Modeling the initial state of early phonetic learning in infants

Maxime Poli
CoML, ENS/PSL/EHESS/CNRS, Paris, France

Thomas Schatz
Université Aix Marseille, CNRS, LIS, Marseille, France

Emmanuel Dupoux
CoML, ENS/PSL/EHESS/CNRS, Paris, France
Meta AI Research, Paris, France

Marvin Lavechin
GIPSA-lab, Université Grenoble-Alpes, Grenoble, France

Abstract: What are the necessary conditions to acquire language? Do infants rely on simple statistical mechanisms, or do they come pre-wired with innate capabilities allowing them to learn their native language(s)? Previous modeling studies have shown that unsupervised learning algorithms could reproduce some aspects of infant phonetic learning. Despite these successes, algorithms still fail to reproduce the learning trajectories observed in infants. Here, we advocate that this failure is partly due to a wrong initial state. Contrary to infants, unsupervised learning algorithms start with little to no prior knowledge of speech sounds. In this work, we propose a modeling approach to investigate the relative contribution of innate factors and language experience in infant speech perception. Our approach allows us to investigate theories hypothesizing a more significant role of innate factors, offering new modeling opportunities for studying infant language acquisition.

Keywords: phonetic learning; language acquisition; deep learning

Corresponding author: Maxime Poli, Centre Sciences des Données, ENS, 45 rue d'Ulm, Paris, France. Email: maxime.poli@ens.psl.eu

ORCID ID: https://orcid.org/0000-0002-9377-9150

Citation: Poli, M., Schatz, T., Dupoux, E. & Lavechin, M. (2025). Modeling the initial state of early phonetic learning in infants. *Language Development Research, 5*(1), 1-34. http://doi.org/10.34842/y89t-6q31

Introduction

The 'statistical learning hypothesis' posits that infants learn their native languages(s) by gradually collecting statistics over their language input (Saffran & Kirkham, 2018). This is strikingly similar to how current AI's Large Language Models (LLMs) learn: building a probabilistic model of sequences of words from the mere observation of these sequences as they occur in their language inputs[1]. How does learning in such models fare in comparison to learning in infants? First, LLMs typically learn from text, while preschool children learn from speech, which constitutes a richer, noisier, and more variable signal. Second, LLMs are trained on exceedingly large amounts of data. For instance, the recent model LLaMA was trained on 1.4T tokens, roughly 800B words (Touvron et al., 2023) while children hear only between 1M and 10M words per year (Gilkerson et al., 2017). At this rate, infants would need to live between 80,000 years and almost a million years to get the same amount of data. Therefore, current language models are outranked by children regarding robustness to input signal variability and data efficiency as already advocated in Lavechin et al. (2023) and Warstadt et al. (2023). One candidate explanation for the incredibly slow learning pace observed in LLMs is their lack of innate language capabilities. Indeed, LLMs have a relatively generic architecture that can be used to learn visual or musical patterns. In contrast, it has been claimed that language learning critically relies on evolution-supplied specialized structures unique to humans (Chomsky, 1957; Hauser et al., 2002).

Far from entering the complicated controversy about the role of innate knowledge in language and cognition, we focus in this paper on an apparently simple yet fundamental subcomponent of language: phonetics. The ability to encode the sounds of language in terms of a relatively invariant representation has been considered one of the first steps of language acquisition in infants. Quite surprisingly, preverbal infants have an excellent ability to discriminate between very subtle sound differences that sometimes escape adults. Contrary to English adult speakers, 10- to 12-month-old English-learning infants can distinguish [ʈ̪a] from [ʈa], which is contrastive in Hindi (Werker et al., 1981). Similarly, Japanese-learning infants can discriminate [ɹa] from [la] as in 'right' versus 'light' (Kuhl et al., 2006), while Japanese adult speakers struggle to hear the difference (Best & Strange, 1992; Yamada & Tohkura, 1992). It is only when infants grow older that their perception specializes to their native language(s) (Kuhl et al., 2006; McMurray et al., 2018).

The early capacities of infants to discriminate speech sounds highlight the *initial state* of their perceptual apparatus, whereas their developmental trajectories emphasize the role of *experience* (Eimas et al., 1971; Kuhl & Iverson, 1995; Kuhl et al., 2003; Maye et al., 2002; Werker & Curtin, 2005).

