Vocal Imitation in Sensorimotor Learning Models: a Comparative Review

Silvia Pagliarini, Arthur Leblois, Xavier Hinaut

Vocal Imitation in Sensorimotor Learning Models: a Comparative Review

Silvia Pagliarini, *1, 2, 3*, Arthur Leblois *, *3*, and Xavier Hinaut *, *1, 2, 3.*

1. INRIA Bordeaux Sud-Ouest, Bordeaux, France. 2. LaBRI, Université de Bordeaux, Institut Polytechnique de Bordeaux, Centre National de la Recherche Scientifique, UMR 5800, Talence, France. 3. Institut des Maladies Neurodégénérative, Université de Bordeaux, France, Centre National de la Recherche Scientifique, UMR 5293, Bordeaux, France.

***Abstract*—Sensorimotor learning represents a challenging problem for natural and artificial systems. Several computational models have been proposed to explain the neural and cognitive mechanisms at play in the brain. In general, these models can be decomposed in three common components: a sensory system, a motor control device and a learning framework. The latter includes the architecture, the learning rule or optimisation method, and the exploration strategy used to guide learning. In this review, we focus on imitative vocal learning, that is exemplified in song learning in birds and speech acquisition in humans. We aim to synthesise, analyse and compare the various models of vocal learning that have been proposed, highlighting their common points and differences. We first introduce the biological context, including the behavioural and physiological hallmarks of vocal learning and sketch the neural circuits involved. Then, we detail the different components of a vocal learning model and how they are implemented in the reviewed models.**

***Index Terms*—Sensorimotor learning, imitative learning, vocal imitation, reinforcement learning, associative learning, inverse model, forward model, songbird, bird song, neural networks, exploration strategy, mirror neurons.**

I. Introduction

HUMANS and animals such as songbirds show imitative vocal learning: they are able to produce a motor command that replicates a previously experienced auditory stimulus [1]–[4]. Imitation implies a causal relationship between the observed stimulus and the produced action, and requires a mechanism to translate the sensory input into motor commands [2]. Humans and animals are able to perform complex imitation, this is illustrated by the imitation of novel action sequences in response to environmental cues [3].

Imitative vocal learning, and more generally sensorimotor learning, are the subject of behavioural, anatomical, physiological and computational studies. Taking into account the biological evidence and constraints revealed by experimental investigations of the underlying brain circuits, many previous studies have attempted to implement imitative learning in computational models. The aim of this review is to identify and compare the various components of existing vocal learning models to provide an integrated and organised view of the literature. While we focus our analysis on vocal learning, the principles addressed here may also apply to sensorimotor learning models in general. To analyse and compare the existing models, we will now define the core components at play in models of vocal learning.

As depicted in Figures 1 and 2, the representations needed for a minimal vocal learning model can be cast into three spaces [5]: motor, sensory and perceptual/internal space. In addition, one needs to define a learning framework and define the connections between the spaces: a motor control function and a sensory response function. The learning framework contains the architecture, the learning algorithm, the evaluation and the exploration strategy (see Table IV). We define the input and output spaces of the learning algorithm as the *learning domain* and the *learning image*[1]. The motor space corresponds either to the muscle activation patterns sent to the vocal organ (e.g. larynx and syrinx for human and birds respectively) or articulatory parameters (e.g. the tongue height) for humans. The sensory space, in the case of vocal learning, represents the physical space of the sound. The perceptual space corresponds to the neural representation of perceived vocalisations in the brain (e.g. acoustic features as pitch in birdsong or first formants in speech). Space representations are implemented as vectors or trajectories (i.e. sequences of vectors) in these multi-dimensional spaces.

Figure 1 shows the canonical model including an action-perception loop: the perceptual space is connected to the motor and sensory spaces through a sensory system and a motor control device. Depending on the modeller's choice, the perceptual and motor spaces may be linked through an inverse model, or both an inverse and a forward model (see definition of internal models and in particular inverse and forward model in Section VI). The learning domain (i.e. input space of the learning algorithm) is the perceptual space in the case of inverse models, and the motor space in the case of forward models.

An internal representation of the goal could lie in the perceptual space, or alternatively as shown in Figure 2, in a separated space if it is encoded independently of the sensory processes of experienced vocalisations. In such a model, an internal representation of the goal is used as the learning domain and hence, it is *non-perceptual*. In the present review, we call *goal-to-motor model* the connections between the

* Corresponding authors that co-supervised the review.
Manuscript draft on November 25th, 2020.

[1] The idea is to conceptualise the learning algorithm as a mathematical function going from the domain, called *learning domain*, to its co-domain, called *learning image*. For simplicity, we will use *learning image* instead of *learning co-domain*.

internal representation and the motor space. The sensory processing of the produced vocalisations may still be implemented downstream from the motor space, for instance to provide a reward signal that guides learning in a reinforcement learning framework.

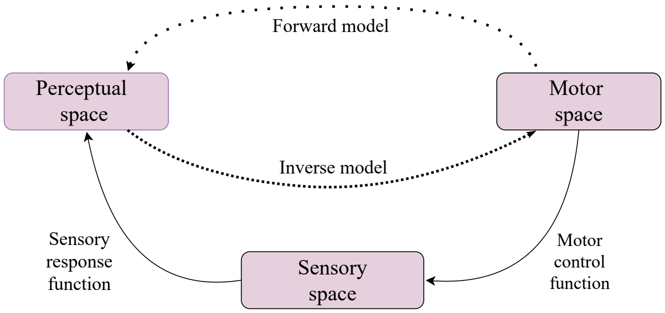


Fig. 1. **Sensorimotor model with a action-perception loop.** The *motor control function* generates a sensory representation (a sound in the more complete models) given the motor command parameters. This kind of sensorimotor model includes an inverse model, and potentially a forward model. One of the advantages of a forward model is that it can bias the perceptual representation in order to facilitate the inverse model learning towards perceptuo-motor representations [6].

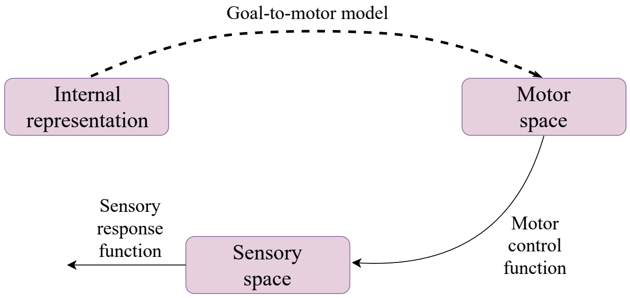
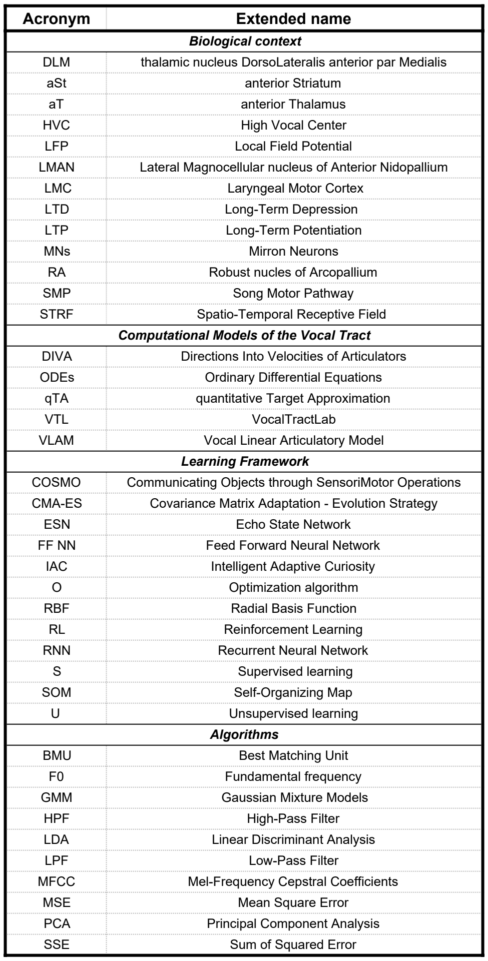


Fig. 2. **Non-perceptual sensorimotor model.** This kind of model is *non-perceptual* because it has a non-perceptual internal representation of goals. The dotted line represent learned connection from goals to motor commands: we call this the *goal-to-motor model*. Sensory response function processes the sound and can be implemented in various ways depending on the learning framework: it could be used to provide a reward or an evaluation of the learning (for this reason there is an arrow starting from the sensory space, but without a specific output space).

Whichever the particular learning framework and mechanisms used, the modelled agent must explore either the goal space or the motor space to later adjust its production. Such a vocal exploration may be purely random or more sophisticated (e.g. intrinsically motivated exploration) [7]. To explore either the motor space or the goal space, and to improve current vocal performance, learning models rely on the evaluation of the produced vocalisation. The aim of the evaluation is to obtain a measure that defines an error signal and/or a reinforcement signal, later used by the learning framework to update the architecture.

Table I contains all the acronyms used along the review. Section II contains an introduction to the neuroanatomy of the human and songbird brains. Additionally, it contains an analysis of the links between biology and the sensorimotor components. Section III details the aim of the reviewed models, giving an overview of the objectives and questions pursued by the modellers. Table II summarises the aims of the models. Section IV describes the motor control device and its components: the motor space, the articulatory model and the connection with the sensory space. Section V introduces the representation of the sensory system and its components: the sensory space, the sensory response function and the perceptual space. Table III contains a summary of the spaces and functions of sensorimotor models. Section VI elaborates on the components of the learning framework and Table IV summarises the implementations used in the models. Section VII contains a discussion about the reviewed models, their relation with the biological framework introduced in Section II, and further directions are proposed.

TABLE I
ACRONYMS: SUMMARY TABLE OF THE ACRONYMS USED IN THE REVIEW.



Acronym	Extended name
Biological context	
DLM	thalamic nucleus DorsoLateralis anterior par Medialis
aSt	anterior Striatum
aT	anterior Thalamus
HVC	High Vocal Center
LFP	Local Field Potential
LMAN	Lateral Magnocellular nucleus of Anterior Nidopallium
LMC	Laryngeal Motor Cortex
LTD	Long-Term Depression
LTP	Long-Term Potentiation
MNs	Mirron Neurons
RA	Robust nucles of Arcopallium
SMP	Song Motor Pathway
STRF	Spatio-Temporal Receptive Field
Computational Models of the Vocal Tract	
DIVA	Directions Into Velocities of Articulators
ODEs	Ordinary Differential Equations
qTA	quantitative Target Approximation
VTL	VocalTractLab
VLAM	Vocal Linear Articulatory Model
Learning Framework	
COSMO	Communicating Objects through SensoriMotor Operations
CMA-ES	Covariance Matrix Adaptation - Evolution Strategy
ESN	Echo State Network
FF NN	Feed Forward Neural Network
IAC	Intelligent Adaptive Curiosity
O	Optimization algorithm
RBF	Radial Basis Function
RL	Reinforcement Learning
RNN	Recurrent Neural Network
S	Supervised learning
SOM	Self-Organizing Map
U	Unsupervised learning
Algorithms	
BMU	Best Matching Unit
F0	Fundamental frequency
GMM	Gaussian Mixture Models
HPF	High-Pass Filter
LDA	Linear Discriminant Analysis
LPF	Low-Pass Filter
MFCC	Mel-Frequency Cepstral Coefficients
MSE	Mean Square Error
PCA	Principal Component Analysis
SSE	Sum of Squared Error

II. Biological context

We present here the biological context of vocal learning. We first highlight the behavioural phases included in imitative vocal learning in humans and songbirds. Then, the main brain circuits related to song (for birds) and spoken language (for humans) are discussed and compared. Finally, we introduce current mechanistic hypotheses and some biological constraints that should be taken into account while defining a vocal learning model: mirror neurons' activity and their putative function in vocal learning, experimental evidence for synaptic plasticity and the sensory representation of vocalisations. In the last three subsections, the literature comes mainly from songbirds, but may also serve as biological support for human studies.

