Decode, Move and Speak! Self-supervised Learning of Speech Units, Gestures, and Sound Relationships Using Vocal Imitation

Marc-Antoine Georges
GIPSA-lab, Université Grenoble Alpes
marc-antoine.georges@grenoble-inp.fr

Marvin Lavechin
GIPSA-lab, Université Grenoble Alpes
marvin.lavechin@grenoble-inp.fr

Jean-Luc Schwartz
GIPSA-lab, Université Grenoble Alpes
jean-luc.schwartz@grenoble-inp.fr

Thomas Hueber*
GIPSA-lab, Université Grenoble Alpes
thomas.hueber@grenoble-inp.fr

Speech learning encompasses mastering a complex motor system to produce speech sounds from articulatory gestures while simultaneously uncovering discrete units that provide entry to the linguistic system. Remarkably, children acquire these associations between speech sounds, articulatory gestures, and linguistic units in a weakly supervised manner, without the need for explicit labeling of auditory inputs or access to target articulatory gestures. This study uses self-supervised deep learning to investigate the respective roles of sounds, gestures, and linguistic units in speech acquisition and control. In a first experiment, we analyzed the quantized representations learned by vector-quantized variational autoencoders (VQ-VAE) from ground truth acoustic and articulatory data using ABX tests. We show an interesting complementarity between acoustic and articulatory modalities that may help in the discovery of phonemes. In a second experiment, we introduce a computational agent that repeats auditory speech inputs by controlling a virtual vocal apparatus. This agent integrates an articulatory synthesizer capable of reproducing diverse speech stimuli from interpretable parameters, along with two internal models implementing the articulatory-to-acoustic (forward) and acoustic-to-articulatory

* Corresponding author.

Action Editors: Marianna Apidianaki, Abdellah Fourtassi, and Sebastian Padó. Submission received: 20 December 2023; revised version received: 1 May 2024; accepted for publication: 19 June 2024.

https://doi.org/10.1162/coli_a_00532

(inverse) mapping, respectively. Additionally, two inductive biases are used to regularize the ill-posed acoustic-to-articulatory inverse mapping. In line with the first experiment, we explore the complementarity between the auditory input and the articulatory parameters inferred by the agent. We also evaluate the impact of discretizing auditory inputs using VQ-VAE. While the majority of the agent's productions are intelligible (according to perceptual evaluations), our analysis highlights inconsistencies in the underlying articulatory trajectories. In particular, we show that the agent's productions only partially reproduce the complementarity between the auditory and articulatory modalities observed in humans.

1. Introduction

Learning to speak involves the acquisition of a complex set of relationships between articulatory gestures, sounds, and meaning. Gestures are the *proximal* variables under the speaker's control to generate a message conveyed by a linguistic sequence; sounds are the *distal* physical output of these gestures, which are captured by the listener's ear and processed by the listener's brain to recover the speaker's message eventually. The interaction between the auditory and motor systems has been at the center of long debates in theories of speech perception. These debates contrast proponents of auditory theories (Diehl, Lotto, and Holt 2004; Kluender 1994), for whom listeners extract acoustic features by comparing them to stored acoustic templates, and proponents of articulatory/motor theories (Liberman et al. 1967; Liberman and Mattingly 1985), for whom perceiving speech is perceiving "gestures," that is, extracting articulatory information from the acoustic signal. Supported by neurocognitive findings on the relationships between perceptual and motor components, recent studies propose unifying auditory and motor theories into a perceptuo-motor framework in which speech sounds and inferred articulatory gestures are combined in the brain to access linguistic content (Bever and Poeppel 2010; Schwartz et al. 2012; Skipper et al. 2007). This interplay between sounds and gestures is also central in the literature on speech production and speech motor control through the concept of internal models (Kawato 1999; Parrell et al. 2019; Houde and Nagarajan 2011; Perrier 2012), which are supposed to be the neurocognitive means enabling prediction of the relation between a sensory input and the underlying motor commands.

An essential aspect of the debates on the use of articulatory information in speech representation learning concerns the existence of invariant features and phoneme categories. Auditory theories propose various types of invariant auditory features (e.g., Stevens 1989), while motor theories assume that invariance cannot be found in the acoustic signal alone due to coarticulation but should instead be found in the inferred articulatory gestures. Still, the existence of invariant features, whatever they are, is itself questionable. As Lindblom suggested in his H&H theory (Lindblom 1990), a phoneme could be defined by domains of acceptable acoustic cues rather than strict invariance. Moreover, the existence of phonemes itself can also be questioned. Considering the lack of clearly distinctive acoustic cues, several authors (e.g., Port and Leary 2005; Bybee 1999) attempt to replace formal phonology with emergent phonetic structures. Here, access to meaning could occur from a combination of phonetic cues, language experience, and contextual information, without requiring phonemes at all in the speech communication process.

These questions on the nature of phonetic representations in relation with the respective roles of articulatory and acoustic information have been largely studied using behavioral and neuroimagery experiments. In the present study, we adopt a

complementary approach based on modeling by studying how a computational agent can learn these relationships from its environment. This allows us to quantitatively analyze the informational content of articulatory gestures and sounds in relation to discrete phonetic categories and to gain new insights into how infants might acquire these relationships.

Recently, large-scale deep learning simulations have been used to design computational models of infant language development (Cruz Blandón, Cristia, and Räsänen 2023). As proposed in Dupoux (2018), a deep neural network fed with a large amount of raw and unlabeled data can be used to "reverse-engineer" speech acquisition in humans. For instance, the developmental trajectories of infants can be approximated by monitoring the evolution of the model's performance with respect to the amount of training data. This paradigm was used recently in Lavechin et al. (2024) to study the perceptual narrowing effect and in Schatz et al. (2021) to study the nature of the discrete units potentially learned by infants at the early stages of their development (which were found to be much shorter and fine-grained than phonemes).

In line with this literature, several studies explicitly considered the role of articulatory information (e.g., Bailly 1997; Kröger, Kannampuzha, and Neuschaefer-Rube 2009; Kröger, Kannampuzha, and Kaufmann 2014; Howard and Messum 2014; Rasilo and Räsänen 2017; Moulin-Frier, Nguyen, and Oudeyer 2014; Philippsen, Reinhart, and Wrede 2014; Philippsen 2021; Patri, Diard, and Perrier 2015; Laurent et al. 2017; Chen, Lammert, and Parrell 2021; Georges et al. 2022; Beguš et al. 2023; Coen 2006; Murakami et al. 2015; Krug et al. 2023a, b; Guenther 1995). In most of them, a computational agent learns to control an artificial vocal tract (often implemented as an articulatory synthesizer) to reach an acoustic target. Similar to children learning speech, the agent does not know the articulatory configuration associated with this acoustic target. Using a trial-and-error procedure, it has to learn the ill-posed acoustic-to-articulatory inverse mapping (Atal et al. 1978) in a self-supervised manner (some studies investigated the contribution of the visual modality, i.e., the vision of the interlocutor's lips, as an additional supervision signal, e.g., Bailly 1997; Murakami et al. 2015; Krug et al. 2023b). A first set of promising results has been reported in the literature with acoustic targets provided by synthetic data generated by an articulatory synthesizer (e.g., Kröger, Kannampuzha, and Neuschaefer-Rube 2009; Chen, Lammert, and Parrell 2021; Philippsen, Reinhart, and Wrede 2014; Philippsen 2021; Moulin-Frier, Nguyen, and Oudeyer 2014; Laurent et al. 2017; Patri, Diard, and Perrier 2015; Murakami et al. 2015; Krug et al. 2023b). In other studies, the acoustic target is given by natural speech recordings, which can be made of isolated vowels (Coen 2006; Rasilo and Räsänen 2017), vowel-consonant-vowel (VCV) sequences (Bailly 1997), or more complex material such as isolated words and full sentences (Krug et al. 2023a; Beguš et al. 2023; Georges et al. 2022).

Crucially, most of these studies do not generally consider the joint learning of the acoustic-to-articulatory mapping on the one hand and of discrete units on the other. Indeed, in some studies, the learning-by-imitation process does not involve any explicit decoding of discrete units (e.g., Beguš et al. 2023; Georges et al. 2022) whereas in others, these units, defined as phonemes, are given to the agent (either as part of the objective function to optimize [Krug et al. 2023b] or as an additional input [Krug et al. 2023a]). To the best of our knowledge, only a few computational models of speech learning address the joint learning problem (Howard and Messum 2014; Rasilo and Räsänen 2017; Philippsen 2021). However, the proposed simulations do not explicitly target the question of the discovery of phonetic units in relation with the respective (and potentially complementary) roles of the acoustic and articulatory modalities, which is the

main goal of the present study. Here, by means of computational modeling and simulation from natural speech data, we investigate (1) to what extent the use of articulatory information can contribute to the discrimination of phonemic categories, and (2) how a computational agent can retrieve such articulatory information from audio-only speech inputs.

To address these complementary research questions, we propose the two following experiments. In the first one, described in Section 2, vector-quantized variational autoencoders (VQ-VAE; Van Den Oord, Vinyals et al. 2017) learn quantized representations either from ground truth articulatory data or from their acoustic counterparts. By extending the ABX methodology (Schatz et al. 2013), we show an interesting complementarity between the representations learned from the acoustic data and those learned from the articulatory data, the former being rather related to the consonant manner of articulation, the latter to the place of articulation. We also show that the discrete units learned from the articulatory data are significantly longer than the ones learned from the acoustic data. Still, both are much shorter than typical phones.

In a second series of experiments, reported in Section 3, we investigate how a child can learn the acoustic-to-articulatory mapping in a self-supervised manner. To that purpose, we develop a computational agent based on deep neural networks that learns to speak by controlling its artificial vocal apparatus to imitate any acoustic input. The proposed agent combines (i) a pre-trained neural articulatory synthesizer able to reproduce complex speech stimuli from a limited set of interpretable articulatory parameters, and (ii) a forward and an inverse internal model representing how the acoustic-articulatory relationships are internalized in the brain. We contrast two versions of the agent, the first one called the "continuous" agent comprising only these two sets of components, and the second one called the "discrete" version, also comprising an additional unit discovery module based on VQ-VAE, similar to that used in the first series of experiments. The speech unit discovery module and the two internal models (forward and inverse) are jointly trained in a self-supervised manner from raw acoustic-only speech data. We also investigated two mechanisms, referred to here as inductive biases, to constrain the inverse model. The first aims to maintain plausible articulatory configurations (independently from any acoustic input), and the second encourages the articulatory gestures to be smooth over time by minimizing their jerk. The evaluation uses the same ABX methodology as for ground truth data and includes a perceptual evaluation of the agent's productions. This evaluation enables us to assess the gap between the agent and human articulatory capabilities, providing interesting perspectives for further developments. We finally discuss in Section 4 these simulation results in relation with the literature on acoustic vs. articulatory invariance and on the development of speech production, and present some major open questions.

All the corresponding source code, including the specification of the articulatory synthesizer, and the complete agent, is made available for further experiments.[1]

2. Experiment 1: Role of Articulatory Features in Learning Phonetic Units

Using self-supervised learning algorithms, we investigate here the role of articulatory knowledge in learning phonetic units. We focus on the categorization of consonants in varying vocalic contexts, known to be difficult from acoustic input, as mentioned in the Introduction. In line with recent studies conducted in the scope of the zero-resource

1 The source code can be downloaded at `https://github.com/georgesma/agent`.

challenge (Dunbar et al. 2019; Morita and Koda 2020), we use the VQ-VAE model to learn quantized representations that can provide a basis for phonetic categories. VQ-VAEs were trained from ground truth articulatory data derived from available datasets for English and French, and compared with VQ-VAEs trained from their acoustic counterpart. The present series of simulations extends those in Georges, Schwartz, and Hueber (2022) with a larger number of speakers and new evaluation metrics.

2.1 Datasets

All experiments reported in this article are based on the two publicly available articulatory-acoustic datasets PB2007 and MOCHA-TIMIT.