[1] More precisely, they learn by predicting the conditional probability of future linguistic units — words or sub-word tokens — based on past units.

In this study, we investigate the respective contribution of initial state abilities and language experience in infant speech perception with computational modeling[2]. Our approach involves pretraining computational models of early phonetic learning to induce initial state sound discrimination capabilities. We then observe how these induced capabilities affect the learning trajectories taken by the model. Our results show that models with strong initial state capabilities better fit the observed data in 6-8 and 10-12 month-old American English and Japanese-learning infants. Our methodology allows us to explore theories positing a greater contribution of initial state factors in infant language acquisition, a theoretical space that has been largely overlooked in computational modeling until now.

Theoretical views on early phonetic learning in infants

The relative contributions of initial abilities versus language experience in phonetic learning have been subject to much debate. Aslin and Pisoni (1980) have outlined three possible theories concerning the development of speech perception in infants – see Rowland (2013) for an overview of the different theories. Those are depicted in Figure 1.

The *universal theory* hypothesizes that infants come pre-equipped with general auditory mechanisms partially shared with other species. According to this theory, newborns could initially discriminate all possible speech sound contrasts. Through exposure to speech, only sensitivity to contrasts to which the child is exposed would persist (maintenance), while sensitivity to contrasts to which the child is not exposed would decline (loss) – see Aslin et al. (2002). There exist at least two observations incompatible with the universal theory. First, infants lose sensitivity for some non-native contrasts but not all of them – see Singh et al. (2022) and Tsuji and Cristia (2014) for meta-analytic evidence. Second, infants are born capable of discriminating many sound contrasts but not all of them – e.g., see Eilers and Minifie (1975) for an example where infants fail to discriminate between [s] as in 'sing' versus [θ] as in 'thing'.

The *attunement theory*, perhaps the prevailing theory nowadays, proposes that infants come pre-equipped with language-specific mechanisms that would enable them to roughly discriminate speech sounds, although not to the same extent as adults in terms of native speech sound discrimination (Kuhl, 2004; Werker & Curtin, 2005). The attunement theory places greater importance on the role of experience by stipulating that the language(s) infants are exposed to reorganize their perceptual abilities. Through exposure to speech, infants' sensitivity to some – mostly native – contrasts would increase (facilitation), while sensitivity to some other – mostly non-native – contrasts would decline (loss). According to this theory, there may be no change in perceptual

[2]Here, we take the initial state to be the state of the perceptual system at birth. Such a system can come about through a combination of evolutionary processes (the true 'innate' components) and prenatal learning in utero. We do not attempt to distinguish these two sources of initial state abilities.