A. Learning phases and behaviour

From a behavioural point of view, speech learning in humans and song acquisition in birds are made up of the same developmental behavioural phases [4], [8], [9]. Figures 3 and 4 show the first year of speech perception (green background) and production (pink background) development in infants (adapted from [4]) and songbirds (adapted from [9]). In babies, as shown in Figure 3, sensory learning starts immediately after birth and allows the infant to discriminate the phonetic contrasts specific to the learned language. This process, also known as categorical learning, is described in Subsection II-C. Vocal production starts with the production of non-speech sounds, also shortly after birth. After this preliminary phase, sensorimotor learning starts: speech-like sounds are first produced erratically, then "canonical babbling" emerges and the first words are produced by the infant around the age of one year [4], [10].

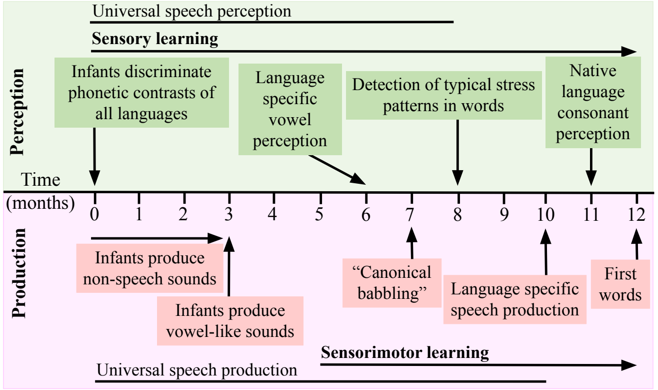

Universal speech perception
Sensory learning
Perception
Infants discriminate phonetic contrasts of all languages
Language specific vowel perception
Detection of typical stress patterns in words
Native language consonant perception
Time
(months)
0
1
2
3
4
5
6
7
8
9
10
11
12
Production
Infants produce non-speech sounds
Infants produce vowel-like sounds
"Canonical babbling"
Language specific speech production
First words
Sensorimotor learning
Universal speech production

Fig. 3. **First year of infant speech-perception and speech-production development.** Speech perception development (green background) is characterised by a sensory learning phase that shapes perception, from an initially universal perception to language-specific phoneme discrimination. Speech production development (pink background) is characterised by some preliminary phases followed by sensorimotor learning, where "canonical babbling" takes place. Image adapted from Kuhl (2004) [4].

In birds, as shown in Figure 4, the sensory learning phase enables juveniles to build a neural representation of adult vocalisations, which would later guide vocal production[8]. Juveniles have a species-specific predisposition and listen to the sounds produced by their parents[9]. Then, during the sensorimotor phase, the young birds start to vocalise, initially producing babbling sounds [11]–[13] and then adapting their vocal output to imitate previously heard vocalisations. Finally, the produced vocalisation becomes more and more stereotyped and vocal plasticity significantly drops. This final phase, when song production converges towards the stereotyped adult song, is called crystallisation in birds [8], [9].

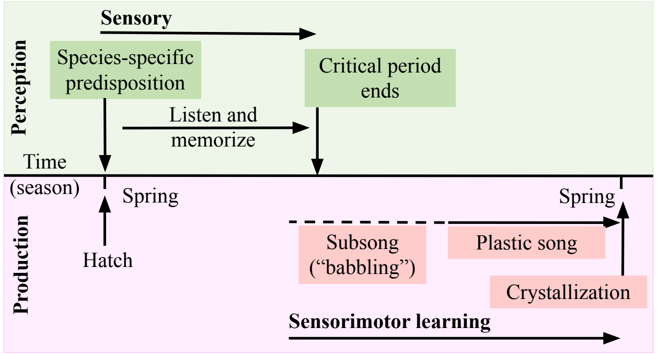


Fig. 4. **Imitative learning phases in birds.** Three main phases characterise imitative learning in songbirds: the sensory learning phase, the sensorimotor learning phase (starting with subsong and continuing with a plastic song), and crystallization of the song (i.e. convergence to adult song). Image adapted from Doupe and Kuhl (1999) [8].

B. Neuroanatomy of human and bird brain

Figure 5 shows the brain pathways controlling song in songbirds (upper panel) and spoken language in humans (lower panel). In both cases, there are two main pathways [14]: the posterior vocal motor pathway (plain black arrows) and the anterior vocal learning pathway (plain white arrows). In addition, there are connections between the two pathways (dashed black arrows) and specialised direct projection to vocal motor neurons (plain red arrows).

The vocal motor pathway in birds (Figure 5a) projects from HVC (used as a proper name[2]) to robust nucleus of arcopallium (RA) (plain black arrows, upper panel). RA and its analogous in humans, represented by the laryngeal motor cortex (LMC), connect directly to vocal motor neurons (plain red arrows, lower panel) providing the motor output (controlling the larynx in humans or syrinx in birds) [14], [15].

The vocal learning pathway is responsible for vocal imitation and plasticity: it forms a basal ganglia-thalamo-cortical loop. In birds, as shown in Figure 5a, it involves the song-related song nucleus Area X, the thalamic nucleus dorsolateralis anterior pars medialis (DLM, sometimes called aDLM) and the lateral magnocellular nucleus of the anterior nidopallium (LMAN, more generally called MAN). The indirect projection onto RA from Area X, through DLM and LMAN is represented by a dashed black arrow [14], [15]. In humans, in Figure 5(b), the vocal learning pathway presumably includes Broca's area (one of the main areas of the language cortex in humans along with Wernicke's area and superior temporal gyrus [16]), the anterior striatum (aSt) and anterior thalamus (aT).

The neuroanatomical structure of the vocal control circuit in human and bird provides the anatomical basis for bio-inspired models of vocal learning that often question the function of specific brain areas and/or the connections between them. Please refer to Section III and Table II for studies making explicit reference to the neuroanatomy of the brain.

[2] HVC was originally High Vocal Center.

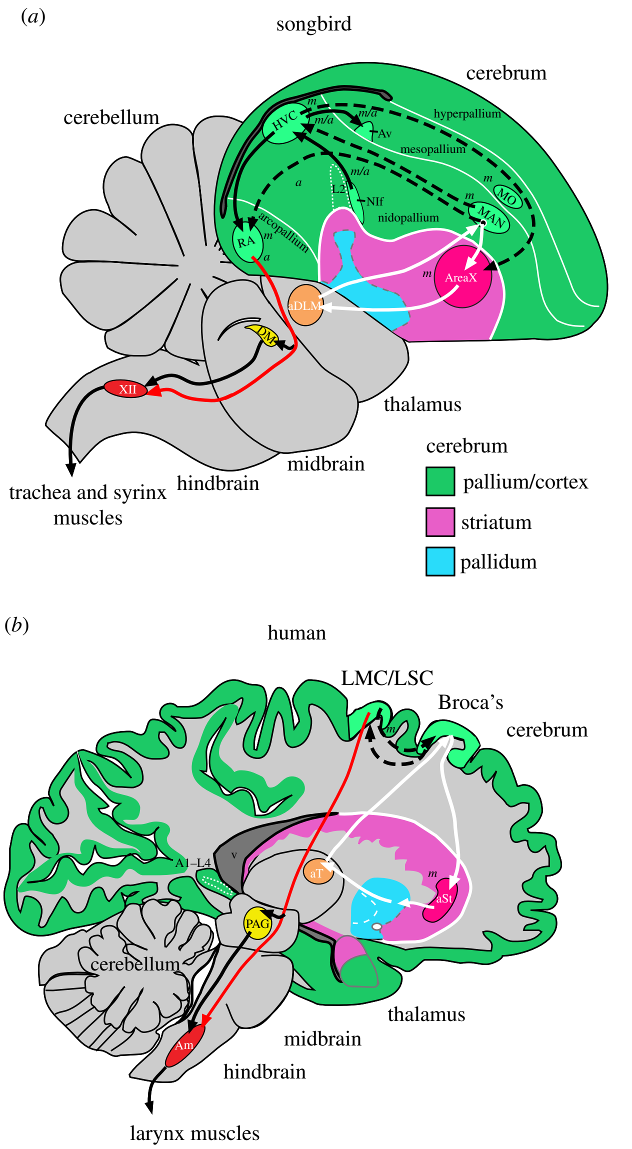


Fig. 5. **Brain pathways controlling (a) song in songbirds and (b) spoken language in humans.** The posterior vocal motor pathway (plain black arrows) is also called vocal production pathway, since it involves direct projections to motor neurons. The anterior vocal learning pathway (plain white arrows) is responsible for vocal imitation and plasticity. In addition, there are the connections between the two pathways (dashed black arrows) and specialised direct projection to vocal motor neurons (plain red arrows). In panel (a), *a* indicates a region where there is an auditory neural activity, *m* a region where there is motor neural activity, *m/a* a region where both auditory and motor neural activities are present. In panel (b), *v* indicates the ventricle space. Image from Chakraborty and Jarvis (2015) [14], CC-BY 4.0 license.

C. Sensory system

The auditory system of mammals and birds builds up selective responses to auditory stimuli. Ultimately, auditory selectivity may give rise to categorical perception, the tendency to perceive a continuous change in sensory space (e.g. sound) as discrete percepts (e.g. phonemes or bird syllables). This ability exists in both humans and birds [4], [9]. During infant first months, the acoustic differences detected influence the selection of phonetic units, and infants become more sensitive to the units that are important for the language they hear [4], [9]. Similarly in birds, neural selectivity for the imitated song develops slowly during song ontogeny [1].

In birds, the auditory system is involved in the discrimination of songs, and relies on the temporal cues and pitch of the song to provide information about the identity of the singer [17]. Song-selective responses (with different responses to the bird's own song and other's vocalisations) have been observed in various high sensory brain areas. The sharp auditory selectivity of neurons in these high sensory areas emerge from a multi-stages auditory pathway that starts from the inner ear. At lower stages of this pathway, sensory responses evoked by the playback of songs or other sounds (including white noise) are well modelled using a linear summation of spatio-temporal receptive fields (STRF) [18]. Higher in the auditory pathway, responses become sparser and more non-linear (i.e. less well modellable by such a linear model) [17].

Interestingly, experimental studies in birds have revealed that some auditory neurons also respond to perturbations of the auditory feedback during singing [17], [19]. This highlights the fact that the whole pathway from high auditory area to motor areas could be involved in the recognition of tutor or conspecific songs and in the evaluation of the bird's auditory feedback.

D. Mirror neurons and perceptuo-motor coherence

Some neurons, called mirror neurons (MNs) show a similar response during the perception and the production of a motor or vocal gesture [20]–[23]. Convergence of sensory and motor signals in the same neurons points to a possible mechanism to enable vocal learning [2]: during vocal production, auditory feedback could activate a sensory neural population directly connected to motor neurons driving song production, leading to a strengthening of connections between sensory and motor neural populations through Hebbian learning [24], [25]. These connections could be the substrate of internal models [26], [27]. However, mirror neurons have only been reported in adult songbirds until now, and it remains unclear whether they are selectively responding to the tutor song following the sensory learning phase. Alternatively, song-related auditory responses may emerge only after the end of the sensorimotor learning phase, ruling out a role for these neurons in song acquisition [24].

In humans, different theories try to explain why there is activation of motor areas during speech perception [28]–[30]. For example, the Perception-for-Action-Control Theory (PACT) [31] highlights how speech percepts are related not only to sounds, but also to motor gestures: speech perception could be biased by articulatory invariant commands.

Syllables are perceptuo-motor by essence: i.e. perception shapes action (e.g. some abstract representation of motor gesture can be recovered to disambiguate perception) and, at the same time, action shapes perception (e.g. motor gestures are ”selected for their functional and perceptual value for communication” [31]). An example is the fact that acoustic features can change abruptly when changing the jaw height or jaw cycle, producing phase transition in the perceptuo-motor phase space diagram [31].

E. Learning rules and synaptic plasticity

Learning rules implemented in artificial neural networks are often inspired by biological synaptic plasticity. Evidence for synaptic plasticity in the songbirds song-related network has been highlighted recently [32]–[35]. The various sites of synaptic plasticity could underlie separate learning processes.