PB2007 dataset (Badin et al. 2022). This corpus contains recordings of 1,109 items including two repetitions of the 14 isolated French vowels (3% of the corpus), two repetitions of the 224 VCV sequences where C and V are, respectively, one of the 16 consonants and 14 vowels of French (51%), two repetitions of 109 pairs of CVC French words (36%), and 88 sentences (10%). Once the silences have been removed, this corpus represents 17 minutes of speech. Movements of the tongue, jaw, and lips during speech production were recorded at 200 Hz using a 2D Carstens electromagnetic articulograph (EMA) synchronously with the speech sound. For this aim, 6 EMA coils were attached on the jaw, lower and upper lips, and tongue (apex, middle, and back of the tongue). The phonetic segmentation of the acoustic recordings was first obtained with decoder based on hidden Markov models and then manually corrected.

MOCHA-TIMIT (Wrench 2000). This corpus consists of the productions of 6 English speakers, 2 male and 4 female (we excluded the 7th, [female] speaker *falh0* from our experiments since there were issues in the articulatory recordings). Each speaker produced the same 460 short sentences for durations varying from 17 to 20 minutes once silences were removed. The articulatory trajectories were acquired with a Carstens articulograph at a sampling rate of 500 Hz using the same setup as for the PB2007 corpus with an additional coil attached to the velum. The phonetic segmentation of the acoustic recordings was obtained automatically using the pre-trained acoustic models and the pronunciation dictionary available in the Montreal Forced Aligner toolkit (McAuliffe et al. 2017).

2.2 Acoustic Features

For both datasets (PB2007 and MOCHA-TIMIT), the spectral content of the acoustic recordings (initially recorded at 16 kHz) was encoded into a sequence of 18 Bark-scale cepstral coefficients (following a short-term Fourier analysis parameterized by a window size of 20 ms and a hop size of 10 ms). This representation was chosen with the use of the LPCNet neural vocoder in mind (Valin and Skoglund 2019). This vocoder explicitly dissociates source (f0 and periodicity) and filter (cepstral coefficients) features, making it well-suited to be interfaced with articulatory data.

2.3 Articulatory Features

First, EMA trajectories were downsampled at 100 Hz in order to match the analysis rate of the acoustic signal. Then, raw EMA coordinates were reduced into lower dimension vectors referred to here as "articulatory features." Those features are expected to reflect

the degrees of freedom of the vocal tract, notably by decoupling the movement of the jaw with respect to the movements of the tongue and the lower lip (which is not the case when considering the raw EMA data). To that purpose, we followed the procedure described in Maeda (1990) and Serrurier et al. (2012) and referred to as a guided Principal Component Analysis (guided-PCA).

Guided-PCA consists of sequentially extracting interpretable articulatory features from restricted sets of coil data. Firstly, a jaw height parameter (JH) is defined as the first component of a PCA of the EMA coil attached to the lower incisor. Its contribution to the movement of the three tongue coils is then estimated using linear regression, and a movement residue for tongue coils is computed by removing this JH contribution. Then two new articulatory features are introduced, namely, tongue body (TB) and tongue dorsum (TD), provided by a PCA applied to the coordinates of the residual movement (removing JH contribution) of the two back EMA tongue coils on the middle and the back of the tongue. The contribution of TB and TD to the residual movements of the coil attached to the tongue tip coil is then estimated by linear regression and removed to compute a final movement residual. PCA applied to the two coordinates of this residual provides the tongue tip parameter TT. In the same way, for lip features, the JH contribution is first removed from the movements of the two lip coils. Then, two PCA analyses separately realized on the set of vertical vs. horizontal movements of the upper and lower lip coils provide the lip height (LH) and lip protrusion (LP) features, respectively. Finally, and only for the MOCHA-TIMIT corpus, an additional articulatory feature VL is obtained by reducing the 2D coordinates of the velum coils into a 1D feature using PCA. An example of extracted articulatory features for corpus PB2007 is given in Figure 1.

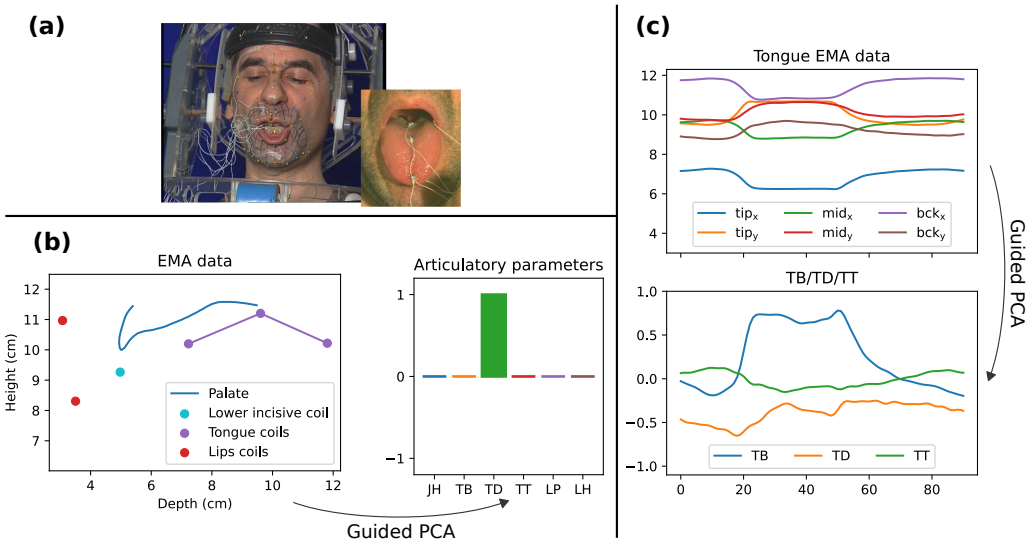

Figure 1
(a) Experimental setup used for the recording of the PB2007 dataset (6 coils attached on tongue, lips, jaw). (b) Extraction of interpretable articulatory features from raw EMA coordinates using a guided-PCA procedure. (c) Trajectories of the 3 EMA coils attached to the tongue (*tip*, *mid*, *bck*). while producing the sequence /ata/ and the corresponding articulatory features: Tongue Body (TB), Tongue Dorsum (TD), and Tongue Tip (TT), bottom.

2.4 Speech Unit Encoder Based on VQ-VAE

A vector quantized variational autoencoder (VQ-VAE; Van Den Oord, Vinyals et al. 2017) can be seen as a discrete version of a variational autoencoder (VAE, Kingma and Welling 2014; Rezende, Mohamed, and Wierstra 2014). It has an encoder-decoder architecture. The decoder defines a posterior distribution $p_\theta(\mathbf{x}|\mathbf{z})$ of the input observation $\mathbf{x}$ given a latent variable $\mathbf{z}$. The parameters of this distribution are provided by a DNN (with weights θ). Symmetrically, the encoder defines a posterior distribution $q_\phi(\mathbf{z}|\mathbf{x})$ of the latent variable $\mathbf{z}$ given an input observation $\mathbf{x}$ (also parameterized by a DNN with weights ϕ). Contrary to the VAE, this posterior distribution is categorical, such as:

$$q_\phi(\mathbf{z} = \mathbf{e}_k \mid \mathbf{x}) = \begin{cases} 1 & \text{for } k = \text{argmin}_j \|z_e(\mathbf{x}) - \mathbf{e}_j\| \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

with $z_e(\mathbf{x})$ the output of the encoder, and $\{\mathbf{e}_i\}$ with $i \in 1, 2, \ldots, K$ a set of K D-dimensional vectors (representations). k is defined as the index of the closest vector $\mathbf{e}_k$ with respect to the output of the encoder (nearest neighbor look-up). Therefore, in a VQ-VAE, the latent variable $\mathbf{z}$ is discrete. Later on, we will refer to the discrete latent variable $\mathbf{z}$ associated with a given $\mathbf{x}$ as $z_q(\mathbf{x})$. This representation vector is then used as the input of the decoder. The codebook (i.e., the set of learned representations) is estimated from the data in an unsupervised manner, in addition to the parameters of both encoder and decoder neural networks. The loss function of the VQ-VAE optimized during training can be expressed as follows:

$$L_{VQ\text{-}VAE} = -\log p_\theta(\mathbf{x}|\mathbf{z} = \mathbf{e}_k) + \|\text{sg}[z_e(\mathbf{x})] - \mathbf{e}_k\|^2 + \beta\|z_e(\mathbf{x}) - \text{sg}[\mathbf{e}_k]\|^2 \tag{2}$$

where sg[] denotes the stop-gradient operation, and β is a weighting term.

As illustrated in Figure 2, for each speaker and each dataset, we trained VQ-VAEs for single modalities, that is, one VQ-VAE from articulatory features only, referred to as the "articulatory VQ-VAE" and a second one from the corresponding acoustic features ("acoustic VQ-VAE"). As a first way of combining both modalities, we investigated an "early fusion" strategy based on the concatenation of articulatory and acoustic feature vectors (the resulting model is named "articulatory-acoustic VQ-VAE"). A second way of combining the modalities consists in "late fusion" in which the outputs of the two single-modality VQ-VAEs jointly define the representation associated with a given frame. We discuss later in Section 2.6.1 how these two components are combined for characterizing the multimodal representation content.

2.5 Implementation Details

First, contextual feature vectors $\mathbf{x_t} \pm \tau = [\mathbf{x}_{t-\tau}^T; \cdots; \mathbf{x}_{t+\tau}^T]^T$ (with T the transpose operator) are built for each feature vector $\mathbf{x_t}$ and used as input of each VQ-VAE. We set $\tau = 2$, bringing the temporal span of a learned discrete unit to 50 ms. This choice was mainly based on previous studies in the field for which a 1D convolutional layer is typically used to downsample the input sequence (as in Tjandra, Sakti, and Nakamura 2020, where a stride parameter equal to 4 is used in the convolution). For each VQ-VAE, the encoder was built with 3 fully connected layers and the hyperbolic tangent used as the activation function (dropout and batch normalization layers were inserted after each fully connected layer with a dropout ratio of 0.25), and with a final linear layer of

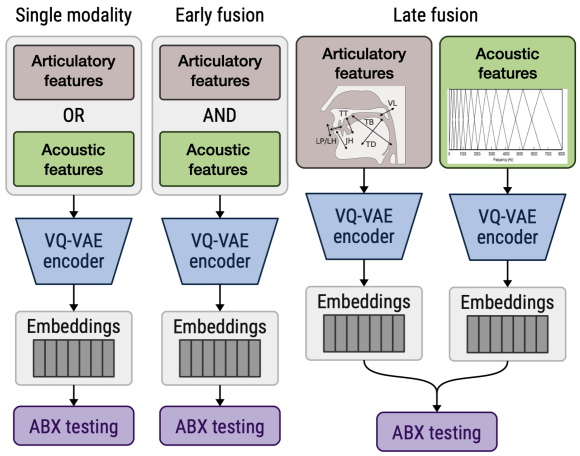

Figure 2
Proposed framework for learning discrete speech units from articulatory and/or acoustic speech data using VQ-VAEs.

the size the dimension of the codebook vectors (i.e., D). A similar structure was used for the decoder, but the final linear layer was adapted to the size of the input data. Similarly to Van Den Oord, Vinyals et al. (2017), the value of the weighting term β in the VQ-VAE loss function (Equation (2)) was set to 0.25. For each experiment, training data were z-scored. Model training was done using back-propagation with the Adam optimizer. Each mini-batch was made of all contextual feature vectors $\mathbf{x_t} \pm \tau$ extracted from 8 training sentences randomly selected. The implementation was done using the *PyTorch* toolkit (Paszke et al. 2019). For each simulation, the datasets were randomly partitioned with 80% of the data used for training and the remaining 20% used for testing; 20% of the training set was used as a validation set to control the early stopping and optimize several hyperparameters with a dedicated optimization procedure described later in Section 2.6.2.