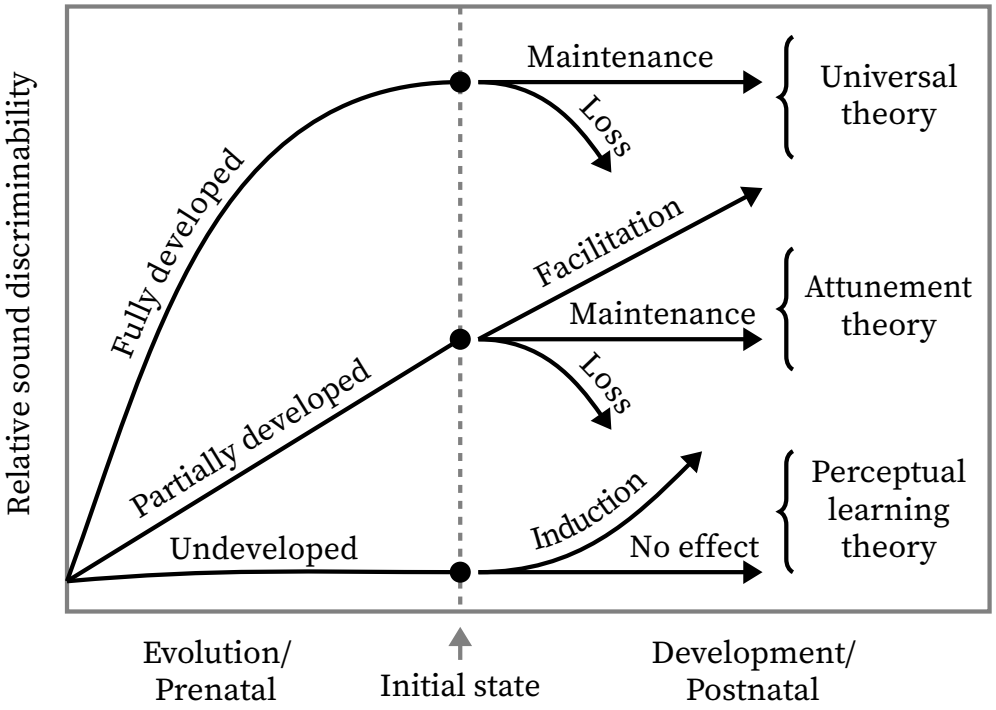

Figure 1. ***The possible effects of innate factors (Evolution/Prenatal) and language experience (Development/Postnatal) in infant speech sound perception. Adapted from Aslin and Pisoni (1980).***

abilities for some native or non-native contrasts (maintenance), which has been reported in many studies in 6- to 12-month-old infants (Best et al., 1995; Eilers & Minifie, 1975; Polka et al., 2001; Tsao et al., 2006) – see Best et al. (2016) for a review on the different speech sound discrimination trajectories observed in infants. Although there may be disagreement on the details of the implementation – e.g., the PRIMIR framework proposed by Werker and Curtin (2005) or the perceptual magnet theory proposed by Kuhl and Iverson (1995) and Kuhl et al. (2008) –, the attunement theory nicely accounts for the large array of developmental patterns observed in infants.

A major critique of both the attunement theory and the universal theory is that we may overestimate infants' capabilities to discriminate speech sounds for two reasons. First, it is common when working with infant participants to exclude those who fail to pay attention, cry, or fall asleep during the experiment. Nittrouer (2001) argues that infants may show uncooperative precisely because they cannot discriminate the stimuli presented. Consequently, excluding infants who fail to meet the criterion of the experimental procedure may result in inflated measures of discriminability. Indeed, testing 6- to 14-month-olds and 2- to 3-year-olds, Nittrouer (2001) found lower discriminability scores than typically reported in the literature – but see Aslin et al. (2002) for counterarguments. The second argument is that sound discrimination experiments use

simplified stimuli in the form of prototypical sounds and cherry-picked contrasts that fail to account for the large variability of spontaneous speech encountered by infants (Pierrehumbert, 2003). Under this view, sound discrimination capabilities measured in controlled laboratory settings would not reflect the actual capabilities of infants in real-world situations (Nittrouer, 2001; Pierrehumbert, 2003; Swingley, 2009).

This brings us to the *perceptual learning theory*, which proposes a scenario where experience plays a more important role. According to this theory, there would be no need to assume innate capabilities, and infants could build the sound system of their native language(s) in a bottom-up manner from sole exposure to speech. This theory seems plausible in light of the experiments attempting to isolate learning mechanisms infants may bring to the task. For instance, Maye et al. (2002) showed that it is possible to induce different discrimination patterns in 6- and 8-month-old infants. Infants exposed to a bimodal distribution of sounds along a [ta]-[da] continuum can discriminate [ta] from [da], while those exposed to a unimodal distribution drawn from the center of the continuum cannot. The perceptual learning theory is further supported by computational modeling studies showing that it is possible to reproduce some developmental patterns in speech perceptual learning using unsupervised learning models (Lavechin et al., 2022; Räsänen et al., 2016; Schatz et al., 2021; Steels & De Boer, 2008; Vallabha et al., 2007)[3].

Current work in modeling early phonetic learning

Computational modeling studies have always been central to the debate on the relative contribution of innate factors and experience, as they shed light on what can be learned from the input signal (Ambridge & Lieven, 2011; Bates et al., 1996; Joanisse & McClelland, 2015). After all, if a model successfully reproduces the observed data in infant perceptual learning of speech sounds, do we need to posit innate factors? Despite successes in reproducing some aspects of early phonetic learning as observed in infants (Antetomaso et al., 2017; Lavechin et al., 2022; Miyazawa et al., 2010; Räsänen, 2012; Schatz et al., 2021; Steels & De Boer, 2008; Vallabha et al., 2007), we argue that computational modeling studies have thus far failed to account for the large array of infant developmental trajectories depicted in Figure 1 and reviewed in Best et al. (2016).