Plasticity in the thalamo-cortical synapse of the learning pathway may subserve early sensory learning [32]. Still, in the learning pathway, long-term potentiation (LTP) in Area X is modulated by dopamine [33]. LTP provides experimental evidence for a three-factor learning rule as those often used to model reinforcement learning processes in neural circuits [36]. Indeed, dopamine often mediates reinforcement signals [37], and several vocal learning models borrow concepts and algorithms from reinforcement learning (RL) theory [38]. In this framework, the progressive improvement of vocalisations observable in the behaviour reflects a trial-and-error strategy guided by the internal evaluation of the produced vocalisations (likely through its comparison with previously experienced adult vocalisations) as well as external rewarding cues directly provided by the adults [8].

Then, the long-term depression (LTD) of RA recurrent collateral synapses (i.e. between projection neurons) is limited to the song learning critical period; this could implement the pruning of unnecessary connections within RA [35]. The connections in the HVC-RA network are thought to be formed by a dense network that provides many paths for the descending motor signals, and only circuits that were active during singing need to be maintained. This is consistent with the high variability of juvenile's song or infant babbling [35], [39].

Finally, recent evidence for synaptic plasticity in the inputs to RA neurons from HVC and LMAN may provide a key element to model the interaction between the motor and learning pathways during learning [34]. Indeed, naturalistic stimulation patterns drive opposing changes in the strength of RAs inputs from HVC or LMAN. The extrapolated learning rule may allow the transfer of motor corrections initially driven by LMAN inputs and later consolidated in the motor pathway [40].

III. Aims of the models

The topics of the reviewed models are either speech perception and production development in humans, or song acquisition in birds. The second column of Table III contains the subject of the study of each paper that is reviewed: either humans (*"H"*) or songbirds (*"SB"*). In both cases, the focus is on early stages of learning, when babbling takes place (see Section II for more details about learning phases in humans and songbirds). Beyond the general topic, there are several objectives and questions that the authors have pursued. An overview of these objectives is shown in Table II: (i) investigate the effects of sensorimotor integration on the model definition, (ii) test the biological plausibility of hypotheses for the function of vocal learning brain areas, (iii) test a particular architecture and/or plasticity rule, (iv) include a realistic vocal tract, (v) test different types of exploration, and (vi) model social interactions.

A. Effects of sensorimotor integration

Some authors aim to study sensorimotor integration and its effect on sensory and motor space representations: Bailly [41] is interested in sensorimotor redundancy given the constraints imposed by the articulatory system; Westermann and Miranda [42] are concerned by the effect of auditory perception and production on the development.

B. Biological plausibility

Many authors are interested in modelling song-related pathways in birds, and in the study of auditory feedback. Troyer and Doupe [43] test several hypotheses about anterior vocal learning pathway including HVC-RA connections, efference copy and auditory feedback. Doya and Sejnowski [44] test the hypothesis that LMAN drives slow exploration in the connection from HVC to RA. Alternatively, Fiete et al. [45] hypothesise that LMAN produces transient song perturbations by driving rapid conductance fluctuations in RA neurons. In the context of speech production and perception, some authors developed models inspired by functions of brain areas [46]. Cohen et al. [47] tested the hypothesis that human brain areas are shared in language understanding and production, and their implication in goal-directed actions using active language learning and social babbling. Barnaud et al. [48] tested the hypothesis of idiosyncrasies (individual specificity) in production and perception; moreover, they tested the inter-individual variability in auditory and motor prototypes within a given language.

C. Learning architectures and algorithms

Some authors test the hypothesis that the anterior vocal learning pathway works as an actor-critic system and implement a gradient-based reinforcement learning rule. This is the case of Doya and Sejnowski [44] and Fiete et al. [45]. Reinforcement learning is implemented also by Howard and Messum [49] and Warlamount et al. [50]: they test the hypothesis that actions are reinforced based on auditory salience. Finally, Troyer and Doupe [43] combine Hebbian learning and a reinforcement learning signal in their architecture.

Alternatively, other authors learn internal models using different learning rules to update the synaptic weights matrix representing the connections between motor commands and goal representations. Howard and Huckvale [51] compare direct inverse mapping and distal supervised learning in the context of speech generated both by a real human subject and by a synthesizer; Philippsen et al. [52] aim to understand how to reduce the need for supervised training using only acoustic examples learning efficiently an inverse and a forward model. Oudeyer [5] and Pagliarini et. al [53] test a normalised Hebbian rule to learn the inverse model. Liu and Xu [54] test if it is possible to develop the acoustic-to-articulatory model by learning inverse kinematics in speech acquisition. More particular cases are presented by the works from Murakami et al. [55] who test imitation learning using a recurrent neural network, Kröger et al. [46] who test a self-organised network (SOM), and Barnaud et al. [48] who test a Bayesian model of speech communication.

TABLE II
AIMS OF THE MODELS: SUMMARY TABLE OF THE MAIN OBJECTIVES AND QUESTIONS PURSUED BY THE AUTHORS OF THE REVIEWED MODELS.

Sensorimotor integration	Brain area functions	Architecture/ plasticity rule	Realistic vocal tract	Exploration	Social interactions
Bailly 1997 [41] Westermann 2002 [42] Moulin-Frier 2015 [63]	Doya 2000 [100] Troyer 2000 [43] Fiete 2007 [45] Cohen 2018 [47] Barnaud 2019 [48]	Doya 2000 [100] Troyer 2000 [93] Fiete 2007 [45] Howard 2005 [51] Oudeyer 2005 [5] Howard 2007 [49] Kröger 2009 [46] Liu 2014 [54] Philippsen 2014 [52] Murakami 2015 [55] Warlaumont 2016 [50] Najnin 2017 [99] Pagliarini 2018 [53] Barnaud 2019 [94]	Doya 2000 [100] Howard 2005 [51] Moulin-Frier 2012 [56] Murakami 2015 [55] Philippsen 2016 [57] Teramoto 2017 [99] Howard 2019 [59]	Moulin-Frier 2012 [56] Moulin-Frier 2014 [10] Philippsen 2016 [56] Forestier 2017 [81] Acevedo-Valle 2018 [61]	Oudeyer 2005 [5] Lyon 2012 [62] Moulin-Frier 2015 [63] Acevedo-Valle 2018 [61]

D. A realistic vocal tract model

One of the objectives of many authors is to take into account anatomical and physiological constraints using a realistic model of the vocal tract for the production of the sound. Many authors want to include such a model in their study: Doya and Sejnowski [44], Howards and Huckvale [51], Moulin-Frier et al. [56], Murakami et al. [55], Philippsen et al. [57], Teramoto et al. [58]. In particular, Howard and Birkholz [59] test two different vocal tract models, with increasing complexity, both in the case of a real human teacher and in the case of an automatic synthesizer of sounds.

E. Exploration strategies

Several authors test whether or not mechanisms of intrinsically motivated exploration can self-organise early developmental stages of learning: Moulin-Frier et al [10], [56] compare different exploration strategies (random motor exploration, random goal selection and curiosity-driven active goal selection) to drive learning; Forestier and Oudeyer [60] focus on body babbling coupling self-generation of goals and imitation learning without any assumptions of capabilities for complex sequencing; Philippsen et al. [57] test goal-directed exploration of the target space and assume that there is no need of visual information. On the contrary Murakami et al. [55] starts and studies the relevance of visual information. Intrinsically motivated exploration is also in the interest of Acevedo-Valle et al. [61]: they formalise a socially reinforced and intrinsically motivated architecture for sensorimotor exploration to study the impact of social reinforcement on prelinguistic development.

F. Social and multi-agent interactions

Many authors are interested in the influence of social interactions during early pre-linguistic development: Acevedo-Valle et al. [61] study the influence of imitation maternal responsiveness; Lyon at al. [62] embed their learning system in a humanoid robot that interacts in real-time with naive participants. Moulin-Frier et al. [63] and Oudeyer [5] study self-organising properties of coupling perception and production within agents and between agents.

IV. MOTOR CONTROL

The first step in defining motor control is to choose an appropriate model mapping a motor space (i.e. muscle command) onto a sensory space (i.e. sound or acoustic representation). This section provides definitions of motor spaces and motor control functions that have been used in models.

A. Motor space

The motor space is used to describe motor articulations parameters (ideally as a function of time). These parameters control the dynamics of vocal tract muscles and glottis (for human control models). A high number of parameters is usually provided but often several can be kept constant, either because they do not have much influence on the sound produced or in order to reduce the number of parameters. The dimension of the motor space depends on the motor control function applied and also on the choices made by the modellers. There is a large variability in the number of dimensions of the motor space: from a low dimensional motor space, which only considers the parameters related to lip and tongue, to high dimensional motor spaces which include almost all the available parameters for the vocal tract and, in addition, the glottis parameters.

B. Motor control function

In humans, vocal motor control involves the respiratory system, the vocal organs (e.g. tongue, lips, jaw, larynx) and the vocal tract. Although some studies have been conducted in the context of vocalisations, neural mechanisms underlying the diversity of respiratory rhythms are largely unknown [64].

A basic model of speech production, therefore, includes a sound source (vocal folds) and a linear acoustic filter (vocal tract) [65]. The sound source is the combination of vocal folds vibration output and noise. Such noise can be due to pressure fluctuations or by activities of other parts of the apparatus (e.g. the glottis). Lumped-element models are a class of self-oscillating biomechanical vocal folds models: these low-dimensional vocal fold models couple airflow and biomechanics [66]. Such low-dimensional models can reproduce characteristics of real vocal fold oscillations [67] and have been largely applied in speech research [68]: the parameters

TABLE III
SUMMARY TABLE OF THE **SPACES AND FUNCTIONS** OF SENSORIMOTOR MODELS.