2.6 Metrics

2.6.1 The Machine ABX Sound Discrimination Test. ABX tests were used to assess the phonetic properties of the latent representations learned by the different VQ-VAEs. ABX tests are based on the idea that the learned representations of two occurrences of the same category (A and X) should be closer to one another than to an occurrence of a different category (B). In this work, we use ABX tests to assess how articulatory vs. acoustic modalities may contribute to the discovery of phonetic units. This is how we implement the concept of "invariance," which corresponds to the search for phonetic (articulatory or acoustic) features likely to distinguish between two phonemic categories. As stated in the Introduction, we focused on consonants in varied left and right vocalic contexts. For each speaker (one for the PB2007 dataset and 6 for the MOCHA-TIMIT dataset), we extracted all VCV sequences. From these sequences, and for all consonants, we built a set of triplets A, B, and X. A and X are the representations associated with two occurrences of the same consonant but potentially in a different vocalic context (all possible vocalic contexts were represented in proportion to their frequency in the corresponding evaluation set). B is the representation associated with an occurrence of

a different consonant. We then compared the distance between A and X on one hand and between B and X on the other.

The distance was defined as the mean frame-wise cosine distance along a dynamic time warping (DTW) path, computed as follows: (1) for each VCV (A, B, and X), we extracted the corresponding sequence of quantized representations from the VQ-VAE; (2) we kept only the frames corresponding to the central consonant relying on the available segmentation of each corpus at the phonetic level; (3) we time-aligned the consonants of A and X on the one hand, and B and X on the other using the DTW algorithm; and (4) we calculated the average cosine distances along the DTW path for both A vs. X ($d_{A,X}$) and B vs. X ($d_{B,X}$). A single ABX test is considered "passed" if $d_{A,X} \leq d_{B,X}$. To reduce the computational cost, this ABX test was not done for all possible A, B, and X triplets in each dataset, but only from a randomly selected subset of 5,000 triplets, ensuring that each (A, B) pair is evenly represented. Then, for the corresponding simulation condition, a global discriminability score ("global ABX score") was defined as the average success rate of all individual ABX tests.

In addition to the early-fusion strategy mentioned above (see Figure 2), we investigated another approach for combining the two modalities for the ABX tests. When processing A and X stimuli, we computed $d^{merge}_{A,X} = \omega.d^{ac}_{A,X} + d^{art}_{A,X}$ where $d^{ac}_{A,X}$ is the distance between A and X obtained for the acoustic VQ-VAE (using the DTW-based procedure described above), $d^{art}_{A,X}$ is the distance between A and X obtained with the articulatory VQ-VAE, and ω a weight factor. We did the same for the (B, X) pair. A single ABX test was then considered "passed" when $d^{merge}_{A,X} < d^{merge}_{B,X}$. This approach is referred to as "late fusion."

2.6.2 Hyperparameter Tuning. Since the aim here is to investigate whether, and to what extent, the articulatory modality can provide additional and complementary information to the problem of finding discriminant phonetic representations, we looked for the model architecture that directly optimizes this aspect, that is, the average ABX score (on the validation set). For that purpose, we used the TPE algorithm (Tree of Parzen Estimators) (Bergstra et al. 2011) implemented in the *hyperopt* package (Bergstra, Yamins, and Cox 2013) to optimize the number of fully connected layers ($N_l \in [1, 2, 3, 4]$), the number of neurons in each layer ($N_n \in [64, 128, 256, 512]$), the dropout ratio, the learning rate, the size of the codebook (with $K \in$ [32, 64, 128, 256, 512, 1,024], and $D \in$ [8, 16, 32, 64, 128, 256]), with respect to the ABX score. For each set of hyperparameters, we trained 5 different models on random splits of the datasets and averaged their performance.

2.6.3 Unit Duration and Phone-normalized Mutual Information. To analyze the extent to which the learned discrete units share similarities with phones, we computed the average duration of the learned discrete units and their phone-normalized mutual information (PNMI), standardly used to investigate clustering quality (e.g., Hsu et al. 2021). Given a sequence of discrete units, the average duration is simply computed as the average duration over which the same unit is repeated.

Regarding the PNMI between the discrete units learned by our VQ-VAE algorithm $\mathbf{z} = [\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_T]$ and their aligned frame-level phonetic transcription $\mathbf{y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_T]$, it is computed as:

$$\text{PNMI}(\mathbf{y}, \mathbf{z}) = \frac{\sum_i \sum_j p_{\mathbf{yz}}(i,j) \log \frac{p_{\mathbf{yz}}(i,j)}{p_{\mathbf{y}}(i) p_{\mathbf{z}}(j)}}{\sum_i p_{\mathbf{y}}(i) \log p_{\mathbf{z}}(j)} \tag{3}$$

Table 1
Optimization of the VQ-VAEs hyperparameters (Experiment 1). Format: number of layers x number of neurons per layer / K / D (K = number of representations and D = dimension of representations).

Speaker ID	Acoustic	Articulatory	Early Fusion
PB2007	2×512 / 512 / 64	3×256 / 512 / 64	2×128 / 512 / 128
MOCHA (fsew0)	2×512 / 512 / 64	2×256 / 512 / 8	2×128 / 512 / 128
MOCHA (mjjn0)	2×512 / 512 / 64	3×64 / 128 / 64	2×128 / 512 / 128
MOCHA (faet0)	2×512 / 512 / 64	3×256 / 512 / 64	2×128 / 512 / 128
MOCHA (fjmw0)	2×512 / 512 / 64	3×256 / 512 / 64	2×128 / 512 / 128
MOCHA (ffes0)	2×512 / 512 / 64	1×128 / 128 / 8	2×128 / 512 / 128
MOCHA (msak0)	2×512 / 512 / 64	1×512 / 128 / 16	2×128 / 512 / 128

where i is the i-th phone category, j is the j-th VQ-VAE category, $p_{\mathbf{y}}$ (resp. $p_{\mathbf{z}}$) is the probability distribution over y (resp. over z), and $p_{\mathbf{yz}}$ is the joint probability distribution between $\mathbf{y}$ and $\mathbf{z}$. The PNMI measures the percentage of uncertainty about the phone label y removed from observing the VQ-VAE label z. Higher PNMI indicates better clustering quality.

2.7 Results

We first present in Table 1 the results of the hyperparameter tuning for both the PB2007 and the MOCHA-TIMIT dataset.

Interestingly, the tuning procedure does not always select the most complex architecture or the largest codebook. This is particularly true for the articulatory VQ-VAEs (with a codebook of limited size for the speakers msak0, ffes0, mjjn0).

We now discuss the average ABX score for the acoustic and articulatory VQ-VAEs and the early fusion strategy presented in Figure 3. We observe a significant difference between the PB2007 and the MOCHA-TIMIT dataset, with, for instance, a 92% vs. 82% consonant discrimination rate for the acoustic VQ-VAEs (similar trends are observed for the articulatory and the early-fusion conditions). This result can be understood in terms of linguistic content, which varies considerably from one dataset to the next: The MOCHA-TIMIT dataset is made up of sentences only, while the PB2007 dataset contains a high proportion of VCVs, isolated vowels, and short words likely to be less prone to coarticulation. Consequently, we now dissociate these two data sets and report in Figure 3 the average ABX scores on the MOCHA-TIMIT dataset only. First, no significant difference was found between the acoustic and the articulatory modalities in terms of ABX score (82.8% vs. 82.36%). However, the ABX score obtained when combining the two modalities is significantly higher (88%). These results suggest that the two modalities carry complementary information that can be combined by the developing child to access phonemic information.

To better understand the complementarity between the acoustic and articulatory modalities, we display in Figure 4 the average ABX discriminability scores for each consonant pair for the PB2007 dataset (a similar tendency is observed on the MOCHA-TIMIT dataset). We observe that for the articulatory VQ-VAE, the discriminability scores are lower for consonant pairs with the same place of articulation, for example, coronal (/s/ *vs.* /d/) or labial (/f/ *vs.* /b/), while for the acoustic VQ-VAE, scores are lower for pairs of consonants with the same manner of articulation, for example, unvoiced

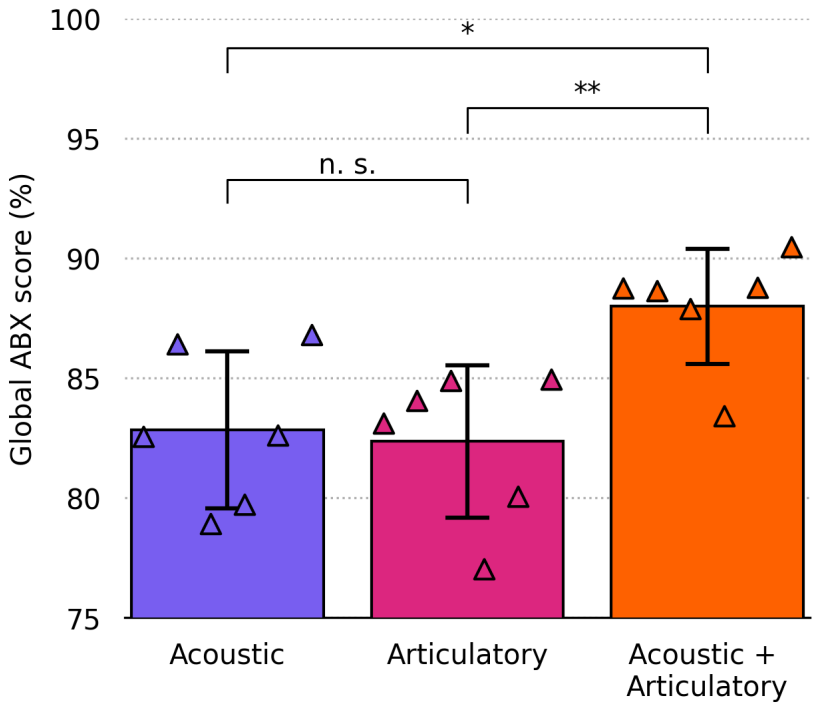

Figure 3
Global ABX scores obtained with respect to the modality considered in the VQ-VAE based experiments (acoustic/articulatory/early-fusion of the two) for the 6 speakers of the MOCHA-TIMIT dataset. Error bars represent standard errors computed across speakers and statistical significance is assessed using paired t-tests on the logits.

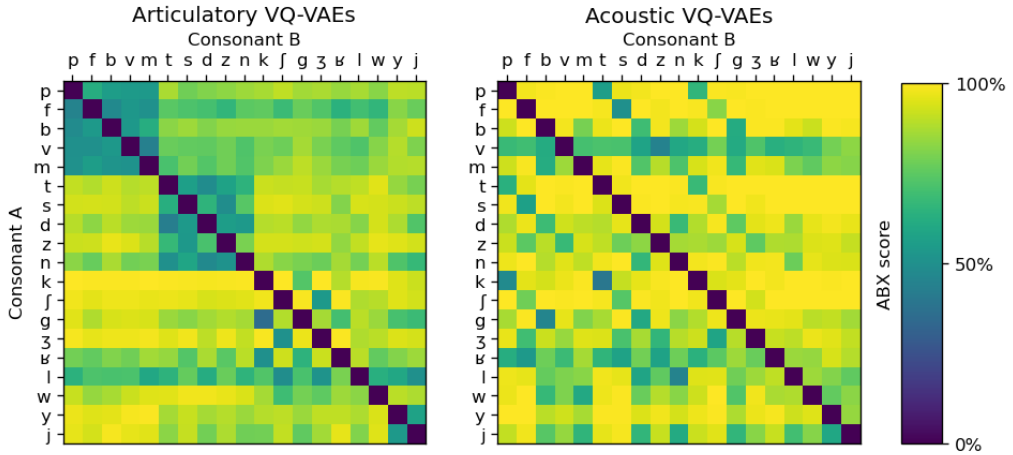

Figure 4
ABX scores for each pair of consonants computed for dataset PB2007.

fricatives (/f/ *vs.* /s/). Therefore, it seems that the learned representations have the same power in terms of phonetic discriminability while being different and complementary in nature.