Let us take the example of the American English [ɹ]-[l] contrast which has received the attention of both infant development and modeling experts. In a seminal study, Kuhl et al. (2006) showed that between 6 and 8 months, Japanese- and American English-

[3]Here, our goal was to provide an overview of the main arguments supporting or challenging the different views but note that most authors do not consider these three theories to be mutually exclusive. In other words, it is unlikely that a single theory explains the development of all speech contrasts. From our perspective, the debate is not about trying to establish a single definitive theory as the absolute truth but more about where the initial state fits on the nature versus nurture continuum (vertical dashed line of Figure 1) and how this initial state influences developmental outcomes.

learning infants are capable of discriminating [ɹ] from [l] with similar performance scores. However, when tested a few months later, these same infants show markedly different perceptual patterns. By 10-12 months, American English infants show an improvement (facilitation) in their ability to discriminate the [ɹ]-[l] contrast, while Japanese infants show a decline (loss). While the effect of language exposure (higher scores for the model for whom the contrast is native) has been reproduced in numerous computational modeling studies and across different pairs of languages – e.g., Lavechin et al. (2022), Li et al. (2020), Matusevych et al. (2023), and Schatz et al. (2021) –, a closer examination of the trajectories taken by the proposed algorithms reveals notable differences with the trajectories observed in infants.

Schatz et al. (2021) used an algorithm based on a mixture of Gaussians applied to mel-frequency cepstral coefficients (MFCCs) with their first- and second-order derivatives. Their results showed that the discrimination score obtained by the Japanese model on the [ɹ]-[l] contrast increases with the quantity of speech available in the training set. In other words, for this contrast, the algorithm follows the inductive trajectory depicted in Figure 1, contrary to the loss observed in infants according to previous studies (Kuhl et al., 2006; Tsushima et al., 1994)[4].

Another example using the same algorithm from Li et al. (2020) showed a slightly different trajectory. When trained on a single speaker, the algorithm exhibits an increase (induction) on the [ɹ]-[l] contrast followed by a decrease (loss), resulting in an inverted U-shaped trajectory which, to the best of our knowledge, has not been documented in infants. Intriguingly, the same U-shaped trajectory is observed on the [w]-[j] pair (as in 'wet' versus 'yet'), which is contrastive in Japanese, and for which current theories predict either a facilitation or maintenance trajectory. This performance loss on the [w]-[j] pair, when the algorithm is trained on a large quantity of speech produced by the same speaker, may indicate that the algorithm overfits that same speaker. Lavechin et al. (2022) report the discrimination accuracy obtained by a Contrastive Predictive Coding (CPC) algorithm trained on raw speech. Although no trajectory is reported for individual contrasts, the overall discrimination accuracy averaged across all English or French contrasts also follows an inductive trajectory.

Statistical learning models, irrespective of whether they operate on handcrafted features or raw speech, are inherently rooted in the perceptual learning theory. Essentially, they begin with limited prior knowledge of speech sounds, and their performance largely tends to exhibit improvement over time. Consequently, current models fail to reproduce the large array of developmental trajectories observed in infants.

[4]In Kuhl et al.'s (2006) study, the observed decline on the [ɹ]-[l] contrast for Japanese infants was not deemed significant, contrary to Tsushima et al. (1994), where a significant decline was noted. When taken together with studies in later childhood and adulthood (e.g., Miyawaki et al. (1975)), it appears reasonable to interpret the cumulative evidence as suggestive of a decline, though additional infant experiments would be advisable.

The present study

In this study, we seek to explore the respective contribution of initial state abilities and experience on the development of speech sound discrimination capabilities. By and large, existing models of early phonetic learning implement the perceptual learning theory, where the proposed model starts with undeveloped or minimally developed discrimination capabilities (first portion of the vertical dashed line in Figure 1). Our primary contribution involves introducing a novel approach, previously used in machine learning but not yet applied to phonetic learning modeling, which consists of inducing 'innate' speech sound discrimination capabilities by pretraining our model. By controlling the initial state, we can now build computational models of early phonetic learning that posit a greater role of innate factors compared to language experience and assess which of these models better aligns with observed data in infants.