	Subject	Motor space		Motor control	Sensory space (SP?)	Pre-processing of the sound	Sensory response	Perceptual space/Internal representation	
		Dim						Dim	
Bailly 1997 [41]	H	8	Lip, larynx, jaw, tongue and apex	DIVA [79,80]	S	4 Formants (in Hz)	Polynomial interpolation + CDA	2	Discriminant space
Doya 2000 [100]	SB	4	Fundamental frequency and peak frequency of sound, sharpness of band-pass filter, gain of the amplifier	Source-filter model	S	--	--	--	Syllable space (localist encoding)
Troyer 2000 [43]	SB	40	Coordinates (arb.)	--	--	--	--	40	Syllable space (localist encoding)
Westermann 2002 [42]	H	29	Interarytenoid, cricothyroid, styloglossus, levator palatini, genioglossus, hyloglossus, mylohyoid, , orbicularis oris, masseter	2D ODE model (Pipes' walls + Air pressure)	S	2 Formants (in Hz)	All-zero filter + Autoregression + Gaussian selectivity	2	Formants (in Hz)
Howard 2005 [51]	H	9	Jaw, tongue, lip, voicing, fundamental frequency and larynx height	VLAM [78]	S	Spectrogram via JSRU vocoder	Autocorrelation	21 * 30 (t)	Autocorrelation estimate for F0 and voicing
Oudeyer 2005 [5]	H	3	Lip rounding, tongue height, tongue position.	de Boer model [70]	--	4 Formants (in Barks)	Linear combination + Gaussian selectivity	2 * ? (t)	Acoustic trajectory in a 2D subspace of the formants
Fiete 2007 [45]	SB	12	Pitch period and height + filter linear predictive coeff.	Source-filter model	S	--	--	720 neurons	Neural activity
Howard 2007 [49]	H	4 to 9	Jaw, tongue, lip, voicing, fundamental frequency and larynx height	VLAM [78]	S	Low pass filtered spectrogram	Differenced narrow-band spectrogram	2	Low frequency power + spectral change
Kröger 2009 [46]	H	2	Back-front, low-high	Motor plan state	S	3 Formants (in Barks)	Rescale to [0,1]	3	Formants (in Barks)
Howard 2011 [83]	H	10	Jaw, tongue, lip, voicing, fundamental frequency, larynx height and nose	VTCalcs [82]	S	--	--	--	Vector of continuous values
Lyon 2012 [62]	H	--	--	eSpeak [84]	S	Phonemes (CMU alphabet)	SAPI 5.4 [98]	4 sec.	Phoneme stream
Moulin-Frier 2012 [56]	H	7	Jaw, tongue, lip, separation and larynx height	VLAM [78]	S	3 Formants (in Hz)	Linear combination	2	Subspace of the formants
Moulin-Frier 2014 [10]	H	7	7 param. from the PCA on the vocal tract shape	DIVA [79,80]	S	2 Formants (in Hz)	Rescale to [-1,1]	3 * 2(t)	2 Formants (in Hz) + intensity
Moulin-Frier 2015 [63]	H	3	Lip, tongue body and dorsum	VLAM [78]	--	3 Formants (in Barks)	--	3	Formants (in Barks)
Liu 2014 [54]	H	3	Target slope and height, rate of target approximation	quantitative Target Approximation	--	F0 continuous traj. (2-dim)	Sampling	5	Syllable space (Time-normalized F0 samples)
Philippsen 2014 [52]	H	26	22 vocal tract arb. param. + glottis param.	VocalTractLab [74]	S	--	Logaritmic energy + 12 MFCC features	39 * ? (t)	Acoustic trajectory
Murakami 2015 [55]	H	20	Tongue, lip, hyoid, jaw, velic, velum shape	VocalTractLab [74]	S	Dual Resonance Non-Linear filter model	Reservoir (1000 units)	4 (+1 empty)	Phoneme classes
Warlaumont 2016 [50]	H	2	Jaw and lip trajectory	Praat [73]	S	--	--	--	--
Philippsen 2016 [57]	H	24	20 vocal tract arb. param. + glottis param.	VocalTractLab [53]	S	3 Formants + 13 MFCC features	PCA + LDA (10 to 2 dim.)	2	Goal vowel embedding
Forestier 2017 [81]	H	7	7 param. from the PCA on the vocal tract shape	DIVA [79,80]	S	--	DIVA [65]	2 * 5(t)	Acoustic trajectory in the 2D space of the first two formants
Najnin 2017 [99]	H	11	11 param. for vocal tract and 2 param. for phonation	DIVA [79,80]	S	--	DIVA [79,80]	4 / 12	Acoustic trajectory in the 2D space of the first three formants and phonation/normalized MFCCs
Teramoto 2017 [58]	M	3	Air pression, vocal fold tension, time constant	Source-filter model	S	--	--	--	--
Acevedo-Valle 2018 [61]	H	13*2	10 position of the articulators + 3 phonation parameters	DIVA [79,80]	S	2 Formants (in Hz)	Average of trajectories	3 * 2(t)	2 Formants (in Hz) + intonation
Cohen 2018 [47]	H	3	Arb.	--	--	IMS	--	--	Specific need (ex: thirst, hunger)
Pagliarini 2018 [53]	SB	3	Arb.	--	--	IMS	Gaussian selectivity	3	Syllable space (localist encoding)
Howard 2019 [59]	H	7	Palate, larynx, pharynx, jaw, lips, teeth, togue	VocalTractLab [74]	S	--	--	--	Vector of continuous values
Barnaud 2019 [48]	H	3	Lip, tongue body and dorsum	VLAM [78]	--	2 Formants (in Barks)	--	2	Formants (in Barks)

--: Not Available; Arb.: arbitrary; Dim: dimension; DIVA: directions into velocities of articulators; IMS: identical to motor space; H: human; JSRU: joint speech research unit; LDA: linear discriminant analysis; M: marmoset; MFCC: mel frequency cepstral coefficients; PCA: principal component analysis; S: sound; SB: songbird; SP?: Sound Production?; VLAM: vocal linear articulatory model

and the structure of lumped-element models can be tuned to sustained vowel simulations to obtain different frequencies. Additionally, it can be tuned to simulate various vocal registers, e.g. to generate a sequence of sounds (to simulate running speech), or to study some pathological phonation conditions (e.g. incomplete glottal closure).

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 resonance frequencies of the vocal tract are called formants [69]. The human vocal tract has been often modelled as a structure of pipes: in the literature, Ordinary Differential Equations (ODEs) models describe air pressure dynamics in the vocal tract. For example, Westermann and Miranda [42] used a synthesizer which models the vocal system as a structure of pipes, each one having four walls represented as mass spring damper models (useful to model non-linearities): the 2D model equations describe the pipe wall physical behaviour, evolution of the movements and air pressure. Similarly, De Boer model [70], [71] has been used by Oudeyer [5]: the synthesizer is based on the interpolation between the formant frequencies of vowels generated by Maeda's articulatory model [72].

Articulatory synthesizers are based on the same idea and control the vocal articulators: (i) *Praat*, a software for speech analysis containing an articulatory synthesizer, developed by Boersma in [73], and used by Westermann and Miranda [42], and by Warlaumont and Finnegan [50]; (ii) *VocalTractLab* (VTL), developed by Birkholz [74], [75], and used by Philippsen et al. [52], [57], Murakami et al. [55], Howard and Birkholz [59], as well as in speech signal filtering [76] or articulatory synthesizer training [77]; (iii) *Vocal Linear Articulatory Model* (VLAM) [78]) has been used by Howard and Huckvale [51], Moulin-Frier et al. [56], [63], Howard and Messum [49]; (iv) *Directions Into Velocities of Articulators* (DIVA) [79], [80] has been used by Bailly [41], Moulin-Frier et al. [10], Forestier and Oudeyer [81] and Acevedo-Valle et al. [61]; (v) *VTCalcs* software proposed by Maeda [82] has been used by Howard and Messum [83].

Taking inspiration from previously developed vocal tract models, Kröger et al. [46] proposed to define two parameters (*back-front* and *low-high*) describing the state of the motor plan and covering the whole articulatory vowel space. Other motor parameters like tongue position and lip parameters are expressed in function of these two motor plan parameters. Two particular cases are given by Lyon et al. [62] which used eSpeak, a synthesizer that uses a formant synthesizer method [84] and Liu and Xu [54], where qTA (quantitative Target Approximation) has been used to mimic the motor control dynamics, controlling them via three parameters related to the target properties.

For modelling song production in birds, an interactive model where nonlinear interaction between timescales enables motor instructions has been proposed [85]–[91]. More recently, Alonso et al. [92] developed a simple time continuous additive neural network model that drives the dynamics of respiratory activity: respiratory patterns can be reproduced and predictions on the timing of HVc activity during the production can be performed. Anatomical properties and small size of birds make the investigation of vocal fold mechanisms difficult. It has been shown that the brain seems unable to control each motor parameter independently but it uses a complex gesture-dependent control scheme to drive the vocal output [86], [93]. Different studies have been looking at the properties of vocal motor control in correlation with acoustic features, such as 3D imaging techniques to investigate the control of sound pitch [94] or neural recordings analysis to investigate the variations in the song [95].

Amador et al. [96], Doya and Sejnowski [44], Fiete et al. [45] model the vocal tract dynamics in birds using ODEs. They include time-dependent constants related to air pressure and syringeal labial tension. The output is the pressure needed to generate the sound. Such a dynamical system is able to synthesise realistic vocalisation sounds if a series of instruction derived from a recorded song input is given [97]. The model from Amador et al. [96] has been used by Teramoto et al. with marmoset [58] in a vocal development study.

C. Sound production

More realistic models generate sound production through the motor control device: Bailly [41], Doya and Sejnowski [44], Westermann ad Miranda [42], Howard and Huckvale [51], Fiete et al. [45], Howard and Messum [49], [83], Howard and Birkholz [59], Lyon et al. [62], Moulin-Frier et al. [10], [56], Forestier and Oudeyer [60], Philippsen et al. [52], [57], Murakami et al. [55], Acevedo-Valle et al. [61], Warlamaunt et al. [50]. Some models rather rely on an abstract representation of the vocal output including a discrete set of features (e.g. formants) as in the works from Troyer and Doupe [43], Oudeyer [5], Moulin-Frier et al. [63] and Barnaud et al. [48].

V. Sensory system

The sensory system processes sensory stimuli and leads to a perceptual representation of those stimuli (in the perceptual space). While sensory stimuli may arise from other subjects (e.g. adult vocalisations to be memorised during the sensory learning phase), the production of vocalisations by the motor control apparatus also leads to the stimulation of the sensory system. As mentioned in Section II-C, it provides a feedback of the motor command that allows to compare the perceived vocal production with previously experienced adult vocalisations (e.g. the memorised tutor song in the case of birds). The evoked sensory responses may also be conveyed to the reinforcement system, where an evaluation of the produced sounds leads to a reward signal.The sensory response function is often modelled as a minimal extraction of a low dimensional feature-based description of sounds in vocal learning models.

This section motivates the choice of the sensory response function, the sensory space and the perceptual space. In Table III we separate the physical space of the sound, which is the sensory space, and its abstract representation, which is composed of a pre-processed sound and the perceptual space. This allows to highlight whether a model has sound production or not, and to compare them in both cases.

A. Sensory space

The sensory space is the output space of the motor control device. As mentioned in Section IV-C not all the models include sound production. The most simplistic models do not define the motor control device, leading to a coincidence between motor and sensory space. For instance, this approach has been used by Cohen et al. [47] and Pagliarini et al. [53] (i.e. Identical to Motor Space (IMS) in Table III).

B. Sensory response function

The sensory response function acts on the sensory space and drives the activity in the perceptual space, where the auditory stimuli are represented with a lower dimension. This process is the result of one or more steps that lead to a filtered, normalised and/or reduced subspace representing the auditory stimulus. The output is an abstract representation of the sound which represents its encoding in the brain. To highlight the fact that the auditory process is in general not a single-step process, a column (*Pre-processing of the sound*) represents an intermediate step between the real sound and the perceptual space. Most models describe first the sound as a trajectory in the formants' space, varying the dimension of the space (usually from 2 to 4) and the measure unit (either Barks or Hertz). Alternatively, other common examples of preprocessed sound are given by a low pass filtered version of the spectrogram, or the trajectory of the fundamental frequency.

A filter on the spectrogram or on the formant space has been applied by Westermann and Miranda [42]. A linear combination has been used in the works from Oudeyer [5] and Moulin-Frier et al. [56]. Furthermore, Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) have been applied by Philippsen et al. [57]. An average over sound trajectories has been used by Acevedo-Valle et al. [61]. Alternatively, some authors extracted different features from the sound to build its representation in the perceptual space, or its internal representation. This is the case for Howard and Messum [49], [83], Philippsen et al. [52], [57], Liu and Xu [54]. Howard and Huckvale [51] estimate the autocorrelation of the fundamental frequency of the sound and of the voicing parameter (from the motor control).

Nonlinearity in the sensory response function can be introduced defining the auditory activity as a bell-shaped function around the target motor pattern. For instance, this choice has been made in the works of Westermann and Miranda [42], Oudeyer [5] and Pagliarini et al. [53]. A few particular cases are given by the work from Lyon et al. [62] where a specific software, called *SAPI 5.4* [98], has been used to encode the stimulus and by the works of Murakami et al. [55] where a phoneme representation of the stimulus is obtained from a Random Recurrent Neural Network (RNN), called a *reservoir*.

C. Perceptual space/Internal representation

The output of the sensory response function is a lower dimensional representation of the sound produced by the vocal apparatus. In the context of humans the sound have been represented in the space of the first 2, 3 or 4 formants (in Hertz or Bark scale) by Westermann and Miranda [42], Oudeyer [5], Kröger et al. [46], Moulin-Frier et al. [10], [56], [63], Forestier and Oudeyer [60], Najnin and Banerjee [99], Philippsen et al. [57] and Barnaud et al. [48], Acevedo-Valle et al. [61]. The latter also consider the intonation as third acoustic parameter. Alternatively, Pagliarini et al. [53] propose a localist encoding for the syllables.

The percepts can be given by the spectral properties of the sound: for instance the frequency powers, the power change, the fundamental frequency, the Mel-Frequency Cepstral Coefficients (MFCC), or also pitch and amplitude. This choice has been made by Howard and Messum [49], [83], Najnin and Banerjee [99], Philippsen et al. [52], [57], Liu and Xu [54] and Fiete et al. [45]. Alternatively, the percepts can be the classes of phonemes, as in the works by Lyon et al. [62] and Murakami et al. [55].