We carried out a more detailed analysis of how each type of representations enables the discrimination of consonants according to their place and manner of articulation. To this end, we defined two distinct ABX discriminability scores, respectively called "ABX manner" and "ABX place." To calculate the former (focused on the manner of articulation), we grouped consonants into three subgroups with similar places of articulation: labial, coronal, and dorsal. We applied the ABX test methodology within each group (for example, for the labial group, A is /abo/, X is /iba/, and B is /uvo/) and calculated the average success rate for the three groups. To calculate the second score (related to place of articulation), we applied the same procedure, but after grouping consonants according to their manner of articulation. We considered the following five subgroups: voiced occlusive consonants, unvoiced occlusive consonants, voiced fricatives/affricates, unvoiced fricatives/affricates, and sonants (i.e., liquids and nasals). The two discriminability scores "ABX manner" and "ABX place," for both the acoustic and articulatory modalities, as well as for the early fusion and late fusion strategies, and for all speakers considered in this study, are presented in Figure 5. The results confirm that the articulatory VQ-VAE provides latent representations that better discriminate the place of articulation, while the acoustic VQ-VAE structures the latent space primarily in terms of manner of articulation. Interestingly, the early-fusion strategy model provides slightly better results than the articulatory-only VQ-VAE in terms of place of articulation and moderately poorer performance than the acoustic-only VQ-VAE in terms of articulation mode. As for the late fusion strategy, by modulating the contribution of representations from the acoustic and articulatory VQ-VAE (i.e., ω varying between 10^{-1} and 10^{1}), performances follow a trajectory with an optimal value that takes almost the best of each modality, namely, with values equaling the articulatory VQ-VAE for

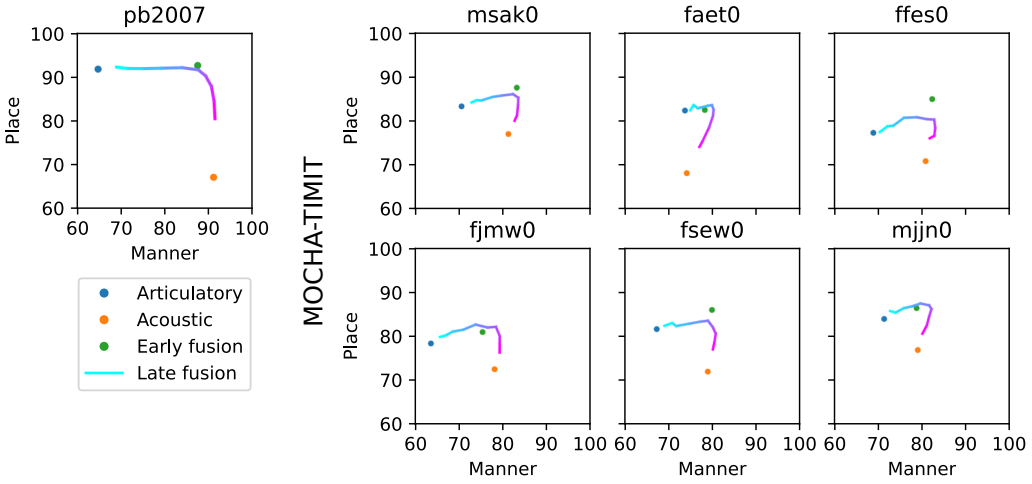

Figure 5
ABX scores with respect to place and manner of articulation for single modalities compared with early and late fusion strategies. Late fusion performance varies with ω ranging from 10^{-1} (cyan) to 10^{1} (purple). Left: Results for the single speaker of PB2007 dataset. Other plots: Results for the 6 speakers of the MOCHA-TIMIT dataset.

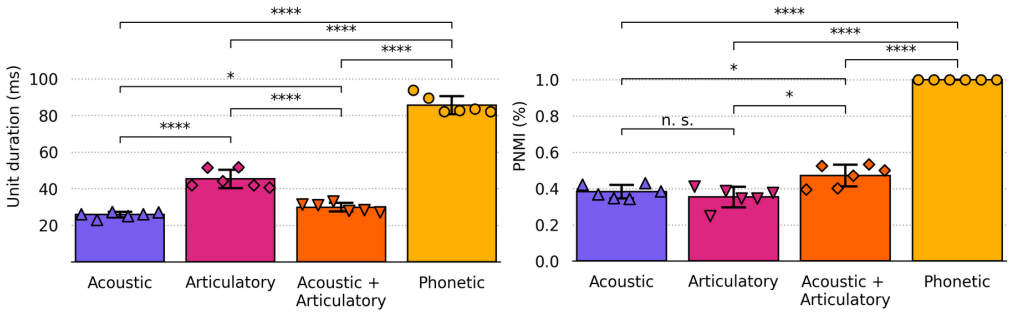

Figure 6
(Left) Average unit duration (in ms) for the acoustic, articulatory, and acoustic-articulatory VQ-VAE with a codebook of size 512 in comparison with the average phone duration. (Right) Phone-normalized mutual information for the same acoustic, articulatory, and acoustic-articulatory VQ-VAEs in comparison with the phone normalized mutual information. Each dot represents an independent training on one of the six speakers from the MOCHA-TIMIT dataset included in this study. Error bars represent standard errors computed across speakers and statistical significance is assessed using paired t-tests on the logits.

place and the acoustic VQ-VAE for manner. Altogether, these experiments show that both fusion strategies can efficiently take advantage of the acoustic and articulatory modalities for phoneme categorization.

We now analyze the temporal similarities between the learned discrete representations and actual phones. An analysis of the duration (left graph of Figure 6) reveals that articulatory units are almost twice longer than acoustic units (45 ms vs. 26 ms). Early fusion of acoustic and articulatory features yields shorter units as compared to articulatory units (30 ms vs. 45 ms). In all conditions, the learned discrete units are much shorter than actual phones with an average duration of 86 ms.

The PNMI scores presented in the right graph of Figure 6 indicate that acoustic units share slightly more information with actual phones than articulatory units (PNMI of .38 for acoustic units versus .35 for articulatory units). Fusing acoustic and articulatory features yields an increase in PNMI, reaching a value of .47. This increase confirms that acoustic and articulatory representations contain complementary information regarding phone categories, as supported by our ABX analysis presented in Figure 5.

2.8 Conclusions

This first experiment focused on the role of the articulatory modality in learning phonetic representations. The results show that the acoustic and articulatory modalities complement each other and can be jointly exploited for the discovery of phonetic units, in line with motor and perceptual-motor theories of speech perception. The challenge for the developing child is to learn to retrieve motor information from a purely auditory stimulus without being explicitly provided with the mapping between speech sounds and underlying articulatory gestures. It is this question of self-supervised learning of this mapping that we are attempting to address with the help of a simulation based on deep learning, and the development of a computational agent able to learn by vocal imitation, presented in Section 3.

3. Experiment 2: Computational Modeling of Speech Learning

Using computer simulation, we investigate here how a child can learn the complex and ill-posed relationship between speech units, speech sounds, and corresponding articulatory gestures. To that purpose, we proposed a computational agent learning to drive its virtual vocal apparatus by imitating auditory stimuli. We proposed two versions of this agent, respectively called the "continuous" and the "discrete" version. The former is similar to the one reported in our previous study (Georges et al. 2022). It is trained to minimize the discrepancies between perceived and repeated auditory stimuli in the spectral domain without explicitly decoding discrete (and potentially less variable) speech units. The discrete version includes two additional modules for learning such units from both the acoustic input and the inferred articulatory trajectories. Similarly to the simulations in Experiment 1 (see Section 2), these two modules are based on VQ-VAEs. It is important to note that, unlike the continuous version of the proposed agent, the discrete one evaluates the distance between perceived and produced auditory stimuli in the unit domain (rather than in the spectral one). In other words, the agent is here expected to explicitly decode and preserve a certain "linguistic content" encoded in a sequence of discrete units. For the two versions of the proposed agent, we also investigated specific mechanisms to constrain the acoustic-to-articulatory mapping, referred to here as inductive biases.

3.1 Design of the Computational Agent

The architecture of the proposed agent is presented in Figure 7. Its core modules and the proposed training algorithm are presented in the following sections.

3.1.1 Neural Articulatory Synthesizer. To avoid modeling the complex (and often speaker-dependent) physical, biomechanical, and acoustic processes underlying speech production, we use a deep neural network to model the relationship between the configuration of the vocal tract on the one hand and the spectral content of the resulting sound on the other, based on data recorded from a real speaker. To that purpose, we proposed a "neural articulatory synthesizer" composed of 2 main modules: a DNN-based articulatory-to-acoustic mapping ϕ and a neural vocoder.

Articulatory-to-Acoustic Mapping. In the proposed approach, a feedforward DNN maps a vector of 6-D articulatory features $\mathbf{a}$ to an 18-D vector of acoustic features (cepstral coefficients in Bark-scale) $\hat{\mathbf{s}} = \phi(\mathbf{a})$. The model used in the present study comprises 4 fully connected layers of 512 neurons each (with batch normalization and dropout after each layer). The hyperbolic tangent was used as activation function for the neurons of the hidden layers. This model was trained on the entire PB2007 dataset using back-propagation with Adam optimizer, on mini-batches of 32 observations and with the mean squared error (*MSE*) as loss function (20% of the training data were used for controlling early-stopping).

Waveform Generation. Acoustic features estimated from articulatory parameters, combined with source parameters (f0 and periodicity) directly extracted from the original signal, are finally fed into a neural vocoder to generate the speech waveform. In the present study, we used the LPCNet neural vocoder (Valin and Skoglund 2019). Starting from a pre-trained version trained on a large acoustic database, we fine-tuned (for 56 epochs) the model to the speaker's voice of the PB2007 dataset. Importantly, this neural

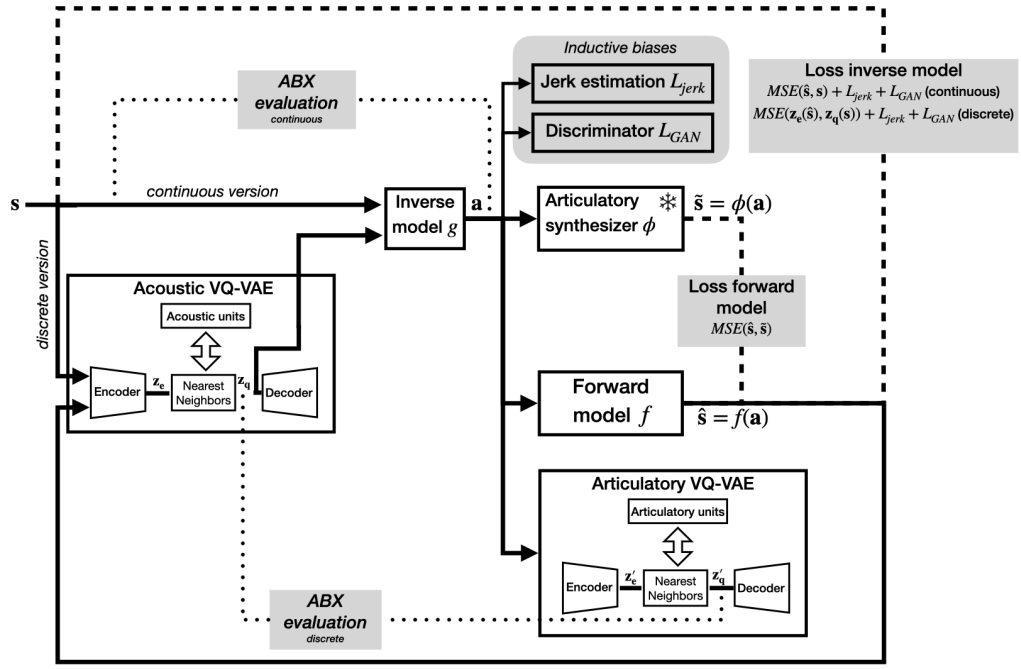

Figure 7
Overview of the proposed computational agent learning to drive an articulatory synthesizer by imitating auditory stimuli. Flows of information in solid lines, loss computations in big-dash lines, ABX measurements in small-dash lines. For the continuous version, audio input **s** is directly sent to the inverse model. For the discrete one, audio input **s** is first quantified from a (learned) codebook of representations $\mathbf{z_q}(\mathbf{s})$ using the acoustic VQ-VAE. Articulatory trajectories $\mathbf{a} = g(\mathbf{z_q}(\mathbf{s}))$ (or $\mathbf{a} = g(\mathbf{s})$ in the continuous case), inferred using the LSTM-based inverse model **g**, are sent both to the pre-trained neural articulatory synthesizer (ϕ, i.e., the plant, assumed to be non-differentiable) which provides the repetition of the input stimulus by the agent $\tilde{\mathbf{s}} = \phi(\mathbf{a})$, and to the forward model f which provides the "mental" simulation of the synthesis process $\hat{\mathbf{s}} = f(\mathbf{a})$. Both internal models (forward and inverse) are trained using the algorithm described in Section 3.1.4. $MSE(.)$ denotes the mean square error. For the discrete version of the agent, the acoustic VQ-VAE is trained only from the auditory input, while the articulatory VQ-VAE is trained only from the inferred articulatory trajectories. The two (optional) inductive biases aim at regularizing the acoustic-to-articulatory inverse mapping. They are respectively based on the minimization of the jerk of the inferred articulatory trajectories and on the use of an adversarial loss expected to prevent the inference of non-plausible vocal tract configurations. ABX tests assess the phonetic properties of auditory inputs and inferred articulatory trajectories, either before (continuous version) or after (discrete version) being quantized by the acoustic and the articulatory VQ-VAEs, respectively.

articulatory synthesizer is used here as a pre-trained module. When used within the computational agent for learning the inverse and forward models, described in the next sections, only the articulatory-to-acoustic mapping part is used, and its parameters are not updated. The waveform generation part is used only for listening tests.

