gaussian Mixture Model; ELM: Extreme Learning Machine; NNA: Nearest Neighbour Algorithm; N.A.: Non Applicable

scope of the present review and we cill focus on the learning architecture and motor control In Table I, we summarize the proerties of sensorimotor models for vocal learning in humans and animals cited throughout the review.

II. Learning architecture

The starting point of a sensorimotor learning model is the input layer. As shown in Fig. 1, it can be given by the definition of the perceptual space or by the internal representation of a goal. The motor output in the motor space is connected with its sensory representation via the motor control system. Finally a sensory response translates the output of the motor control into the input representation. The differentiation in the input and the connections between spaces depend on the type of model one wants to implement and the learning algorithm used. Reinforcement learning mechanisms can be implemented defining the internal representation of the goal and the reward (dashed-dot line in Fig. 1). The introduction of internal models adds the definition of the perceptual space. Inverse models have the aim to provide an appropriate motor command for a given perceptual goal, which is driven by the sensory response (dotted lines in Fig. 1). Forward models describe a causal relationship between motor commands and their corresponding perceptual representations (dashed lines in Fig. 1).

Several learning architectures have been proposed in the literature that connect the input (perceptual or internal goal) to the motor output (in the motor space) through an artificial neural network. The structure of the network varies between models. A Random Recurrent Neural Network (called a *reservoir*) have been used in Warlaumont and Finnegan [9] and Philippsen et al. [10]. Some other approaches also use neural networks as a modeling architecture, but not a recurrent one ([18], [7], [19]).

The learning rule used to update connections between nodes in the network also differ between models, spacing from reinforcement learning algorithms to associative learning rules. Warlaumont and Finnegan [9] implement reinforcement learning based on auditory salience. The salience of a sound is computed from the sound spectrogram ([20]) that determines the value of the reward signal. Izhikevich's DA-modulated STDP [21] updates the connections between the reservoir and

motor neurons [1]. Doya and Sejnowski [18] update connections using a stochastic gradient ascent algorithm and taking account critic evaluation. Associative learning rules are usually used for building internal models (inverse or forward). For instance, Oudeyer [7] uses a Hebbian correlation rule involving the mean activation of neurons over a certain time interval [22].

A theoretical model of inverse learning in songbird has been proposed by Hahnloser and Ganguli [23]. We recently did an implementation [19] of this theoretical model and we studied how variations in non-linearity and learning rules affect performance. The model includes two neural populations (motor and auditory neurons). Learning is driven by a postdictive Hebbian learning rule. Our model makes an implementation of such theoretical model and extend it. It introduces nonlinearity in the sensory function, which is coherent with the fact that sensory responses in brain are sparse and, indeed, non-linear. In particular, to represent selective responses as observed in various high sensory brain areas (e.g. auditory regions of the pallium in birds display responses selective to tutor syllables or to the bird's own syllables), the auditory activity is defined as a bell-shaped function around a target motor pattern. Learning is driven by a normalized Hebbian learning rule.

Forward and inverse models can also be used together, as in Philippsen et al [10]. He moves from supervised self-training (thanks to the availability of a forward model) to unsupervised learning when imitation of novel contexts is includes (after the training).

Sensorimotor learning requires a phase of motor exploration, whether in reinforcement learning model or for the implementation of internal (e.g. inverse) models. Different strategies for motor exploration have been studied in the context of vocal learning or in other types of sensorimotor learning. The simpliest exploration mechanism is driven by random motor exploration [13], while moresophisticated exploration include random and active motor babbling [24], intrinsically motivated goal exploration [25], or [26], active-goal exploration [15].

III. Motor control

The first step in defining motor control is to choose an appropriate model mapping a motor space onto a sensory space (e.g. muscle command to sound/acoustic representation). Then, we need an appropriate parameter space to describe motor articulations (ideally as function of time).

Air pressure causes vocal folds vibration, which results into a sound wave output. The sound source is the combination between the output of vocal folds vibration and noise. The latter can be due to pressure fluctuations or by the activity of other components of the apparatus (e.g. the glottis). Downstream from the sound source, the vocal tract acts as a resonator, filtering the sound as it travels to the outside world. It modifies the original sound wave and changes the balance between its frequency components. The resonances of the vocal tract are called formants [27]. A basic model of speech production therefore includes a sound source (vocal folds) and a linear acoustic filter (vocal tract) [28].