For any species, it is likely that the representation of a given sound in the perceptual space keeps changing during development, thus making the learning of inverse model even more difficult until the moment when the perceptual space "converged". That is why the vast majority of the models have a sensory response function that does not change during learning and is kept fixed. This can be justified based on the assumption that learning the inverse model only starts at the end of the "universal sensory period", which is the case for some species of birds like sparrows.

Some studies do not have a sensory response function that leads to a perceptual space. These are models with a non-perceptual internal representation of the goals, such as the general model shown in Figure 2. As for the perceptual space, the choices made by the author can be found in the last column of Table III. This is the case for the reinforcement learning models proposed by Doya and Sejnowski [100], Troyer and Doupe [43], Fiete et al. [45], Warlamount et al. [50] Cohen et al. [47] and Howard and Birkholz [59]. In the context of songbirds, a typical choice is to use an arbitrary syllable space given by a localist encoding, as in the works from Doya and Sejnowski [44] and Troyer et al. [43]. Alternatively, Fiete et al. [45] use the neural activity of a spiking neural network. Finally, in the work from Cohen et al. [47] goals are specific needs of the agent (e.g. thirsty, hunger).

VI. Learning framework

This section introduces the different types of architectures, the learning domains (i.e. perceptual, motor or goal spaces), the learning rules or optimisation algorithms, the exploration strategies that could drive learning, and finally the evaluation measures. Table IV summarises how the reviewed models implement the learning framework.

A. Architecture

The architecture linking the learning domain to the learning image varies between models and different architectures can be used. Biological hypotheses made in a particular model are important to understand the choice of the architecture and the learning rule. For more details about the biological hypothesis refer to II-D and II-E).

Inverse and forward models (i.e. internal models) are both predictor models: they provide a bi-directional link between the perceptual space and the motor space, when both a forward and an inverse model are included. Inverse models have the aim to provide an appropriate motor command for a given perceptual goal, which is driven by the sensory response; the learning domain is defined by the perceptual space. Forward models describe a causal relationship between motor commands and their corresponding perceptual representations; the

TABLE IV
SUMMARY TABLE OF THE **LEARNING FRAMEWORKS**.

	Subject	Architecture				Exploration			Learning	Evaluation	
		Internal model (if present)	RNN	FF NN	Other	Goal-directed	Random	Dimension		Measure	Space
Bailly 1997 [41]	H	I + F		1-layer perceptron			X	Motor	Gradient inversion + deviation measure	Deviation	Real forward measure and its estimation by the interpolator
Doya 2000 [100]	SB			2-layer perceptron with 4 subnetworks			Dynamic perturbation	Motor	Reinforcement learning via stochastic gradient ascent	Correlation	Gaussian filter + normalization of the spectrogram
Troyer 2000 [43]	SB			2-layer perceptron			X	Motor	Reinforcement signal + Hebbian rule	Correlation coefficient	Matrices of co-fluctuations in activity over syllable epochs
Westermann 2002 [42]	H	I + F		1-layer perceptron			X	Motor	Hebbian Covariance rule	--	--
Howard 2005 [51]	H	I + F		2-layer perceptron			X	Motor	Back-propagation + gradient descent	Similarity	Spectrogram of the sound
Oudeyer 2005 [5]	H	F		1-layer perceptron			Uniform	Motor	Hebbian Correlation rule	--	--
Fiete 2007 [45]	SB			2-layer perceptron			Dynamic perturbation	Motor	Reinforcement learning via stochastic gradient ascent	MSE	Delayed estimate of performance (pitch and amplitude)
Howard 2007 [49]	H		--	--	--		X	Motor	Reinforcement learning via gradient descent	Auditory salience + effort (voicing degree in VLAM)	Spectral properties of the sound, motor properties
Kröger 2009 [46]	H	I + F			SOM		BMU	Motor	Hebbian normalized rule	Distance	Motor pattern estimation
Howard 2011 [83]	H				Optimization	X		Sensory, Motor	Quasi-Newton gradient ascent	Auditory salience, diversity and effort	Spectral properties of the sound, motor properties
Lyon 2012 [62]	H		--	--	--	Syllable probability		Perceptual	--	F-measure	Perceptual
Moulin-Frier 2012 [56]	H	I + F			Optimization	Competence progress		Goal	Reaching algorithm	Competence progress	Goal
Moulin-Frier 2014 [10]	H	I			Bayesian	Competence progress		Goal	GMM over motor variables	Distance	Perceptual
Moulin-Frier 2015 [63]	H	I + F			Bayesian		X	Motor	COSMO [106]	Dispersion Theory formula	Perceptual
Liu 2014 [54]	H	I		2-layer perceptron		X		Goal	Online learning, Backpropagation.	SSE	Perceptual
Philippsen 2014 [52]	H	I + F	ESN (firing rate reservoir)				Uniform	Goal	Linear regression	MSE	Perceptual
Murakami 2015 [55]	H				Optimization	Confidence levels		Goal	CMA-ES	Confidence levels	Goal
Warlaumont 2016 [50]	H		Spiking Reservoir				X	Motor	Reinforcement learning via reward-modulated STDP	Auditory salience	Perceptual
Philippsen 2016 [57]	H	I + F		2-layer RBF		Dynamic perturbation + GMM		Goal	Distance in goal space + auditory salience	Competence	Goal
Forestier 2017 [81]	H				Optimization	Random + Exploration noise		Goal	Reaching algorithm	Distance	Perceptual
Najnin 2017[99]	H		3-layered RNN		Optimization		X	Goal	Autoencoder and actor-critic network	Distance	Perceptual
Acevedo-Valle 2018 [61]	H		--	--	--	Interest model		Goal	ilGMM	Distance	Perceptual
Cohen 2018 [47]	H			1-layer perceptron		Caregiver choice		Goal	Maximization of the reward	Moving average of the reward	--
Pagliarini [53]	SB	I		1-layer percepron			Uniform	Motor	Hebbian normalized rule	Distance	Perceptual
Howard 2019 [59]	H		--	--	--		X	Motor	Reinforcement learning via gradient descent	Auditory salience, diversity	Spectral properties of the sound, motor properties
Barnaud 2019 [48]	H				Bayesian		X	Motor	COSMO [106]	Dispersion Theory formula	Perceptual

--: Not Available; BMU: Best Matchin Unit; CMA-ES: covariance matrix adaptation - evolution strategy; COSMO: communicating objects through sensorimotor operations; ESN: Echo State Network; F: forward model; FF NN:feedforward neural network; GMM: gaussian mixture models; H: human; I:inverse model; ilGMM: incremental learning GMM; Int: Internal representation; MSE: mean square error; O: optimization algorithm; RL: reinforcement learning; RBF: radial basis function; RNN: recurrent neural network; S: supervised; SB: songbird; SSE: sum of squared error; SOM: self-organizing maps; STDP:spike timing dependent plasticity; U: unsupervised; X: distribution not specified

learning domain is defined by the motor space. As mentioned in Section I, sensorimotor integration leads to redundancy: this is a fundamental problem with inverse models since introducing such kind of model leads to non-convex problems [101]. This problem can be approached using the combination of an inverse and a forward model [6], [102]. Indeed, forward modelling can be used to facilitate the estimation of the current state enabling the learner to modify its action and match the prediction: this switches action and perception representations, and explain the effects of perception on action [103]. Else, the non-convexity problem can be solved using a combination of an inverse model and a feedback controller [104] or "goal-babbling" to drive learning [105].

Alternatively, as shown in Figure 2 other models define a non-perceptual internal representation of goals and learn not the connections between the perceptual and the motor space, but the link between the internal representation and the motor space. These models include a sensory space if there is a motor control producing a real sound as output, and a sensory response function, which is used to process the sound and build a reward or an evaluation of the learning.

The structure of the network varies among the models. Some approaches involve feedforward neural networks (FF NN) in the learning architecture. For instance, a 1-layer perceptron has been used in the works by Pagliarini et al. [53], Westermann and Miranda [42], Oudeyer [5] and Cohen et al. [47]. A multi-layer perceptron has been used in the works by Doya and Sejnowski [100], Troyer and Doupe [43], Howard and Huckvale [51] and Liu and Xu [54]. A Radial Basis Function (RBF) network has been used by Philippsen et al. [57]. Alternatively, a reservoir has been used by some authors: Warlaumont et al. [50] uses a reservoir as a kind of biological implementation of reinforcement learning for high-level control of sequential motor production; Philippsen et al. [52] use two reservoirs to learn both the forward and the inverse models that link motor space with perceptual space; Najnin and Banerjee [99] use a 3-layered RNN to define the predictive model that uses a generative network to predict the proprioceptive sensory (representing the perceptual dimension) from a causal state (representing the motor dimension). A Bayesian architecture has been proposed in the works from Moulin-Frier et al. [10], [63] and Barnaud et al. [48]. Finally, three Self-Organizing Maps (SOM) have been used in the work by Kröger et al. [46].

B. *Learning domain*

The learning domain, in the case of inverse models, coincides with the perceptual space, which contains the representation of the stimuli (how the brain encodes sensory stimuli) and is obtained through the sensory response. In the case of forward models the perceptual space coincides with the output, and the learning domain coincides with the motor space. Alternatively, the learning domain is defined as the internal representation of goals and the sensory response drives a reward that modulates the learning rule. In the latter case, the learning domain may be an abstract representation of the goal, that could represent for instance a sound trajectory. Of course, the internal representation could be considered as a component and not the domain of the learning framework, but we prefer to consider it as the domain of the learning mechanism (or architecture), in order to know what is needed for the learning or optimisation to be available. The full learning domain, or a sub-part, could be also called goal space in models using goal-driven exploration.

The learning domain might encode a whole song (that is a sequence of syllables) or a single syllable, depending on the choices and aims of the model. For instance, the learning domain can be defined as a syllable space that might encode features, as in the works from Troyer and Doupe [43], Liu and Xu [54]. Or again, using localist encoding (i.e. one-hot encoding)[3] as in the work from Pagliarini et al. [53].

Some authors use sound features to describe the perception of a stimulus, for example intensity in the work from Moulin-Frier et al. [10], fundamental frequency in the work from Howard and Huckvale [51], Frequency power and spectral change in the works from Howard and Messum [49], [83] or pitch and amplitude in the work from Fiete et al. [45]. Alternatively, the learning domain has been defined as a subspace of the formants in the works from Oudeyer [5], Moulin-Frier et al. [10], [63], Kröger et al. [46], Acevedo-Valle et al. [61], Forestier et al. [81] and Barnaud et al [48]. Philippsen et al. [57] define the learning domain as the first two dimensions of the embedding space. Here the stimulus has been modelled using the sound trajectory lying in the correspondent space.

Finally, the learning domain can be identified with the output of specific neural networks architecture or particular software. For instance, in the work from Lyon et al. [62] the goal is represented by a phoneme stream computed using the software SAPI 5.4 [98], and the work from Murakami et al. [55] where an intrinsic learning is defined to build the goal space. In the latter, an Echo State Network (ESN, a specific Recurrent Neural Network (RNN), called reservoir) has been used to learn in advance the goals, and an auditory memory encodes the knowledge of each goal (called auditory memory function).

C. *Learning rule*

Different types of learning have been used to model sensorimotor learning: supervised and unsupervised learning, and reinforcement learning. In a few models, an optimization algorithm (instead of a learning rule) has been used to improve motor production.

1) Unsupervised learning: Biologically, as seen in Section II-E, specificity, cooperativity and associativity are expressed in the neural activity. Computationally, this can be modelled using associative learning rules, which are usually used for building internal models (inverse or forward) and are unsupervised. Hebbian-inspired learning algorithms typically implement associative learning and shape the excitatory links between perceptual and motor representations [2].

A theoretical inverse model has been proposed by Hahnloser and Ganguli [25], where an Hebbian-inspired learning rule drives learning. Hebbian-inspired learning rules have been used in the works from Troyer and Doupe [43], Kröger et al. [46] and Pagliarini et al. [53]. Also, a Hebbian correlation rule involving the mean activation of neurons over a certain time interval [106] has been used in the works from Westermann and Miranda [42] and Oudeyer [5]. Otherwise, to define

[3] Localist and one-hot encoding is probably the simplest orthogonal representation one can have. It consists of a binary encoding where an input is represented by one feature at 1 and all the other features at 0: e.g. 4-dimensional vectors [0 1 0 0] and [0 0 0 1] could represent two different inputs with localist encoding.

a learning rule one can use the distance between the target and the actual production in the goal space as in the work from Philippsen et al. [57].