3.1.2 Internal Models

Forward Model. The agent's forward internal model f predicts the acoustic consequences $\tilde{\mathbf{s}} = f(\mathbf{a})$ of the execution of a sequence of articulatory commands **a**. Similarly to the

neural articulatory synthesizer, it is implemented as a feed-forward DNN. However, unlike the neural articulatory synthesizer, kept frozen all along the learning phase, the forward model parameters are randomly initialized. Hence, the forward model contains no prior knowledge of the properties of the agent's vocal apparatus, and it must be trained by "listening" to the outputs of the agent's plant in order to learn to provide good estimates of the acoustic result of articulatory commands. We used the same hyperparameter tuning strategy already described in Section 2 to find the optimal architecture for the forward model, though the criterion here was based on the minimization of the *MSE* on the spectral distance between the synthesized and the forward model outputs. The optimal network turned out to be a 4-layer FF-DNN with 512 neurons in each layer.

Inverse Model. The inverse internal model g is used to estimate the sequence of articulatory feature vectors $\mathbf{a} = [\mathbf{a}_1, \cdots, \mathbf{a}_T]$ to be sent to the synthesizer in order to approximate the auditory input $\mathbf{s} = [\mathbf{s}_1, \cdots, \mathbf{s}_T]$. It is implemented as a unidirectional LSTM network with two recurrent layers and one final linear layer. Approximation is evaluated either directly in the continuous version or after discretization by the acoustic VQ-VAE in the discrete version. Hyperparameter tuning was based on the minimization of the corresponding *MSE* between the auditory input to imitate and the forward model output, directly in the spectral components in the continuous version or after discretization in the discrete version. We selected an architecture providing optimal or quasi-optimal performance in both the continuous and discrete versions of the agent, comprising 2 LSTM layers with a 32 hidden state dimensionality and a dropout parameter $p = 0.25$.

3.1.3 Inductive Biases. In order to better tackle the ill-posed nature of the acoustic-to-articulatory mapping and to guide the proposed agent to plausible articulatory trajectories, we investigated the use of specific mechanisms to constrain the learning process in our algorithms, referred to as inductive biases. We considered two mechanisms: The first one is inspired by optimal control theory and focuses on the dynamics of inferred articulatory features, while the second aims to keep these features within plausible ranges. These inductive biases are detailed in the following paragraphs.

Minimum Jerk. Optimal control deals with finding the control inputs of a system that optimize a certain performance criterion. In order to guide the agent to smooth trajectories, we use the minimization of the jerk as an additional criterion to optimize during the learning phase. For each articulatory feature, the agent is encouraged to find a minimum jerk trajectory from the beginning to the end of the sentence. This inductive bias can be introduced through the following additional loss, defined as the mean square jerk along the trajectory (as in Rajpal and Patil 2016) expressed as:

$$L_{jerk} = \frac{1}{T}\int_0^T \left(\frac{d^3\mathbf{a}_t}{dt^3}\right)^2 dt \tag{4}$$

GAN-based Geometrical Constraint. In the absence of any constraints, the inverse model can generate articulatory configurations that do not reflect those adopted by our speaker used to train the neural articulatory synthesizer. For instance, some articulators' positions can be outside the limits of the EMA data, or two articulators can result in

unrealistic configurations (e.g., the lower lip above the upper lip). To alleviate this issue, we condition the generated articulatory configurations using a generative adversarial network (GAN)-based approach. To do so, we introduce a discriminator network that is trained to discriminate between generated and real articulatory configurations by maximizing the following loss:

$$L_D = \frac{1}{T}\sum_{t=1}^{T}[\log(D(\hat{\mathbf{a}}_t)) + \log(1 - D(\mathbf{a}_t))] \tag{5}$$

where $\hat{\mathbf{a}}_t$ is a randomly sampled real articulatory frame, $\mathbf{a}_t$ is a generated articulatory frame, and $D(\hat{\mathbf{a}}_t)$ (resp., $D(\mathbf{a}_t)$) is the probability returned by the discriminator for the generated (resp., real) articulatory frame.

Conversely, the inverse model is trained to fool the discriminator by minimizing:

$$L_{GAN} = \frac{1}{T}\sum_{t=1}^{T}\log(1 - D(\mathbf{a}_t)) \tag{6}$$

3.1.4 Learning Algorithm. The objective pursued by this agent during its learning process is threefold. It must (1) build a repertoire of discrete units representing speech compactly, (2) learn to recognize these units in the sounds it perceives, and (3) learn to control its plant (i.e., the neural articulatory synthesizer) so as to produce sounds containing as much as possible the same units as those perceived. Starting from a random initialization, the multi-task learning procedure used to jointly train the VQ-VAE, and both the forward and inverse models is presented in Algorithm 1.

Importantly, freezing the weights of the forward model in step 13 ensures that the acoustic error $MSE(\widehat{\mathbf{s}}, \mathbf{s})$ (in the continuous version) or $MSE(\mathbf{z_e}(\widehat{\mathbf{s}}), \mathbf{z_q}(\mathbf{s}))$ (in the discrete version) between the target acoustic input and the output of the inverse-forward model chain generates a signal error only in the inverse model, in charge for adequate control of the forward model. In practice, we train the proposed agent with batch size $B = 8$. Early stopping was used to prevent overfitting and stop the training. The inverse and forward internal models were trained for at least 50 epochs and the patience parameter was set to 25 epochs. For the discrete version of the agent, it proved necessary to pre-train the acoustic VQ-VAE before the inverse model, as the latter's loss function is calculated from representations of the learned units, which must therefore be more or less already stabilized. The acoustic VQ-VAE is trained only from the acoustic input s. The parameters λ_j and λ_s used to weight the influence of static and dynamic inductive biases were set to 0.05 after some preliminary tests on the validation set. However, the search for the most efficient weighting strategy proved to be a difficult task that should be further explored in future work. Finally, in the discrete version of the agent, an articulatory VQ-VAE, extracting speech units from the inferred articulatory trajectories, is trained separately (with the same early-stopping criterion as for the acoustic VQ-VAE). Hence, at this stage of our work, the articulatory VQ-VAE is not involved in the perception-production loop *per se*, it just enables assessment of the articulatory content of the discrete agent at the end of the simulation.

3.1.5 Experiments. In the proposed series of simulations, the agent received auditory stimuli extracted from the test partition of the PB2007 dataset, which is the dataset used to train the neural synthesizer (i.e., the "voice" of the agent). Admittedly, this

Algorithm 1 Learning algorithm of the proposed computational agent (continuous and discrete version)

Initialize: Random initialization of parameters of both the forward f and inverse model g, and the acoustic VQ-VAE (discrete version)
for each training iteration **do**
  Sample a mini-batch of B sequences of acoustic features $\{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_B\}$ from the dataset
  for $b = 1$ to B **do**
    Compute acoustic VQ-VAEs decoder input (discretized version of the encoder output) $\mathbf{z_q}(\mathbf{s}_{b,t\pm\tau})$ for each time step t, with τ the size of the context considered around the frame of interest (discrete version)
    Compute inverse model output: $\mathbf{a}_t = g(\mathbf{s}_{b,t\pm\tau})$ (continuous version) or $\mathbf{a}_t = g(\mathbf{z_q}(\mathbf{s}_{b,t\pm\tau}))$ (discrete version) for each time step t
    Compute neural articulatory synthesizer output: $\widetilde{\mathbf{s}}_{b,t} = \phi(\mathbf{a}_{b,t})$ for each time step t
    Compute forward model output: $\widehat{\mathbf{s}}_{b,t} = f(\mathbf{a}_{b,t})$ for each time step t
    Compute the output of the acoustic VQ-VAE's encoder $\mathbf{z_e}(\widehat{\mathbf{s}}_{b,t})$ for each time step t (discrete version)
  end for
  Update the acoustic VQ-VAE using Equation (2) (discrete version)
  Update forward model f to minimize $L_{forward} = MSE(\widehat{\mathbf{s}}, \widetilde{\mathbf{s}})$
  Update inverse model g to minimize $L_{inverse} = MSE(\widehat{\mathbf{s}}, \mathbf{s}) + \lambda_j L_{jerk} + \lambda_g L_{GAN}$ (continuous version) or $L_{recons} = MSE(\mathbf{z_e}(\widehat{\mathbf{s}}), \mathbf{z_q}(\mathbf{s})) + \lambda_j L_{jerk} + \lambda_g L_{GAN}$ (discrete version) while keeping the forward model unchanged.
end for

experimental setup is not realistic since a child at this early stage of life has never received perfectly articulated speech from their own voice. However, this setup can provide a topline since it alleviates the problem of speaker normalization, that is, the ability to decode phonologically identical utterances uttered by different speakers despite large acoustic variability. An experiment with several speakers is left open for future work (this point is briefly discussed in the conclusive section 4).

For each simulation, the data sets were randomly distributed with 80% of the data used for training and the remaining 20% for testing; 20% of the training set was used as a validation set (to control early stopping). The data were centered and reduced each time, using the mean and variance calculated on the training set. We trained 5 agents, each time with a different random partitioning of the data sets. Performance measures reported here consist of the average performance over these 5 evaluations.

3.2 Metrics

3.2.1 ABX Evaluation of the Acoustic and Articulatory Content of the Agent's Productions. First, as in Experiment 1, we assessed the phonetic discriminability power of the acoustic and articulatory content of the agent's productions using the ABX methodology described in Section 2.6.1. In the continuous version, evaluation was done directly by evaluating the ABX score from the raw audio input and the inferred articulatory data (i.e., from **s** and **a**). In the discrete version enabling the quantification of the effect of the discretization done by the VQ-VAEs, ABX scores were computed on the representations

learned by the acoustic and articulatory VQ-VAEs (i.e., $\mathbf{z_q}(\mathbf{s})$ and $\mathbf{z'_q}(\mathbf{a})$). ABX scores computed on the agent's productions were compared to ground truth data, that is, ABX scores computed on the acoustic and articulatory content of the PB2007 dataset. For the discrete version, this corresponds to the data already presented in the previous section (Figure 5), called "quantized gold spectral" and "quantized gold articulatory" in the following. For the continuous version, this provides new ABX scores directly computed from the raw acoustic and articulatory data in PB2007, called "gold spectral" and "gold articulatory" in the following.

3.2.2 Perceptual Evaluation of the Agent's Productions. Second, we tested the quality of the phonemic content of the speech signal produced by the proposed agent (discrete version only) using a subjective evaluation by human listeners. For this aim, we extracted from the test corpus a set of isolated oral vowels in {i, y, u, e, ø, o, a} and a set of non-nasal consonants in {b, d, g, p, t, k, f, v, s, z, ʃ, ʒ} in a VCV context, with V one of the three vowels {i, a, u} (since the PB2007 dataset does not contain information about the velum). For each item, we synthesized 5 audio stimuli using LPCNet by combining the original source parameters (f0 and periodicity feature) with different sets of articulatory features: (1) ground truth from the PB2007 dataset (in order to evaluate the neural articulatory synthesizer independently from the agent), (2) the (discrete) agent's predictions when considering no inductive bias, (3) same but with the static bias (GAN-based), (4) same but with the dynamic bias (Jerk minimization), and (5) same but with both biases. These sets, comprising a total of 195 synthesized stimuli, were evaluated by 30 participants who self-identified as native French speakers. The stimuli were presented in an online test through the Prolific platform. The full set of isolated-vowel (respectively, VCV) stimuli was presented, and participant had to listen to each stimulus and to indicate the perceived vowel category (forced-choice within all possible vowel categories, hence 7 possibilities), respectively, consonant category (possibility for the participant to indicate the absence of the central consonant in a VCV, hence altogether 13 possibilities). Participants indicated their choices by clicking on a series of buttons associated with each vowel (respectively, consonant). The presentation order was randomized across subjects and conditions.