To demonstrate the relevance of our approach in modeling early phonetic learning, we simulate the learning process of American English- and Japanese-learning infants using CPC, an algorithm that learns from raw speech in an unsupervised manner already proposed in Lavechin et al. (2022, 2024) and Nguyen et al. (2020) – see Matusevych et al. (2023) for a comparison of different models. To induce 'innate' speech sound discrimination capabilities and propose models more aligned with the attunement or universal theories, we pretrain models on ambient sounds in Experiment 1, and on multilingual speech in Experiment 2. Following Schatz et al. (2021), we evaluate the model's capability to discriminate American English and Japanese contrasts using the machine ABX sound discrimination task and test whether the simulated learning trajectories align with the observed data in infants. In particular, we focus on the [ɹ]-[l] pair, which is contrastive in English but not in Japanese and for which existing data indicate a facilitation effect over the first year of life for American English-learning infants and a loss effect for Japanese-learning infants (Kuhl et al., 2006; Tsushima et al., 1994). We also analyze the performance obtained on the [w]-[j] control pair (as in 'well' versus 'yell'), contrastive in both languages, for which prevailing theories predict either a maintenance or facilitation effect over the first year of life for both American English- and Japanese-learning infants. Although fewer observations are available on the [w]-[j] contrast, see Tsushima et al. (1994) whose results are compatible with a maintenance or facilitation trajectory in Japanese-learning infants.

Experiment 1: inducing initial speech sound discrimination capabilities through pretraining on ambient sounds

In this first experiment, we ask whether it is possible to induce 'innate' speech sound discrimination capabilities in our model and how the resulting initial state affects its developmental trajectory. Following Lavechin et al. (2024), we chose a learning algorithm relying on auditory predictive coding at the core of the predictive brain hypothesis that has gained attention in the neuroscience community (Huang & Rao,

2011; Hueber et al., 2020). The algorithm learns by predicting future representations of audio based on present and past ones (see Methods).

We consider two types of models. One model starts from random initialization, which is akin to assuming little initial discrimination capabilities except those brought by the architecture which has been optimized to process human speech (see Rivière et al., 2020) and corresponds to how computational models of early phonetic learning are typically trained (e.g., see Lavechin et al., 2024; Matusevych et al., 2023; Schatz et al., 2021). This is our *no-pretraining* condition, which aligns with the perceptual learning theory. The other model follows a pre-exposure or evolutionary phase during which it is pretrained on ambient sounds (e.g., animal vocalizations, vehicles, raindrops) yielding an initial state optimized to process ambient sounds. We predicted that such a pretrained model would learn the temporal dynamics of non-speech sounds and show some initial discrimination capabilities that are not specific to any language. This is our *pretrained* condition, which aligns with attunement or universal theories.

These two types of models (no pretraining vs. pretrained) undergo an exposure phase, during which they receive the exact same language experience in the form of either Japanese or American English recordings. We then compare their learning trajectories in terms of speech sound discrimination capabilities.

Methods

Pretraining dataset

To build the dataset of ambient sounds, we started with the Animal Sound Archive (Frommolt et al., 2006; GBIF.org, 2023), which consists of 78 hours of field recordings of animals. We supplemented it with 422 hours from AudioSet (Gemmeke et al., 2017), excluding utterances annotated as human vocalizations or music and retaining only animal sounds or everyday environmental sounds. Additionally, we filtered out the remaining speech segments missed by human annotators using the model proposed in Bredin et al. (2020). Our pretraining set comprises 500 hours of ambient sounds.

Training datasets

The Japanese training set is derived from the Corpus of Spontaneous Japanese (Maekawa, 2003), and the American English corpus is made of audiobooks (Kahn et al., 2020; Kearns, 2014). For both corpora, non-speech segments were removed (Bredin et al., 2020). We then selected a subset of American English audiobooks to match the characteristics of the Japanese corpus in two aspects: the number of speakers and the duration of speech per individual speaker. Ultimately, both corpora are made of approximately 500 hours of speech data. This quantity of speech is compatible with what infants hear during their first year as current estimates vary from 60 hours (Cristia et al., 2019) to 1,000 hours (Cristia, 2023).