2) Reinforcement learning: Reinforcement learning (RL) is a mechanism to learn an action policy to maximize the expected reward, where the reward function encodes the goal. The goal space (internal representation of goals in Figure 2) defines the learning domain. The definition of the learning domain (given in Table III in column "*Perceptual space/Internal representation*") and of the reward function (given in Table IV in column"*Evaluation*") are important to determine the complexity of the learning and the biological plausibility of the model.

Among the reviewed models there are models which implement classical RL: Troyer and Doupe [43] used a plasticity rule which combines an associative learning rule and a reinforcement signal. Doya and Sejnowski [100], Fiete et al. [45] and Howard and Messum [83] implemented reinforcement learning using a gradient ascent or descent algorithm. Similarly, gradient descent has been used in the work from Howard and Birkholz [59]. In these studies, the reward was computed as the correlation between spectrograms by Doya and Sejnowski [44], or based on the feature of the song (the delayed estimation of the sum of the squares of pitch and amplitude) in the work from Fiete et al. [45]. In these models, the reward function, which is treated in Subsection VI-E, encodes the goal and contributes to the learning. Alternatively, the reward can be driven by the auditory salience as in the works from Warlaumont et al. [50], Howard and Messum [49] and Philippsen et al. [57], or by the caregiver choice which defines any novel situation that the agent must learn [47].

Other authors did not choose to maximise a classical reward function but other quantities encoding the goal (e.g. a competence function, auditory salience). In this sense, RL has been implemented introducing intrinsically motivated exploration and active-goal selection. A Competence Progress algorithm which updates the internal representation of the goal and drives the exploration has been used by Moulin-Frier et al. [10], [56]: in the particular case of Moulin-Frier et al.[10] the learning algorithm is based on Gaussian Mixture Models(GMM) updated via Bayesian inference in a self-supervised paradigm. Intrinsic motivation has been used by Forestier and Oudeyer [81] and different types of goal selection have been proposed by Moulin-Frier [10], [56]. See Subsection VI-D for details.

3) Optimisation algorithm: Learning can also be driven by an optimisation algorithm that tunes the motor parameters: this is an exception and hence we did not use a more general category *Parameter tuning* instead of *Learning* in Table in Fig. IV. The optimisation procedure can aim to maximise the ability of the agent in reproducing a selected goal via a reaching algorithm [56], [60], or to maximise the reward [47]. Alternatively, a gradient inversion has been proposed by Bailly [41], and a quasi-Newton gradient descent algorithm has been proposed by Howard and Messum [83] to maximise a reward given by the combination of auditory salience, a diversity measure in the sensory space and en effort measure in the motor space. A particular example of optimisation algorithm is the Covariance Matrix Adaptation - Evolution Strategy (CMA-ES) [55], that is a searching algorithm to maximise the confidence level of each goal. Finally, Najnin and Banerjee [99] propose an actor-critic network to obtain the optimal sequence of actions to reach the target.

4) Supervised learning: Some works use supervised learning to learn the sensorimotor map. This could be implemented using an online algorithm via backpropagation as proposed by Liu and Xu [54]. Otherwise, this could be implemented combining backpropagation and gradient descent as proposed by Howard and Huckvale [51]. Supervised and unsupervised learning can also be used in combination with forward and inverse models, as in the work from Philippsen et al [52]. They move from supervised self-training (thanks to the availability of a forward model) to unsupervised learning when imitation of novel contexts is included (after the training).

5) Other types of learning: Alternatively, incremental learning Gaussian Mixture Models (ilGMM) has been proposed by Acevedo-Valle et al. [61] or GMM updated using Bayesian inference has been proposed by Moulin-Frier et al. [10]. A probability-based model has been proposed by Barnaud et al. with COSMO (Communicating Objects through SensoriMotor Operations) [48] architecture. This architecture was proposed by Moulin-Frier et al. [63] and represents a Bayesian framework to approach vocal learning.

D. Exploration strategies

Different exploration strategies have been studied in the context of vocal learning or in other types of sensorimotor learning. Exploration can take place either in the motor space or in the goal space (perceptual space or internal representation). The simplest exploration mechanism is driven by uniform random exploration. Pure random exploration does not take into account (1) the memory of perceived stimuli (e.g. the distribution of percept vectors in the perceptual space), (2) the history of what has already been explored in the past. Several works use this approach to explore the motor space: Troyer et al. [43], Westermann and Miranda [42], Howard and Huckvale [51], Howard and Messum [49], Howard and Birkholz [59], Oudeyer [5], Moulin-Frier et al. [63], Warlaumont et al. [50], Pagliarini et al. [53] and Barnaud et al. [48]. Alternatively, dynamic perturbation around a motor configuration has been used in the works from Doja and Sejnowski [44] and Fiete et al. [45] while implementing RL. A few authors used random exploration in the goal space: Forestier and Oudeyer [60], Najnin and Banerjee [99], Moulin-Frier et al. [56] and Philippsen et al. [52].

More sophisticated strategies are inspired by the nature of human development, which is progressive, incremental, autonomous and active. Behavioural analysis evidences how the actions of an agent are motivated by an internal or external reward. Following this idea, intrinsic motivation makes the agent choose an action basing the decision on the level of novelty, on the challenge it represents and on an internal reward. An example of such a strategy is called Intelligent Adaptive Curiosity (IAC) [7]: using a similarity-based progress maximisation the exploration is driven by the aim of maximising the learning progress, while the agent goes towards novel situations. Intrinsic motivation can drive motor babbling, defining a goal-directed exploration strategy. Usually, a competence function drives the choice of the next goal estimating the error or the reward or the level of knowledge relative to the goal. Different goal-directed strategies have been proposed in kinematic motor control learning by Forestier et al. [81], [107], Baranes et al. [108] and Rolf et al. [105].

Studies in the speech domain take inspiration from kinematic studies and introduce goal babbling as exploration strategy. This strategy allows the agent to do intermediate productions in the direction of the selected goal: that is, for any chosen goal the agent can define and make use of intermediate sub-goals to adapt the production. Goal babbling has been used by Liu and Xu [54] and proposed in unsupervised learning driven by a measure of confidence to reproduce a sound as in the works from Philippsen et al. [57] and Murakami et al. [55], a competence progress as in the works from Moulin-Frier et al. [10], [56], an interest model as in the work from Acevedo-Valle et al. [61] or the intrinsic reward as in the works from Forestier et al. [60].

E. Evaluation

Evaluation of learning (or reward computation) can take place in the perceptual space, in the internal representation or in an additional space defined *ad hoc*. In models using the reinforcement learning (RL) paradigm it is common to have such *ad hoc* definitions: in such a case the evaluation is called *reward* and is computed by a *critic*. For example, the reward can be given by the correlation between the target and the output songs represented as a filtered, vectorized version of the sound spectrogram as in the work from Doya and Sejnowski [100], or by the sum of the squares of pitch and amplitude as in the work from Fiete et al. [45]. In the work of Troyer et al. [43], the quality of learning can be computed using the correlation coefficients between matrices representing the co-fluctuation of activity at different syllable epochs.

In the case of intrinsically motivated agents, evaluation guides exploration, even if it does not contribute directly to the learning algorithm. These examples are related to evaluation computed in the goal space (i.e. perceptual or internal space). It can be computed using *competence progress* as proposed by Moulin-Frier et al. [56] and Philippsen et al. [57], or defining the confidence level of each goal as proposed by Murakami et al. [55]. Alternatively, other distance measures can be used to evaluate the learning in the perceptual space. For instance, Mean Square Error (MSE) has been used by Philippsen et al. [52], an intensity measure has been used by Moulin-Frier et al. [10], the distance between sound trajectories in the formant space is used by Forestier and Oudeyer [60]. Sum of Squared Error (SSE) has been used by Liu, Xu [54] and Euclidean distance has been used by Acevedo-Valle et al. [61] and Pagliarini et al. [53]. A particular example of evaluation performed in the perceptual space is the work from Lyon et al.[62] where a measure (called *F-measure*) is used to check the performance in learning the phonemes' dictionary.

Although evaluation is not usually implemented in the motor space, it is possible that some motor properties are used (e.g. articulator speed to compute the cost of a movement) to compute the reward. Kröger et al. [46] compute the error value estimating the distance between the initial motor pattern and the estimated one. Interestingly, the works from Howard and Messum [49] [83] and Howard and Birkholz [59] contain an example of a reward computed combining motor properties (voicing, effort, diversity) and auditory salience (computed using the spectral properties of the sound such as acoustic power, high to low frequency ratio and vice-versa). Auditory salience has been used also in the work from Warlaumont et al. [50].

Two particular cases are given by the work from Howard and Huckvale [51], where a spectrographic analysis is used to determine similarity between target and produced sound, and the work from Moulin-Frier et al. [63] and Barnaud et al. [48], where simulations are evaluated using the Dispersion Theory formula [109] in the COSMO architecture [63].

VII. Discussion

To provide an accurate representation of the vocal learning process in humans or songbirds, a model should implement the biological mechanisms revealed by past experimental investigations at the behavioural, anatomical and physiological level (see the biological context introduced in Section II). The various models presented here are about song learning in songbirds and speech development in humans, but have been built to answer different questions (as highlighted in Section III). However, we believe that comparing the various frameworks used to model different aspects of vocal learning will help to identify the important components and the links between them. Ultimately, such comparison may also reveal the next steps required to build a common model schema to study various questions about vocal learning and to account for a large number of experimental findings.

In Section I we introduced two kinds of sensorimotor learning models (see Figure 1 and Figure 2), the different spaces characterising a vocal learning model, and the functions going from one space to another. Table III and Table IV highlight all the components we discussed in the review for all the considered models. However, it is not always possible to clearly identify each model component as they are missing in some models (indicated by "–" in the tables).

In general, motor control in vocal learning models is often based on pre-existing biologically-inspired models of vocal production and include the production of sound. The motor parameters are usually related either to sound properties (e.g. fundamental frequency, pitch period) or to anatomical parts of the body (e.g tongue and lips in humans, air pressure in birds). Models of sound production (e.g. VTL, DIVA) may not be able to reproduce perfectly the distribution of sounds that could be obtained from real data. Therefore, they may not have the same *perceptuo-motor phase space* than the target (e.g. infant's brain) they are trying to model. Indeed, the *perceptuo-motor phase space* is shaped by the fact that *"some regions of the motor command do almost not change the sound, while others change it abruply"* [31]. Thus, we suggest that, in their computational experiments, modellers control for the potential discrepancy between produced sounds and target sounds. More generally, they should check for *perceptuo-motor phase space* discrepancies. Such an issue could impact learning efficiency.

Some learning frameworks do not take inspiration from biology. Indeed, some reinforcement learning algorithms and Hebbian learning rules used to implement synaptic plasticity are coherent with biology (as described in Section II-E), but some authors proposed biologically implausible learning algorithms (e.g. optimisation algorithms to implement trial-and-error strategies, or particular ways of training internal models). Moreover, it is not easy to cast learning algorithms into clear-cut categories: the ambiguity comes from the fact that different readers might have different definitions or categorisations. For instance, one can think about an architecture where a supervised algorithm is incorporated into a reinforcement learning framework.

The dimensions of the sensory, perceptual and motor spaces greatly vary among models, and the learning architectures do not deal with the same task complexity. Performance can thus not be directly compared between models. The choice of learning framework may constrain the authors to reduce the space dimensions: many learning frameworks and exploration strategies cannot deal with high-dimensional spaces, and brains likely reduce complexity because they cannot control all muscle fibers [110], [111].

In order to find an evaluation strategy and reward function definition, it is convenient to have a low-dimensional preprocessed representation of the sound. To obtain such a representation, several reduction techniques have been used in the reviewed models: PCA and LDA (e.g. [57]), formant extraction (e.g. [5], [10]), or scaling and normalization techniques (e.g. [44] [46]). Ongoing studies try to use Variational Autoencoder (VAE) to help exploration strategies, reducing the goal space to a low-dimensional space while keeping an important part of the information encoded. For instance, Laversanne et al. [112] use a particular type of VAE, called *β-VAE* to achieve this aim.