3.3 Results

3.3.1 Qualitative Description of the Learned Articulatory Trajectories. As a qualitative evaluation of the effect of the two proposed inductive biases, we present in Figure 8 the articulatory trajectories predicted by the inverse model of the proposed agent (discrete version) for the sounds [aga] and [aba]. First, without any constraints ($\lambda_j = 0$ and $\lambda_g = 0$; i.e., no inductive bias, orange dotted curve), the inferred articulatory parameters seem to oscillate from one timestep to the next, and sometimes achieve very high or low values. To some extent, the two inductive biases address these issues and complement each other. The dashed green curve ($\lambda_j = 0$ and $\lambda_g = 0.05$), partly overlaying the dotted orange curve, illustrates how the static inductive bias (GAN-based) functions as a gating mechanism. It prevents the inverse model from inferring out-of-domain targets, such as implausible vocal tract configurations (e.g., parts of the tongue above the palate). The dynamic bias, focused on minimizing jerk, smoothens trajectories to enhance compatibility with real ones, albeit at times excessively. When combined (purple curve), the two inductive biases result in more accurate articulatory dynamics, particularly for the tongue tip and tongue body. Nevertheless, the inferred trajectories occasionally deviate from the true ones. Strikingly, we observe, for example, that while the initial tongue dorsum upward movement for closing the vocal tract in [g] is correctly captured, the

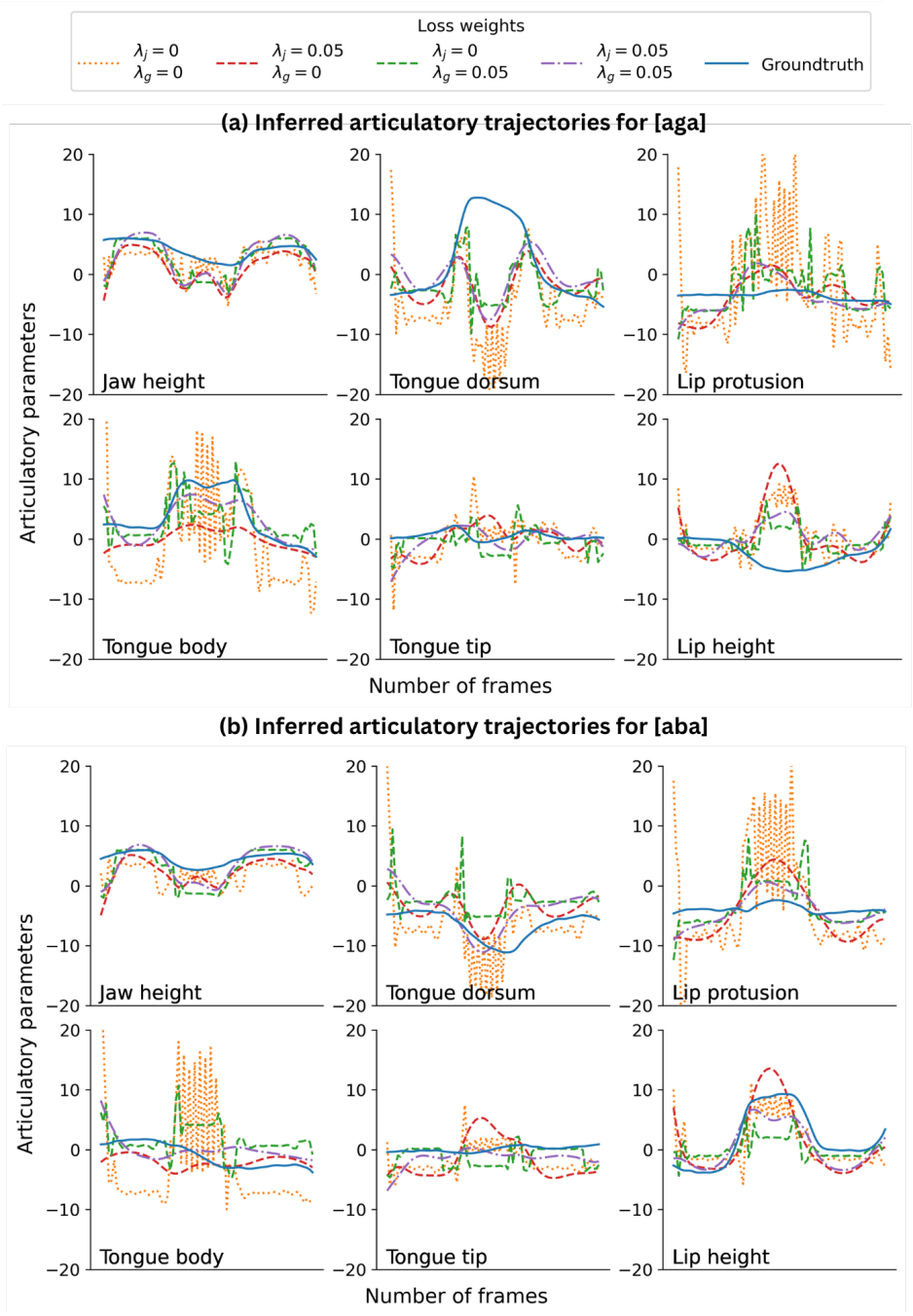

Figure 8
Example of articulatory trajectories inferred by the agent (discrete version), with no inductive bias ($\lambda_g = 0$, $\lambda_j = 0$, orange), with the static one only ($\lambda_g = 0.05$, $\lambda_j = 0$, red), with the dynamic one only ($\lambda_g = 0$, $\lambda_j = 0.05$, green), or with both ($\lambda_g = 0.05$, $\lambda_j = 0.05$, purple), for the sounds [aga] (a) and [aba] (b).

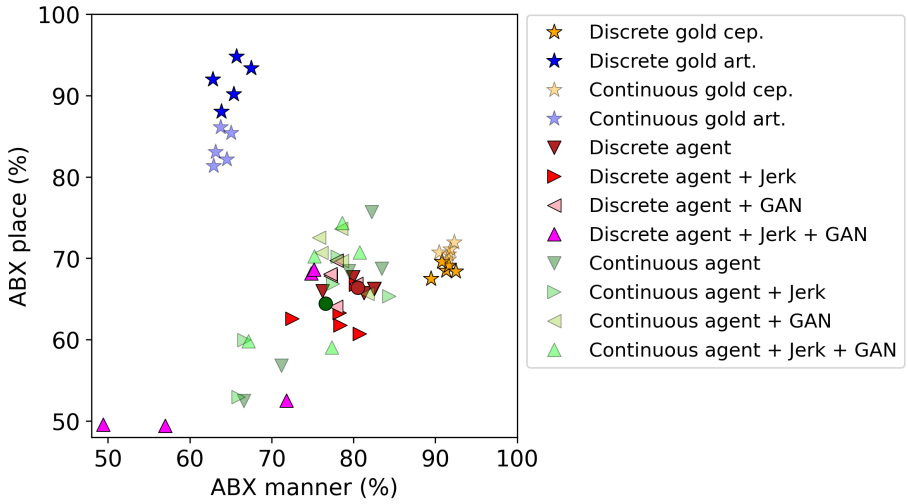

Figure 9
ABX scores with respect to place and manner of articulation computed on gold cepstral coefficients (yellow stars), gold articulatory trajectories (blue stars), or articulatory trajectories inferred by the discrete (red triangles) or the continuous agent (green triangles). Transparency (resp. non-transparency) indicates discrete (resp. continuous) features. Shades of red and green indicate the presence or absence of inductive biases. The red (resp. green) circle indicates the average score obtained by the discrete (resp. continuous) agent without biases. Each dot represents an independent training/validation/test split of the PB2007 dataset.

movement is suddenly totally modified during the plosive closure, the dorsum coming down and the closure being actually realized by a raising movement of the lips. This is likely because the model has found that the sound of plosive closure was rather similar, whatever the plosive, and hence it has decided to realize all closures with the same labial movement, whatever the plosive. This is indeed coherent with the pattern for the sound [aba] also displayed on Figure 8, for which the same global pattern among the various bias conditions emerges, but here with a clear labial movement rather coherent with ground truth patterns. All these tendencies are also observed for the continuous version of the proposed agent.

3.3.2 Quantitative Analysis of the Simulation Results. In Figure 9 we display the ABX scores for place and manner of articulation provided by the various types of agents (i.e., discrete vs. continuous and with or without the two biases) compared with gold scores from ground truth data. The average and standard deviation values computed over the 5 variants of each condition are provided in Table 2. The following trends emerge from this pattern of simulation results:

1. For articulatory ground truth data, discretization improves the ABX score for place of articulation, increasing from 83.9% to 91.7% on average (Table 2). On the contrary, discretization slightly degrades the ABX score for manner on acoustic ground truth data.
2. All simulations, regardless of the agent and the presence of inductive biases, provide articulatory outputs that fall far apart from ground truth articulatory data in terms of ABX place of articulation scores (Figure 9).

Table 2
Global, place, and manner ABX scores computed on gold cepstral coefficients, gold articulatory trajectories, or articulatory trajectories inferred by the discrete or continuous agent, with or without our two biases. Standard deviations are computed on 5 training/validation/test splits of the PB2007 dataset.

Features	ABX global (%)	ABX place (%)	ABX manner (%)
Discrete gold cep.	89.8 ± 0.9	68.8 ± 1.1	91.1 ± 1.1
Continuous gold cep.	88.0 ± 0.9	71.0 ± 0.3	91.3 ± 0.8
Discrete gold art.	84.0 ± 2.3	91.7 ± 3.2	65.0 ± 1.9
Continuous gold art.	79.4 ± 1.0	83.9 ± 1.8	63.7 ± 0.3
Discrete agent	80.2 ± 1.0	66.4 ± 0.8	80.5 ± 2.6
+ Jerk	77.7 ± 1.9	63.0 ± 2.3	78.0 ± 3.3
+ GAN	79.6 ± 0.4	67.3 ± 2.1	78.1 ± 1.3
+ Jerk + GAN	67.3 ± 12.2	57.7 ± 9.9	65.6 ± 11.7
Continuous agent	77.4 ± 9.7	64.4 ± 9.5	76.6 ± 7.3
+ Jerk	76.1 ± 8.3	63.0 ± 6.8	74.5 ± 8.0
+ GAN	82.0 ± 1.4	70.5 ± 3.1	78.1 ± 2.4
+ Jerk + GAN	79.1 ± 4.6	66.8 ± 7.0	75.8 ± 5.2

Overall, we do not observe a higher ABX score for the predicted articulatory data than for the acoustic data, as was the case in humans (Experiment 2).

3. Discretization slightly improves the agents' ABX scores, though more on manner (4% gain for agents with no bias) than on place (2% gain for agents with no bias; see Table 2). It also decreases the variability of simulation outputs (see standard deviations in Table 2) and hence stabilizes simulated articulatory trajectories.

4. Inductive biases can lead to increase but also decrease in ABX scores (Figure 9). It appears that for the continuous version the main gain is obtained with the GAN bias, particularly for place (6% gain), while for the discrete version the biases decrease performance. Interestingly, the GAN bias also decreases the variability of simulation outputs, particularly for the continuous version (see standard deviations in Table 2).

5. The variability of simulation results, either due to conditions (versions and biases) or to fluctuations in performance (among the 5 occurrences of each type of agent) may lead to ABX scores for place that are higher than the ABX place score for the gold acoustic condition. This suggests that some complementarity may be found in the agents' outputs, even though globally, their performance remains far from the gold articulatory score, as mentioned in the first bullet point.