A sound wave as source is a common starting point of many works, such as [10], [29], [18], [15] and [13]. Usually sound source is modeled taking inspiration from the mas model of sound production [30], coupling vocal fold tension and air pressure.

The vocal tract has been often modeled as a structure of tubes in literature. For instance Boersma in [31] modeled the vocal tract as a structure of tubes with moving walls. This synthesizer has been used by Warlaumont and Finnegan [9]. "VocalTractLab" system developed by Birkholz [11] defines the vocal tract using several cylindrical sections. This has been used by Philippsen et al. to generate the sound in a vocal learning model [10]. It has also been used in speech signal filtering [32] or articulatory synthesizer training [33]. Amador et al. [29] and Doya and Sejnowski [18] model the vocal tract dynamics using ordinary differential equations and a filter. The output is the pressure needed to generate the sound. Amador et al. [29] has been used by [17] with marmoset. Moulin-Frier et al. [15] [13] model vocal production using VLAM (Vocal Linear Articulatory Model [14]) or DIVA(Directions Into Velocities of Articulators).

The output of motor control defines the sensory space. It can be defined by the sound wave, as in [9], [17] and [18]. The sensory space can also be defined by the formants, obtained tuning the parameter of the motor space (for instance the input of VLAM and DIVA model, as in the works from Moulin-Frier et al. [15] [13]). More simplistic models do not define the motor control system, leading to a coincidence between motor and sensory space [8], [19].

IV. Discussion

This short review shows that it is difficult to compare the diversity of existing models. One needs to go into the details of each components of the model and the a priori assumptions. In particular, it is not always possible to clearly identify each component: for instance, sometimes the internal representation (e.g. goal) substitutes the perceptual space. The precise features used to represent internal and perceptual space seems to be specific to each model: the comparison between models is thus complex.

The dimensions of the sensory, perceptual and motor spaces also greatly vary among models: thus the learning architectures do not deal with the same task complexity. Actually, the choice of the architecture may constraint the authors to reduce the space dimensions if the architecture could not deal with high-dimensional spaces.

The various models try to answer different questions. However, we believe it could be useful to have a common model capable to answer multiple questions and enable the comparison of different experiments. Moreover, it should include more and more the general biological mechanisms. Some models include a developmental aspect (that we did not include in our summarizing table) which changes model parameters over time. In general, modelers tend to define motor control inspired

[1] Actually, they are using only the long term potentiation.

by preexisting models having some biological foundations. However, learning architectures do not always take inspiration from biology.

Acknowledgment

This work was supported by the Inria CORDI-S PhD fellowship grant.

References

[1] M. S. Brainard and A. J. Doupe. What songbirds teach us about learning. *Nature*, 417(6886):351, 2002.

[2] A. J. Doupe and P. K. Kuhl. Birdsong and human speech: common themes and mechanisms. *Annual review of neuroscience*, 22(1):567–631, 1999.

[3] R. Mooney. Neural mechanisms for learned birdsong. *Learning & Memory*, 16(11):655–669, 2009.

[4] E. Oztop, M. Kawato, and M. Arbib. Mirror neurons and imitation: A computationally guided review. *Neural Networks*, 19(3):254–271, 2006.

[5] J. F. Prather, S. Peters, S. Nowicki, and R. Mooney. Precise auditory–vocal mirroring in neurons for learned vocal communication. *Nature*, 451(7176):305, 2008.

[6] N. Giret, J. Kornfeld, S. Ganguli, and R. HR Hahnloser. Evidence for a causal inverse model in an avian cortico-basal ganglia circuit. *PNAS*, 111(16):6063–6068, 2014.

[7] PY Oudeyer. The self-organization of speech sounds. *Journal of Theoretical Biology*, 233(3):435–449, 2005.

[8] L. Cohen and A. Billard. Social babbling: The emergence of symbolic gestures and words. *Neural Networks*, 2018.