Models for sensorimotor learning with different motivations (e.g. grasping, recognition) are important complementary studies to take into account while studying vocal learning model. Indeed, these studies contain many important discussions about exploration strategies, target definition, motor space identification [108] [105] that can be useful to take inspiration for future investigation of vocal learning mechanisms. Perceptuo-motor skills, typical of speech production, do also exist in non-vocal gestures [29]. In some of the mentioned studies, other modalities than vocal were used. For example, Forestier and Oudeyer [60] propose two sensorimotor models: a vocal learning model and an action motor learning model. Cohen et al. [47] propose a model of symbol acquisition via active language learning (which combines vocal learning and symbol recognition).

We did not discuss previous modelling of the developmental aspects of vocal learning that investigate the effects of slow changes in the motor control apparatus or sensory system related to growth in the present review. It is, however, important to consider how such slow changes influence vocal production and interact with the learning process [113].

We provided different diagrams, tables, along with segmentation of spaces and functions, as a conceptual tool to analyse and compare existing models of vocal learning. We believe it provides several benefits: to understand the choices of the authors, to look at the biological plausibility of a model or part of it, to compare models systematically, and to give a baseline to build new models. We hope that researchers in the field will agree with our attempt of categorisations and comparisons, and that it will help in further studies to make descriptions more explicit and comparable.

Acknowledgment

This work was supported by the Inria CORDI-S PhD fellowship grant. We really want to thank the anonymous reviewers for all the pertinent remarks, questions and corrections they provided, which was of great help. We thank Jean-Luc Schwartz and Clément Moulin-Frier for helpful discussions. We thank Anthony Strock and Bhargav Teja Nallapu for proofreading the paper.

References

[1] M. Brainard and A. Doupe, "What songbirds teach us about learning," *Nature*, vol. 417, no. 6886, p. 351, 2002.

[2] C. Heyes, "Causes and consequences of imitation," *Trends in cognitive sciences*, vol. 5, no. 6, pp. 253–261, 2001.

[3] ——, "What's social about social learning?" *Journal of Comparative Psychology*, vol. 126, no. 2, p. 193, 2012.

[4] P. Kuhl, "Early language acquisition: cracking the speech code," *Nature reviews neuroscience*, vol. 5, no. 11, p. 831, 2004.

[5] P. Oudeyer, "The self-organization of speech sounds," *Journal of Theoretical Biology*, vol. 233, no. 3, pp. 435–449, 2005.

[6] D. Wolpert and M. Kawato, "Multiple paired forward and inverse models for motor control," *Neural Networks*, vol. 11, no. 7-8, pp. 1317–1329, 1998.

[7] P. Oudeyer, F. Kaplan, and V. Hafner, "Intrinsic motivation systems for autonomous mental development," *IEEE transactions on evolutionary computation*, vol. 11, no. 2, pp. 265–286, 2007.

[8] A. Doupe and P. Kuhl, "Birdsong and human speech: common themes and mechanisms," *Annual review of neuroscience*, vol. 22, no. 1, pp. 567–631, 1999.

[9] P. Kuhl, "A new view of language acquisition," *Proceedings of the National Academy of Sciences*, vol. 97, no. 22, pp. 11 850–11 857, 2000.

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

[11] T. Imada, Y. Zhang, M. Cheour, S. Taulu, A. Ahonen, and P. Kuhl, "Infant speech perception activates broca's area: a developmental magnetoencephalography study," *Neuroreport*, vol. 17, no. 10, pp. 957–962, 2006.

[12] S. Robinson, M. Blumberg, M. Lane, and L. Kreber, "Spontaneous motor activity in fetal and infant rats is organized into discrete multilimb bouts." *Behavioral neuroscience*, vol. 114, no. 2, p. 328, 2000.

[13] P. Wallace and I. Whishaw, "Independent digit movements and precision grip patterns in 1–5-month-old human infants: hand-babbling, including vacuous then self-directed hand and digit movements, precedes targeted reaching," *Neuropsychologia*, vol. 41, no. 14, pp. 1912–1918, 2003.

[14] M. Chakraborty and E. Jarvis, "Brain evolution by brain pathway duplication," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 370, no. 1684, p. 20150056, 2015.

[15] E. Jarvis, "Evolution of vocal learning and spoken language," *Science*, vol. 366, no. 6461, pp. 50–54, 2019.

[16] A. Friederici, "The brain basis of language processing: from structure to function," *Physiological reviews*, vol. 91, no. 4, pp. 1357–1392, 2011.

[17] R. H. Hahnloser and A. Kotowicz, "Auditory representations and memory in birdsong learning," *Current opinion in neurobiology*, vol. 20, no. 3, pp. 332–339, 2010.

[18] F. Theunissen, K. Sen, and A. Doupe, "Spectral-temporal receptive fields of nonlinear auditory neurons obtained using natural sounds," *Journal of Neuroscience*, vol. 20, no. 6, pp. 2315–2331, 2000.

[19] G. Keller and R. H. Hahnloser, "Neural processing of auditory feedback during vocal practice in a songbird," *Nature*, vol. 457, no. 7226, p. 187, 2009.

[20] V. Gallese, L. Fadiga, L. Fogassi, and G. Rizzolatti, "Action recognition in the premotor cortex," *Brain*, vol. 119, no. 2, pp. 593–609, 1996.

[21] E. Oztop, M. Kawato, and M. Arbib, "Mirror neurons and imitation: A computationally guided review," *Neural Networks*, vol. 19, no. 3, pp. 254–271, 2006.

[22] J. Prather, S. Peters, S. Nowicki, and R. Mooney, "Precise auditory–vocal mirroring in neurons for learned vocal communication," *Nature*, vol. 451, no. 7176, p. 305, 2008.

[23] G. Rizzolatti, L. Fadiga, V. Gallese, and L. Fogassi, "Premotor cortex and the recognition of motor actions," *Cognitive brain research*, vol. 3, no. 2, pp. 131–141, 1996.

[24] A. Tramacere, K. Wada, K. Okanoya, A. Iriki, and P. Ferrari, "Auditory-motor matching in vocal recognition and imitative learning," *Neuroscience*, 2019.

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

[26] N. Giret, J. Kornfeld, S. Ganguli, and R. H. Hahnloser, "Evidence for a causal inverse model in an avian cortico-basal ganglia circuit," *PNAS*, vol. 111, no. 16, pp. 6063–6068, 2014.

[27] M. Kawato, "Internal models for motor control and trajectory planning," *Current opinion in neurobiology*, vol. 9, no. 6, pp. 718–727, 1999.

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

[29] C. Fowler, "Speech perception as a perceptuo-motor skill," in *Neurobiology of Language*. Elsevier, 2016, pp. 175–184.

[30] S. Wilson, A. Pinar Saygin, M. Sereno, and M. Iacoboni, "Listening to speech activates motor areas involved in speech production," *Nature Neuroscience*, vol. 7, no. 7, pp. 701–702, Jun. 2004.

[31] 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, Sep. 2012.

[32] C. Boettiger and A. Doupe, "Developmentally restricted synaptic plasticity in a songbird nucleus required for song learning," *Neuron*, vol. 31, no. 5, pp. 809–818, 2001.

[33] L. Ding and D. Perkel, "Long-term potentiation in an avian basal ganglia nucleus essential for vocal learning," *Journal of Neuroscience*, vol. 24, no. 2, pp. 488–494, 2004.

[34] W. Mehaffey and A. Doupe, "Naturalistic stimulation drives opposing heterosynaptic plasticity at two inputs to songbird cortex," *Nature neuroscience*, vol. 18, no. 9, p. 1272, 2015.

[35] M. Sizemore and D. Perkel, "Premotor synaptic plasticity limited to the critical period for song learning," *Proceedings of the National Academy of Sciences*, vol. 108, no. 42, pp. 17 492–17 497, 2011.

[36] R. Legenstein, S. Chase, A. Schwartz, and W. Maass, "A reward-modulated hebbian learning rule can explain experimentally observed network reorganization in a brain control task," *Journal of Neuroscience*, vol. 30, no. 25, pp. 8400–8410, Jun. 2010.

[37] W. Schultz, "Predictive reward signal of dopamine neurons," *Journal of neurophysiology*, vol. 80, no. 1, pp. 1–27, 1998.

[38] J. Goldberg, M. Farries, and M. Fee, "Basal ganglia output to the thalamus: still a paradox," *Trends in neurosciences*, vol. 36, no. 12, pp. 695–705, 2013.

[39] R. Darshan, W. Wood, S. Peters, A. Leblois, and D. Hansel, "A canonical neural mechanism for behavioral variability," *Nature communications*, vol. 8, p. 15415, 2017.

[40] A. Andalman and M. Fee, "A basal ganglia-forebrain circuit in the songbird biases motor output to avoid vocal errors," *PNAS*, vol. 106, no. 30, pp. 12 518–12 523, 2009.

[41] G. Bailly, "Learning to speak. sensori-motor control of speech movements," *Speech Communication*, vol. 22, no. 2-3, pp. 251–267, 1997.

[42] G. Westerman and E. Miranda, "Modelling the development of mirror neurons for auditory-motor integration," *Journal of new music research*, vol. 31, no. 4, pp. 367–375, 2002.

[43] T. Troyer and A. Doupe, "An associational model of birdsong sensorimotor learning i. efference copy and the learning of song syllables," *Journal of Neurophysiology*, vol. 84, no. 3, pp. 1204–1223, 2000.

[44] K. Doya and T. Sejnowski, "A computational model of birdsong learning by auditory experience and auditory feedback," in *Central auditory processing and neural modeling*. Springer, 1998, pp. 77–88.

[45] I. Fiete, M. Fee, and H. Seung, "Model of birdsong learning based on gradient estimation by dynamic perturbation of neural conductances," *Journal of neurophysiology*, vol. 98, no. 4, pp. 2038–2057, 2007.

[46] B. 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.

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

[48] M. Barnaud, J. Schwartz, P. Bessière, and J. Diard, "Computer simulations of coupled idiosyncrasies in speech perception and speech production with cosmo, a perceptuo-motor bayesian model of speech communication," *PloS one*, vol. 14, no. 1, p. e0210302, 2019.

[49] I. Howard and P. Messum, "A computational model of infant speech development," in *SPECOM*, 2007, pp. 756–765.

[50] A. Warlaumont and M. Finnegan, "Learning to produce syllabic speech sounds via reward-modulated neural plasticity," *PloS one*, vol. 11, no. 1, p. e0145096, 2016.

[51] I. Howard and M. Huckvale, "Training a vocal tract synthesiser to imitate speech using distal supervised learning," in *SPECOM*, vol. 2. University of Patras, Wire Communications Laboratory, 2005, pp. 159–162.

[52] A. Philippsen, R. Reinhart, and B. Wrede, "Learning how to speak: Imitation-based refinement of syllable production in an articulatory-acoustic model," in *ICDL-EpiRob*. IEEE, 2014, pp. 195–200.

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

[54] H. Liu and Y. Xu, "Learning model-based f0 production through goal-directed babbling," in *ISCSLP*. IEEE, 2014, pp. 284–288.

[55] M. Murakami, B. Kröger, P. Birkholz, and J. Triesch, "Seeing [u] aids vocal learning: Babbling and imitation of vowels using a 3d vocal tract model, reinforcement learning, and reservoir computing," in *ICDL-EpiRob*. IEEE, 2015, pp. 208–213.

[56] C. Moulin-Frier and P. Oudeyer, "Curiosity-driven phonetic learning," in *ICDL-EpiRob*. IEEE, 2012, pp. 1–8.

[57] A. Philippsen, R. Reinhart, and B. Wrede, "Goal babbling of acoustic-articulatory models with adaptive exploration noise," in *ICDL-EpiRob*. IEEE, 2016, pp. 72–78.

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

[59] I. Howard and P. Birkholz, "Modelling vowel acquisition using the birkholz synthesizer," *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2019*, pp. 304–311, 2019.

[60] S. Forestier and P. Oudeyer, "A unified model of speech and tool use early development," in *CogSci 2017*, 2017.