3.3.3 Intelligibility of the Agent's Synthetic Speech. The results of the perceptual evaluation are presented in Figure 10.[2] The performance of the neural articulatory synthesizer can

2 Audio samples can be found at `https://georges.ma/publications/agent/`.

be considered as a topline for the proposed agent (since the speech stimuli are here synthesized from ground truth articulatory data). The average correct score is 90% for isolated vowels and 55% for the consonants in VCV sequences (which are more complicated to synthesize). Confusions such as "d vs. z" and "d vs. n" can respectively be partly due to lack of information on the tongue tip (the EMA coil used to track the tip of the tongue is glued 1 cm behind the actual tip) or on the velum. Although this result may seem relatively low, it is important to note that participants had the option to indicate that they were unable to identify any consonants in the VCV sequence. Some may have exercised this option to express that a particular consonant was slightly less accurately reproduced than the others. Analysis of the agent's speech productions shows that, as with the articulatory synthesizer, overall performance on vowels is much better than on consonants. A positive relative difference is observed for the vowels (e.g., 90% for the synthesizer vs. 95% for the agent with both the static and dynamic biases). In contrast, a negative one is observed for the consonants (55% vs. 49%). As a result, the agent's overall performance remains close to that of the synthesizer, which may act as a bottleneck in the proposed computational model. Moreover, we do not observe any significant effect of inductive biases on the agent's performance, which therefore only seem to impact the plausibility of the articulatory trajectories. Several confusions are made between groups of phonemes sharing the same manner of articulation (e.g., "ʒ vs. z", "s vs. g"). This further indicates that the agent has yet to completely solve the consonant invariance problem mentioned in the Introduction.

4. Discussion

Let us discuss the results of Experiments 1 and 2 in relation with the questions introduced in Section 1.

The first question concerned the potential role of articulatory knowledge in the discovery of phonemes. Two major aspects emerge from the analysis of acoustic and articulatory data of all the studied speakers in Experiment 1. Firstly, there is an interesting complementarity concerning the representation of consonants, regardless of the vocalic context. The quantized representations of the acoustic VQ-VAE better encode the manner of articulation while those of the articulatory VQ-VAE better encode the place of articulation, as assessed by the ABX testing method. The amount of complementarity is rather large, with a difference between modalities (acoustic vs. articulatory) of around 10% for the English speakers and up to 30% for the French one. It is important to notice that the articulatory data used in the present study is only a partial description of the vocal tract shape, with only few flesh point positions tracked by the EMA sensors. It does not include aerodynamic information on the tongue position with respect to the palate, nor muscular information on the state of the orofacial muscles. It also lacks information on the glottal source. All these additional sources of information could actually play a role in the representation of phonological categories and possibly improve ABX scores. Interestingly, it appears that introducing a discretization stage enhances the complementarity between the acoustic and the articulatory modalities (e.g., for speaker PB2007, the ABX place score for the articulatory data and for the discrete case is higher than that obtained for the continuous case; see Figure 9). This suggests that discretization contributes to decrease articulatory variability in spite of variations of the vocalic context. A possible complementarity between articulatory and acoustic representations of consonantal features comes in resonance with the old but still vivid debate between auditory (e.g., Diehl, Lotto, and Holt 2004; Kluender 1994) vs.

articulatory/motor invariance (e.g., Liberman and Mattingly 1985; Galantucci, Fowler, and Turvey 2006). In fact, the present data seem to support the nuanced view that both acoustic and articulatory features should play a role in the representation of phonological units, as underpinned by perceptuo-motor theories such as the Perception-for-Action-Control Theory (PACT, Schwartz et al. 2012). The second result in Section 2 concerns the dynamics of speech representations. It appears that articulatory trajectories are more stable than acoustic ones, with less variations in position among the available VQ-VAE units (see Figure 6). This result may be of interest for a computational system attempting to learn phonological categories from underlying sensory or motor trajectories. It is actually not completely surprising, since it is well known that the articulatory trajectories are smoother than their acoustic counterparts because of strong nonlinearities in the articulatory-to-acoustic mapping (Stevens 2000).

The second question raised in the Introduction concerned the joint learnability of acoustic trajectories, speech gestures, and speech units from raw acoustic data. The results presented in Section 3 provide here a much more mixed set of answers. Indeed, the simulations with the architecture presented in Figure 7 appear rather successful in their ability to learn to control a given articulatory synthesizer (see Figure 10), but only for auditory inputs extracted from the same dataset as the one used to train the neural synthesizer (i.e., "the agent listens to its own voice"). Preliminary tests (not reported here) showed that our model was not yet able to correctly address the problem of speaker normalization. This problem is actually quite classical and the proposed inductive biases were a priori not sufficient to deal with it adequately. Indeed, a number of studies have proposed deep architectures able to disentangle the linguistic content from other attributes such as the speaker identity (e.g., Benaroya, Obin, and Roebel 2023), and

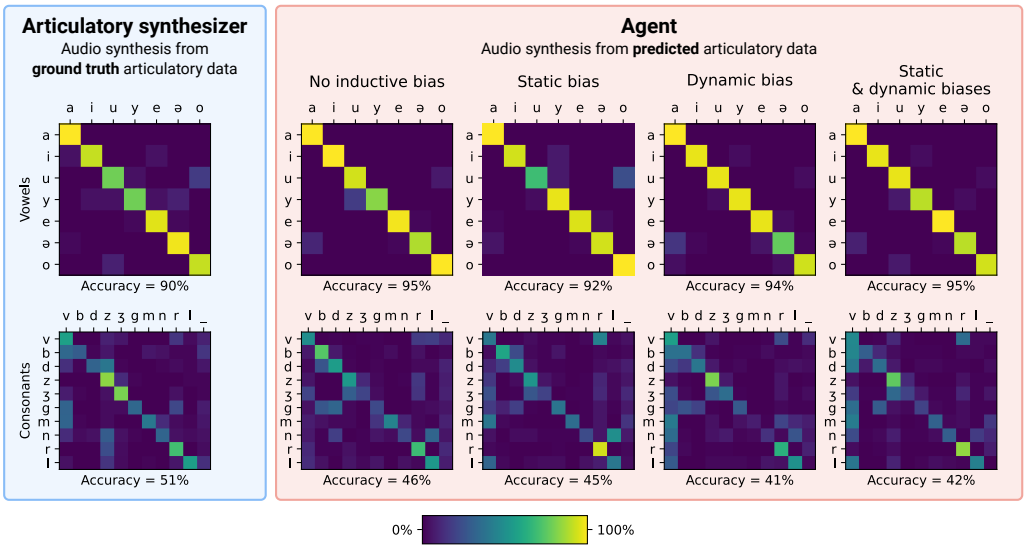

Figure 10
Perceptual evaluation of the acoustic output of the neural articulatory synthesizer (blue) and of the proposed agent (discrete version, pink), for both isolated vowels and central consonants in VCV sequences, with or without constraining the acoustic-to-articulatory inverse mapping with the static and dynamic biases.

a possible extension of this work would consist in introducing such techniques to learn speaker-independent acoustic representations. But more importantly, it appears that the articulatory trajectories inferred by the agent, whether in its continuous or discrete version, are not satisfying in the sense that they do not allow us to recover the acoustic/articulatory complementarity observed in our first experiment (i.e., from ground truth acoustic/articulatory data). This behavior is observed even in the presence of both static (GAN-based geometrical constraint) and dynamic (Jerk minimization) inductive biases. These biases improve the inferred trajectories by eliminating implausible vocal tract configurations and adding smoothness, but do not provide articulatory trajectories with the same level of complementarity with their acoustic counterparts, as seen in the first experiment. Independently on these rather disappointing results, the evaluation methodology we propose in the present work could benefit to other computational developments in the field. Indeed, a number of neural models of speech imitation and speech learning emerge in the literature (e.g., Beguš et al. 2023; Krug et al. 2023a, b; Siriwardena, Espy-Wilson, and Shamma 2022), sometimes accompanied by qualitative evaluation of the underlying articulatory dynamics. Still, a systematic evaluation of articulatory information based on ABX place and manner scores, in various kinds of categorization tasks, could interestingly complement the quantitative evaluation of these computational models.

A major perspective for this work should in our view involve better integration of the developmental schedule in the learning process at work in the model. Barnaud et al. (2019) showed that the endogenous babbling process specified by the so-called Frame-Content Theory introduced by MacNeilage (1998) could provide a solution to this problem. Indeed, this theory proposes that infants enter orofacial babbling by a set of "vertical" gestures associating the jaw with specific single articulators (tongue tip, tongue dorsum, lips) without moving the tongue from the consonant to the next vowel. This could provide a bootstrapping mechanism in the discovery of articulatory invariance, thanks to which infants would then favor such "vertical" gestures for speech motor control, likely to focus articulatory gestures towards invariant commands for consonants. A next stage for our work will hence consist in assessing whether adding this kind of babbling scenario to our computational agent could improve the self-supervised learning of the acoustic-to-articulatory inverse model.

References

Atal, Bishnu S., Jih Jie Chang, Max V. Mathews, and John W. Tukey. 1978. Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique. *The Journal of the Acoustical Society of America*, 63(5):1535–1555. https://doi.org/10.1121/1.381848, PubMed: 690333

Badin, Pierre, Gérard Bailly, Atef Ben Youssef, Frédéric Elisei, Christophe Savariaux, and Thomas Hueber. 2022. *PB2007 French Acoustic-articulatory Speech Database*. Zenodo, 6390598.

Bailly, Gérard. 1997. Learning to speak. Sensori-motor control of speech movements. *Speech Communication*, 22(2–3):251–267. https://doi.org/10.1016/S0167-6393(97)00025-3

Barnaud, Marie-Lou, Jean-Luc Schwartz, Pierre Bessière, and Julien Diard. 2019. Computer simulations of coupled idiosyncrasies in speech perception and speech production with COSMO, a perceptuo-motor Bayesian model of speech communication. *PloS ONE*, 14(1):e0210302. https://doi.org/10.1371/journal.pone.0210302, PubMed: 30633745

Beguš, Gašper, Alan Zhou, Peter Wu, and Gopala K. Anumanchipalli. 2023. Articulation GAN: Unsupervised modeling of articulatory learning. In *Proceedings of ICASSP*, pages 1–5.