[9] A. S. Warlaumont and M. K. Finnegan. Learning to produce syllabic speech sounds via reward-modulated neural plasticity. *PloS one*, 11(1):e0145096, 2016.

[10] 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 *ICDL-Epirob, 2014*, pages 195–200. IEEE, 2014.

[11] P. Birkholz. Vocaltractlab–towards high-quality articulatory speech synthesis, 2014.

[12] B. De Boer. *The origins of vowel systems*, volume 1. Oxford University Press on Demand, 2001.

[13] C. Moulin-Frier and PY Oudeyer. Curiosity-driven phonetic learning. In *Development and Learning and Epigenetic Robotics (ICDL), 2012 IEEE International Conference on*, pages 1–8. IEEE, 2012.

[14] 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*, pages 131–149. Springer, 1990.

[15] C. Moulin-Frier, S. M. Nguyen, and P. Oudeyer. Self-organization of early vocal development in infants and machines: the role of intrinsic motivation. *Frontiers in psychology*, 4:1006, 2014.

[16] P. Gaussier and S. Zrehen. Perac: A neural architecture to control artificial animals. *Robotics and Autonomous Systems*, 16(2-4):291–320, 1995.

[17] Y. Teramoto, D. Y. Takahashi, P. Holmes, and A. A. Ghazanfar. Vocal development in a waddington landscape. *eLife*, 6:e20782, 2017.

[18] K. Doya and T. J. Sejnowski. A computational model of birdsong learning by auditory experience and auditory feedback. In *Central auditory processing and neural modeling*, pages 77–88. Springer, 1998.

[19] S. Pagliarini, X. Hinaut, and A. Leblois. A bio-inspired model towards vocal gesture learning in songbird. In *ICDL-Epirob, 2018*. IEEE, 2018.

[20] M. Coath and S. L. Denham. The role of transients in auditory processing. *Biosystems*, 89(1-3):182–189, 2007.

[21] E. M. Izhikevich. Solving the distal reward problem through linkage of stdp and dopamine signaling. *Cerebral cortex*, 17(10):2443–2452, 2007.

[22] T. J. Sejnowski. Storing covariance with nonlinearly interacting neurons. *Journal of mathematical biology*, 4(4):303–321, 1977.

[23] R. Hahnloser and S. Ganguli. Vocal learning with inverse models. *Principles of Neural Coding*, pages 547–564, 2013.

[24] S. Forestier and PY Oudeyer. Curiosity-driven development of tool use precursors: a computational model. In *38th annual conference of the cognitive science society (cogsci 2016)*, pages 1859–1864, 2016.

[25] S. Forestier, Y. Mollard, and PY Oudeyer. Intrinsically motivated goal exploration processes with automatic curriculum learning. *arXiv preprint arXiv:1708.02190*, 2017.

[26] A. Baranes and PY Oudeyer. Active learning of inverse models with intrinsically motivated goal exploration in robots. *Robotics and Autonomous Systems*, 61(1):49–73, 2013.

[27] P. Ladefoged. *Elements of acoustic phonetics*. University of Chicago Press, 1996.

[28] G. Fant. *Acoustic theory of speech production: with calculations based on X-ray studies of Russian articulations*, volume 2. Walter de Gruyter, 2012.

[29] A. Amador, Y. S. Perl, G. B. Mindlin, and D. Margoliash. Elemental gesture dynamics are encoded by song premotor cortical neurons. *Nature*, 495(7439):59, 2013.

[30] I. R. Titze. Mechanical stress in phonation. *NCVS Status and Progress Report*, 4:291–301, 1993.

[31] P. P. G. Boersma et al. *Functional phonology: Formalizing the interactions between articulatory and perceptual drives*, volume 11. Holland Academic Graphics The Hague, 1998.

[32] J. Gudhnason, D. D. Mehta, and T. F. Quatieri. Evaluation of speech inverse filtering techniques using a physiologically based synthesizer. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 4245–4249. IEEE, 2015.

[33] S. Prom-on, P. Birkholz, and Y. Xu. Training an articulatory synthesizer with continuous acoustic data. In *INTERSPEECH*, pages 349–353, 2013.