[61] J. Acevedo-Valle, V. Hafner, and C. Angulo, "Social reinforcement in artificial prelinguistic development: A study using intrinsically motivated exploration architectures," *IEEE Transactions on Cognitive and Developmental Systems*, 2018.

[62] C. Lyon, C. Nehaniv, and J. Saunders, "Interactive language learning by robots: The transition from babbling to word forms," *PloS one*, vol. 7, no. 6, p. e38236, 2012.

[63] C. Moulin-Frier, J. Diard, J. 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.

[64] C. Scharff and F. Nottebohm, "A comparative study of the behavioral deficits following lesions of various parts of the zebra finch song system: implications for vocal learning," *Journal of Neuroscience*, vol. 11, no. 9, pp. 2896–2913, 1991.

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

[66] P. Birkholz, "A survey of self-oscillating lumped-element models of the vocal folds," *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2011*, pp. 47–58, 2011.

[67] K. Ishizaka and J. Flanagan, "Synthesis of voiced sounds from a two-mass model of the vocal cords," *Bell system technical journal*, vol. 51, no. 6, pp. 1233–1268, 1972.

[68] B. Erath, M. Zanartu, K. Stewart, M. Plesniak, D. Sommer, and S. Peterson, "A review of lumped-element models of voiced speech," *Speech Communication*, vol. 55, no. 5, pp. 667–690, 2013.

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

[70] B. De Boer, “Self-organization in vowel systems,” *Journal of phonetics*, vol. 28, no. 4, pp. 441–465, 2000.

[71] ——, *The origins of vowel systems*. Oxford University Press on Demand, 2001, vol. 1.

[72] S. Maeda, “Compensatory articulation in speech: analysis of x-ray data with an articulatory model,” in *First European Conference on Speech Communication and Technology*, 1989.

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

[74] P. Birkholz, “Vocaltractlab–towards high-quality articulatory speech synthesis,” *http://www.vocaltractlab.de/*, Accessed Sept. 2019.

[75] P. Birkholz, D. Jackèl, and B. Kroger, “Construction and control of a three-dimensional vocal tract model,” in *ICASSP*, vol. 1. IEEE, 2006.

[76] J. Gudhnason, D. Mehta, and T. Quatieri, “Evaluation of speech inverse filtering techniques using a physiologically based synthesizer,” in *ICASSP*. IEEE, 2015, pp. 4245–4249.

[77] S. Prom-on, P. Birkholz, and Y. Xu, “Training an articulatory synthesizer with continuous acoustic data.” in *INTERSPEECH*, 2013, pp. 349–353.

[78] 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.

[79] F. Guenther, S. Ghosh, A. Nieto-Castanon, and J. Tourville, “A neural model of speech production,” *Speech production: Models, phonetic processes and techniques*, pp. 27–40, 2006.

[80] J. Tourville and F. Guenther, “The diva model: A neural theory of speech acquisition and production,” *Language and cognitive processes*, vol. 26, no. 7, pp. 952–981, 2011.

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

[82] S. Maeda, “Vtcalcs,” *http://ed268. univ-paris3. fr/lpp/index. php? page= ressources/logiciels*.

[83] I. Howard and P. Messum, “Modeling the development of pronunciation in infant speech acquisition,” *Motor Control*, vol. 15, no. 1, pp. 85–117, 2011.

[84] F. S. Foundation, “http://espeak.sourceforge.net/.accessed 2011 dec 13,” 2007.

[85] L. Alonso, J. Alliende, F. Goller, and G. Mindlin, “Low-dimensional dynamical model for the diversity of pressure patterns used in canary song,” *Physical Review E*, vol. 79, no. 4, p. 041929, 2009.

[86] C. Elemans, J. Rasmussen, C. Herbst, D. Düring, S. Zollinger, H. Brumm, K. Srivastava, N. Svane, M. Ding, O. Larsen *et al.*, “Universal mechanisms of sound production and control in birds and mammals,” *Nature communications*, vol. 6, p. 8978, 2015.

[87] M. Fee, B. Shraiman, B. Pesaran, and P. Mitra, “The role of nonlinear dynamics of the syrinx in the vocalizations of a songbird,” *Nature*, vol. 395, no. 6697, p. 67, 1998.

[88] R. Laje, T. Gardner, and G. Mindlin, “Neuromuscular control of vocalizations in birdsong: a model,” *Physical Review E*, vol. 65, no. 5, p. 051921, 2002.

[89] G. Mindlin, “The physics of birdsong production,” *Contemporary physics*, vol. 54, no. 2, pp. 91–96, 2013.

[90] M. Trevisan, J. Mendez, and G. Mindlin, “Respiratory patterns in oscine birds during normal respiration and song production,” *Physical Review E*, vol. 73, no. 6, p. 061911, 2006.

[91] I. Yildiz and S. Kiebel, “A hierarchical neuronal model for generation and online recognition of birdsongs,” *PLoS Computational Biology*, vol. 7, no. 12, p. e1002303, 2011.

[92] R. Alonso, M. Trevisan, A. Amador, F. Goller, and G. Mindlin, “A circular model for song motor control in serinus canaria,” *Frontiers in computational neuroscience*, vol. 9, p. 41, 2015.

[93] K. Srivastava, C. Elemans, and S. Sober, “Multifunctional and context-dependent control of vocal acoustics by individual muscles,” *Journal of Neuroscience*, vol. 35, no. 42, pp. 14 183–14 194, 2015.

[94] D. Düring, B. Knörlein, and C. Elemans, “In situ vocal fold properties and pitch prediction by dynamic actuation of the songbird syrinx,” *Scientific reports*, vol. 7, no. 1, p. 11296, 2017.

[95] S. Sober, M. Wohlgemuth, and M. Brainard, “Central contributions to acoustic variation in birdsong,” *Journal of Neuroscience*, vol. 28, no. 41, pp. 10 370–10 379, 2008.

[96] A. Amador, Y. Perl, G. Mindlin, and D. Margoliash, “Elemental gesture dynamics are encoded by song premotor cortical neurons,” *Nature*, vol. 495, no. 7439, p. 59, 2013.

[97] S. Boari, Y. Perl, . Amador, . Margoliash, and . Mindlin, “Automatic reconstruction of physiological gestures used in a model of birdsong production,” *Journal of neurophysiology*, vol. 114, no. 5, pp. 2912–2922, 2015.

[98] I. Yildiz and S. Kiebel, “The cmu pronouncing dictionary,” *http://www.speech.cs.cmu.edu/cgi-bin/cmudict*.

[99] S. Najnin and B. Banerjee, “A predictive coding framework for a developmental agent: Speech motor skill acquisition and speech production,” *Speech Communication*, vol. 92, pp. 24–41, 2017.

[100] K. Doya and T. Sejnowski, “A computational model of avian song learning,” in *The new cognitive neurosciences (2nd ed.) Gazzaniga, M. S. (Ed.)*. Cambridge, MA, US: The MIT Press., 2000.

[101] R. Reinhart, *Reservoir computing with output feedback*. PhD Thesis. Bielefeld University, Germany, 2011.

[102] M. Jordan and D. Rumelhart, “Forward models: Supervised learning with a distal teacher,” *Cognitive science*, vol. 16, no. 3, pp. 307–354, 1992.

[103] M. Pickering and S. Garrod, “An integrated theory of language production and comprehension,” *Behavioral and brain sciences*, vol. 36, no. 4, pp. 329–347, 2013.

[104] M. Kawato, “Feedback-error-learning neural network for supervised motor learning,” in *Advanced neural computers*. Elsevier, 1990, pp. 365–372.

[105] M. Rolf, J. Steil, and M. Gienger, “Goal babbling permits direct learning of inverse kinematics,” *IEEE Transactions on Autonomous Mental Development*, vol. 2, no. 3, pp. 216–229, 2010.

[106] T. Sejnowski, “Storing covariance with nonlinearly interacting neurons,” *Journal of mathematical biology*, vol. 4, no. 4, pp. 303–321, 1977.

[107] S. Forestier and P. Oudeyer, “Curiosity-driven development of tool use precursors: a computational model,” in *CogSci 2016*, 2016, pp. 1859–1864.

[108] A. Baranes and P. Oudeyer, “Active learning of inverse models with intrinsically motivated goal exploration in robots,” *Robotics and Autonomous Systems*, vol. 61, no. 1, pp. 49–73, 2013.

[109] J. Liljencrants, B. Lindblom *et al.*, “Numerical simulation of vowel quality systems: The role of perceptual contrast,” *Language*, vol. 48, no. 4, pp. 839–862, 1972.

[110] A. Dhawale, M. Smith, and B. Ölveczky, “The role of variability in motor learning,” *Annual review of neuroscience*, vol. 40, pp. 479–498, 2017.

[111] D. Wolpert, Z. Ghahramani, and J. Flanagan, “Perspectives and problems in motor learning,” *Trends in Cognitive Sciences*, vol. 5, no. 11, pp. 487–494, Nov. 2001.

[112] A. Laversanne-Finot, A. Péré, and P. Oudeyer, “Curiosity driven exploration of learned disentangled goal spaces,” *arXiv preprint arXiv:1807.01521*, 2018.

[113] A. Ghazanfar and D. Liao, “Constraints and flexibility during vocal development: insights from marmoset monkeys,” *Current opinion in behavioral sciences*, vol. 21, pp. 27–32, 2018.

Silvia Pagliarini received her Bachelor degree in Applied Mathematics and her Master Degree in Mathematics from University of Verona (Italy) in 2014 and 2017. From 2017 she is pursuing her Ph. D. degree in the Inria project Team Mnemosyne, at Inria Bordeaux Sud-Ouest and University of Bordeaux. She is currently currying out research in the area of vocal learning models and generative models.

Arthur Leblois was born in Paris, France, in 1980. He received B.S. and M.S. degrees in theoretical physics from the Ecole Normale Suprieure of Paris in 1999 and 2002 respectively, and a PhD degree in Physics from the Universit Pierre et Marie Curie, Paris, France, in 2006. During his PhD, he studied normal and pathological function of the basal ganglia, deep brain nuclei involved in Parkinsons disease. Since his PhD, he studies questions from integrative neurophysiology and combines experimental neuroscience methods and concepts borrowed from physics or applied mathematics. Theoretical models are built based on the available experimental data, and the predictions emerging from models are then tested experimentally. In 2007, he joins the Department of Otolaryngology at the University of Washington, Seattle, USA, as a research fellow to study the function of basal ganglia circuits in song learning and production in songbirds. In 2010, he joins the Center for Neurophysics and Physiology (CNRS, UMR 8119), in Paris, France, and is hired as a full-time researcher at the CNRS in 2012. There, he starts his own research group to study the neural correlates of song learning in songbirds, and more generally to reveal the cellular mechanisms of sensorimotor learning. He also continues the investigation of the pathological dynamics in the basal-ganglia-thalamo-cortical loop in several pathologies, including Parkinsons disease, dystonia as well as absence epilepsy. In 2018, he moves to Bordeaux, France, to join the Institut des Maladies Neurodgnratives (CNRS, UMR 5293) where he studies the normal and pathological function of brain circuits involved in sensorimotor learning.

Xavier Hinaut is research scientist in computational neuroscience, bio-inspired machine learning and developmental language using robots. He graduated his first MSc. in Computer Engineering from the University of Technology of Compiegne, France in 2008 and then obtained a second MSc. in Cognitive Sciences and Artificial Intelligence from the Practical School of High Studies (EPHE) in Paris in 2009. He defended his PhD thesis in Lyon in 2013 under the direction of Peter Dominey on a neurocomputational model processing in sentence templates and complex action sequences. In 2014, he was a postdoctoral researcher at Paris-Saclay Institute of Neuroscience (NeuroPSI) in C. del Negros group to work on the neural encoding of syntax in songbirds. Then, he was awarded a EU Marie Skodowska Curie IF postdoctoral fellowship in 2015 to work in S. Wermters lab, at University of Hamburg (Germany), on cognitive architectures for Human-Robot Interaction through language. Since 2016, he is a permanent research scientist at Inria, Bordeaux, France, in the Mnemosyne team.