For each language, we built smaller datasets by partitioning them into mutually exclusive subsets of varying sizes: 1 hour, 4 hours, 20 hours, and 100 hours. In all conducted experiments, whether on Japanese or English, we trained separate models on 15 subsets for the 1-hour, 4-hour, and 20-hour splits and 5 separate models for the 100-hour split.

The learner model

We chose Contrastive Predictive Coding (CPC) as our core learning mechanism (Oord et al., 2018; Rivière et al., 2020). In the Zero Resource Speech Challenge 2021 on unsupervised representation learning, CPC achieved the best speech sound discrimination scores (Dunbar et al., 2021). This model takes as input the raw waveforms. It is designed to predict future states of a sequence from its past in an autoregressive manner. In other words, given a sequence of observations, the model aims to accurately predict the next state of the input sequence based on its past context. This predictive task is achieved through a contrastive objective, where the model learns to distinguish between positive samples — the actual future states — and a set of negative samples — sampled from other parts of the dataset — during training (see implementation details in Appendix "Contrastive Predictive Coding").

Measuring speech sound discrimination

To assess the model's ability to discriminate contrasts, we conducted the same machine ABX sound discrimination task as used by Schatz et al. (2013). This evaluation procedure presents the model with three triphone stimuli pronounced by the same speaker labeled as A, B, and X. A and X are two instances of the same triphone (e.g., 'boot'), while B differs only in the central phone while maintaining the same context (e.g., 'beet'). We compute the corresponding representations R_A, R_B, and R_X for these stimuli and calculate the pairwise distances $d(R_A, R_X)$ and $d(R_B, R_X)$, with d the angular distance. As stimuli can have different durations, we perform Dynamic Time Warping (DTW) to obtain a time alignment before computing the average angular distance along the shortest DTW path. The representations of A and X returned by the model are more similar than those of B and X if $d(R_A, R_X) < d(R_B, R_X)$, in which case the model is considered to be correct in discriminating the contrasts. The ABX accuracy is computed as the average number of times the model provides the correct answer across all possible triphones and all possible contrasts. Alternatively, the accuracy can be computed across all possible triphones containing a specific contrast (e.g., [ɹ]-[l]).

Evaluation sets

We used the same evaluation sets as in Schatz et al. (2021). These sets consist of two Japanese corpora – the left-out subset of the CSJ and the Globalphone corpus of Japanese (GPJ) (Schultz, 2002) — and two American English corpora – a subset of the Wall Street

Journal corpus (WSJ) (Paul & Baker, 1992) and the Buckeye corpus (Pitt et al., 2005). The CSJ and Buckeye corpora contain more spontaneous speech, while GPJ and WSJ are composed of read speech. The CSJ evaluation set was built from the speech of speakers absent in the training set. All four evaluation sets are made of approximately ten hours of speech along with their forced-aligned phonetic transcripts.. Across registers (read or spontaneous speech), the number of speakers, the proportion of male and female speakers, and the cumulated duration per speaker are matched.

We compute the ABX accuracy in the *native* and the *non-native* condition. In the native condition, models are evaluated on the same language they have been exposed to (e.g., the Japanese model evaluated on our two Japanese evaluation sets). In the non-native condition, models are evaluated on the language they have not been exposed to (e.g., the Japanese model evaluated on our two American English evaluation sets). When mutually exclusive training sets of the same duration are available, we consider the mean and the standard deviation of the ABX accuracy in either the native or non-native condition.

Results and discussion

We begin by measuring the ABX accuracy of both our unpretrained and pretrained learners to assess their initial speech sound discrimination capabilities. We then compare the learning trajectories displayed by our two types of learners during the language exposure phase. To gain deeper insights into the nature of our two initial states (no pretraining vs. pretrained), we visualize the separability of the representations according to phonetic categories. Finally, we reflect on how the learning trajectories exhibited by our learners on the [w]-[j] and [ɹ]-[l] contrasts fare with the data observed in infants.