https://doi.org/10.1109/ICASSP49357.2023.10096800
Benaroya, Laurent, Nicolas Obin, and Axel Roebel. 2023. Manipulating voice attributes by adversarial learning of structured disentangled representations. *Entropy*, 25(2):375. https://doi.org/10.3390/e25020375, PubMed: 36832741
Bergstra, James, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. 2011. Algorithms for hyperparameter optimization. *Advances in Neural Information Processing Systems*, 24, 9 pages.
Bergstra, James, Daniel Yamins, and David Cox. 2013. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In *International Conference on Machine Learning*, pages 115–123.
Bever, Thomas G. and David Poeppel. 2010. Analysis by synthesis: A (re-) emerging program of research for language and vision. *Biolinguistics*, 4(2–3):174–200. https://doi.org/10.5964/bioling.8783
Bybee, Joan. 1999. Usage-based phonology. *Functionalism and Formalism in Linguistics*, 1:211–242. https://doi.org/10.1075/slcs.41.12byb
Chen, Taijing, Adam Lammert, and Benjamin Parrell. 2021. Modeling sensorimotor adaptation in speech through alterations to forward and inverse models. *Proceedings of Interspeech*, pages 3201–3205. https://doi.org/10.21437/Interspeech.2021-1746
Coen, Michael H. 2006. Self-supervised acquisition of vowels in American English. In *AAAI*, volume 6, pages 1451–1456.
Cruz Blandón, María Andrea, Alejandrina Cristia, and Okko Räsänen. 2023. Introducing meta-analysis in the evaluation of computational models of infant language development. *Cognitive Science*, 47(7):e13307. https://doi.org/10.1111/cogs.13307, PubMed: 37395673
Diehl, Randy L., Andrew J. Lotto, and Lori L. Holt. 2004. Speech perception. *Annual Review of Psychology*, 55:149–179. https://doi.org/10.1146/annurev.psych.55.090902.142028, PubMed: 14744213
Dunbar, Ewan, Robin Algayres, Julian Karadayi, et al. 2019. The Zero Resource Speech Challenge 2019: TTS Without T. In *Proceedings of Interspeech*, pages 1088–1092. https://doi.org/10.21437/Interspeech.2019-2904
Dupoux, Emmanuel. 2018. Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner. *Cognition*, 173:43–59. https://doi.org/10.1016/j.cognition.2017.11.008, PubMed: 29324240
Galantucci, Bruno, Carol A. Fowler, and Michael T. Turvey. 2006. The motor theory of speech perception reviewed. *Psychonomic Bulletin & Review*, 13(3):361–377. https://doi.org/10.3758/BF03193857, PubMed: 17048719
Georges, Marc Antoine, Julien Diard, Laurent Girin, Jean-Luc Schwartz, and Thomas Hueber. 2022. Repeat after me: Self-supervised learning of acoustic-to-articulatory mapping by vocal imitation. In *Proceedings of ICASSP*, pages 8252–8256. https://doi.org/10.1109/ICASSP43922.2022.9747804
Georges, Marc-Antoine, Jean-Luc Schwartz, and Thomas Hueber. 2022. Self-supervised speech unit discovery from articulatory and acoustic features using VQ-VAE. In *Proceedings of Interspeech*, pages 774–778. https://doi.org/10.21437/Interspeech.2022-10876
Guenther, Frank H. 1995. Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production. *Psychological Review*, 102(3):594. https://doi.org/10.1037/0033-295X.102.3.594, PubMed: 7624456
Houde, John F. and Srikantan S. Nagarajan. 2011. Speech production as state feedback control. *Frontiers in Human Neuroscience*, 5. https://doi.org/10.3389/fnhum.2011.00082, PubMed: 22046152
Howard, Ian S. and Piers Messum. 2014. Learning to pronounce first words in three languages: An investigation of caregiver and infant behavior using a computational model of an infant. *PLoS ONE*, 9(10):e110334. https://doi.org/10.1371/journal.pone.0110334, PubMed: 25333740
Hsu, Wei Ning, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460. https://doi.org/10.1109/TASLP.2021.3122291
Kawato, Mitsuo. 1999. Internal models for motor control and trajectory planning. *Current Opinions in Neurobiology*, 9(6):718–727. https://doi.org/10.1016

/S0959-4388(99)00028-8, PubMed: 10607637

Kingma, Diederik P. and Max Welling. 2014. Auto-Encoding Variational Bayes. In *Proceedings of ICLR*.

Kluender, Keith R. 1994. Speech perception as a tractable problem in cognitive science. In M. A. Gernsbacher, editor, *Handbook of Psycholinguistics*. Academic Press, pages 173–217.

Kröger, Bernd J., Jim Kannampuzha, and Emily Kaufmann. 2014. Associative learning and self-organization as basic principles for simulating speech acquisition, speech production, and speech perception. *EPJ Nonlinear Biomedical Physics*, 2(1):1–28. https://doi.org/10.1140/epjnbp15

Kröger, Bernd J., Jim Kannampuzha, and Christiane Neuschaefer-Rube. 2009. Towards a neurocomputational model of speech production and perception. *Speech Communication*, 51(9):793–809. https://doi.org/10.1016/j.specom.2008.08.002

Krug, Paul K., Peter Birkholz, Branislav Gerazov, Daniel R. Van Niekerk, Anqi Xu, and Yi Xu. 2023a. Self-supervised solution to the control problem of articulatory synthesis. In *Proceedings of Interspeech*, pages 4329–4333. https://doi.org/10.21437/Interspeech.2023-2173

Krug, Paul Konstantin, Peter Birkholz, Branislav Gerazov, Daniel Rudolph Van Niekerk, Anqi Xu, and Yi Xu. 2023b. Artificial vocal learning guided by phoneme recognition and visual information. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:1734–1744. https://doi.org/10.1109/TASLP.2023.3264454

Laurent, Raphaël, Marie-Lou Barnaud, Jean-Luc Schwartz, Pierre Bessière, and Julien Diard. 2017. The complementary roles of auditory and motor information evaluated in a Bayesian perceptuo-motor model of speech perception. *Psychological Review*, 124(5):572–602. https://doi.org/10.1037/rev0000069, PubMed: 28471206

Lavechin, Marvin, Maureen de Seyssel, Marianne Métais, Florian Metze, Abdelrahman Mohamed, Hervé Bredin, Emmanuel Dupoux, and Alejandrina Cristia. 2024. Modeling early phonetic acquisition from child-centered audio data. *Cognition*, 245:105734. https://doi.org/10.1016/j.cognition.2024.105734, PubMed: 38335906

Liberman, Alvin M., Franklin S. Cooper, Donald P. Shankweiler, and Michael Studdert-Kennedy. 1967. Perception of the speech code. *Psychological Review*, 74(6):431–461. https://doi.org/10.1037/h0020279, PubMed: 4170865

Liberman, Alvin M. and Ignatius G. Mattingly. 1985. The motor theory of speech perception revised. *Cognition*, 21(1):1–36. https://doi.org/10.1016/0010-0277(85)90021-6, PubMed: 4075760

Lindblom, Björn. 1990. On the communication process: Speaker-listener interaction and the development of speech. *Augmentative and Alternative Communication*, 6(4):220–230. https://doi.org/10.1080/07434619012331275504

MacNeilage, Peter F. 1998. The frame/content theory of evolution of speech production. *Behavioral and Brain Sciences*, 21(4):499–511. https://doi.org/10.1017/S0140525X98001265, PubMed: 10097020

Maeda, Shinji. 1990. 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, pages 131–149. https://doi.org/10.1007/978-94-009-2037-8_6

McAuliffe, Michael, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal Forced Aligner: Trainable text-speech alignment using Kaldi. In *Proceedings of Interspeech*, pages 498–502. https://doi.org/10.21437/Interspeech.2017-1386

Morita, Takashi and Hiroki Koda. 2020. Exploring TTS without T using biologically/psychologically motivated neural network modules (ZeroSpeech 2020). In *Proceedings of Interspeech*, pages 4856–4860. https://doi.org/10.21437/Interspeech.2020-3127

Moulin-Frier, Clément, Sao Mai Nguyen, and Pierre-Yves Oudeyer. 2014. Self-organization of early vocal development in infants and machines: The role of intrinsic motivation. *Frontiers in Psychology*, 4:1006. https://doi.org/10.3389/fpsyg.2013.01006, PubMed: 24474941

Murakami, Max, Bernd Kröger, Peter Birkholz, and Jochen Triesch. 2015. Seeing [u] aids vocal learning: Babbling and imitation of vowels using a 3D vocal tract model, reinforcement learning, and reservoir computing. In *Proceedings of*

ICDL-EpiRob, pages 208–213. https://doi.org/10.1109/DEVLRN.2015.7346142

Parrell, Benjamin, Vikram Ramanarayanan, Srikantan Nagarajan, and John Houde. 2019. The facts model of speech motor control: Fusing state estimation and task-based control. *PLoS Computational Biology*, 15(9):e1007321. https://doi.org/10.1371/journal.pcbi.1007321, PubMed: 31479444

Paszke, Adam, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An open source deep learning platform. https://pytorch.org/

Patri, Jean François, Julien Diard, and Pascal Perrier. 2015. Optimal speech motor control and token-to-token variability: A Bayesian modeling approach. *Biological Cybernetics*, 109(6):611–626. https://doi.org/10.1007/s00422-015-0664-4, PubMed: 26497359

Perrier, Pascal. 2012. Gesture planning integrating knowledge of the motor plant's dynamics: A literature review from motor control and speech motor control. In Susanne Fuchs, Mélanie Weirich, Pape Daniel, and Pascal Perrier, editors, *Speech Planning and Dynamics*, speech pro edition. Peter Lang Publishers, pages 191–238.

Philippsen, Anja. 2021. Goal-directed exploration for learning vowels and syllables: A computational model of speech acquisition. *KI-Künstliche Intelligenz*, 35(1):53–70. https://doi.org/10.1007/s13218-021-00704-y

Philippsen, Anja Kristina, René Felix Reinhart, and Britta Wrede. 2014. Learning how to speak: Imitation-based refinement of syllable production in an articulatory-acoustic model. In *Proceedings of ICDL-EpiRob*, pages 195–200. https://doi.org/10.1109/DEVLRN.2014.6982981

Port, Robert F. and Adam P. Leary. 2005. Against formal phonology. *Language*, 81(4):927–964. https://doi.org/10.1353/lan.2005.0195

Rajpal, Avni and Hemant A. Patil. 2016. Jerk Minimization for acoustic-to-articulatory inversion. In *Proceedings of Speech Synthesis Workshop*, pages 82–87. https://doi.org/10.21437/SSW.2016-14

Rasilo, Heikki and Okko Räsänen. 2017. An online model for vowel imitation learning. *Speech Communication*, 86:1–23. https://doi.org/10.1016/j.specom.2016.10.010

Rezende, Danilo Jimenez, Shakir Mohamed, and Daan Wierstra. 2014. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of ICML*, pages 1278–1286.

Schatz, Thomas, Naomi H. Feldman, Sharon Goldwater, Xuan-Nga Cao, and Emmanuel Dupoux. 2021. Early phonetic learning without phonetic categories: Insights from large-scale simulations on realistic input. *Proceedings of the National Academy of Sciences*, 118(7):e2001844118. https://doi.org/10.1073/pnas.2001844118, PubMed: 33510040

Schatz, Thomas, Vijayaditya Peddinti, Francis Bach, Aren Jansen, Hynek Hermansky, and Emmanuel Dupoux. 2013. Evaluating speech features with the minimal-pair ABX task: Analysis of the classical MFC/PLP pipeline. In *Proceedings of Interspeech*, pages 1–5. https://doi.org/10.21437/Interspeech.2013-441

Schwartz, Jean Luc, Anahita Basirat, Lucie Ménard, and Marc Sato. 2012. The perception-for-action-control theory (pact): A perceptuo-motor theory of speech perception. *Journal of Neurolinguistics*, 25(5):336–354. https://doi.org/10.1016/j.jneuroling.2009.12.004

Serrurier, Antoine, Pierre Badin, Anna Barney, Louis-Jean Boë, and Christophe Savariaux. 2012. The tongue in speech and feeding: Comparative articulatory modelling. *Journal of Phonetics*, 40(6):745–763. https://doi.org/10.1016/j.wocn.2012.08.001

Siriwardena, Yashish M., Carol Espy-Wilson, and Shihab Shamma. 2022. Learning to compute the articulatory representations of speech with the MIRRORNET. *arXiv preprint arXiv:2210.16454*. https://doi.org/10.21437/Interspeech.2023-562

Skipper, Jeremy I., Virginie Van Wassenhove, Howard C. Nusbaum, and Steven L. Small. 2007. Hearing lips and seeing voices: How cortical areas supporting speech production mediate audiovisual speech perception. *Cerebral Cortex*, 17(10):2387–2399. https://doi.org/10.1093/cercor/bhl147, PubMed: 17218482

Stevens, Kenneth N. 1989. On the quantal nature of speech. *Journal of Phonetics*, 17(1–2):3–45. https://doi.org/10.1016/S0095-4470(19)31520-7

Stevens, Kenneth N. 2000. *Acoustic Phonetics*, volume 30. MIT Press. https://doi.org/10.7551/mitpress/1072.001.0001

Tjandra, Andros, Sakriani Sakti, and Satoshi Nakamura. 2020. Transformer VQ-VAE for unsupervised unit discovery and speech synthesis: Zerospeech 2020 challenge. In *Proceedings of Interspeech*, pages 4851–4855. https://doi.org/10.21437/Interspeech.2020-3033

Valin, Jean Marc and Jan Skoglund. 2019. LPCNet: Improving neural speech synthesis through linear prediction. In *Proceedings of ICASSP*, pages 5891–5895. https://doi.org/10.1109/ICASSP.2019.8682804

Van Den Oord, Aaron, Oriol Vinyals, and Koray Kavukcuoglu. 2017. Neural discrete representation learning. *Proceedings of NIPS*, 30:6309–6318.

Wrench, Alan A. 2000. A multi-channel/multi-speaker articulatory database for continuous speech recognition research. *Phonus*, 5:1–13.