Initial speech sound discrimination capabilities and developmental trajectories

Panel **a)** of Figure 2 shows the average American English and Japanese ABX accuracy obtained by our two initial states: with no pretraining (in blue) or with pretraining on ambient sounds (in orange). Our randomly initialized model, which has not been pretrained, obtains an average ABX accuracy of 62.3% (μ_{JP} = 65.2%, μ_{EN} = 59.4%). In contrast, our model pretrained on ambient sounds obtains 92.4% ABX accuracy (μ_{JP} = 93.1%, μ_{EN} = 91.8%) showing better discrimination capabilities. We interpret the surprisingly high ABX accuracy obtained by our model pretrained on non-speech sounds as evidence that learning generic representations not specific to any language is enough to discriminate most human speech sounds.

Now that our first goal – inducing initial speech sound discrimination capabilities in our model – has been achieved, we analyze the learning trajectory exhibited by our model after exposure to either American English or Japanese in panel **b)**. Here, we

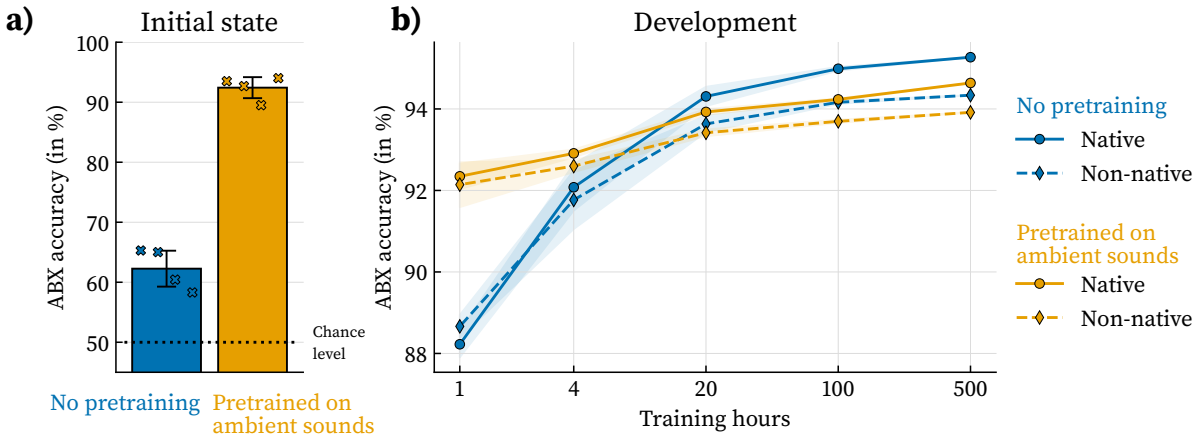

Figure 2. ***Comparison of our learner trained with no pretraining (in blue) or with pretraining on ambient sounds (in orange). Panel a) shows the average American English and Japanese ABX accuracy obtained by both types of learners before language experience (initial state). Panel b) shows the same information for native (same training and test language; dashed line) and non-native (different training and test languages; solid line) models as a function of the quantity of speech available in the training set (development). Error bars in panel a) represent +/- the standard deviation computed across our four evaluation sets. Shaded areas in panel b) represent +/- the standard deviation computed across mutually exclusive training sets whose number depends on the quantity of data.***

distinguish between native (same training and test language, solid line) and non-native (different training and test languages, dashed line) models.

Let us first focus on the trajectory exhibited by our model that has not been pretrained in blue. For low data quantities (between 1 and 4 hours), the native and non-native models obtain similar ABX accuracies, indicating that models have not yet learned language-specific representations. In other words, the American English model discriminates American English sounds as accurately as the Japanese model (and vice-versa). It is only after substantial exposure to their 'native' language (20 hours) that models start learning language-specific representations. Overall, we observe a positive effect of data quantity on the sound discrimination performance of our models. This is true for both the native and the non-native models. The more speech the model receives, the better it discriminates sounds. While this is expected in the native condition (e.g., exposure to Japanese makes the model better at discriminating Japanese sounds), this might be more surprising in the non-native condition. This is because there are many shared sounds across the two languages and the results reported in panel **b)** are computed across all possible contrasts – similar to what has been observed by Lavechin et al. (2022), Matusevych et al. (2023), and Schatz et al. (2021).

We now turn to the model pretrained on ambient sounds in orange. The pretrained model starts with better sound discrimination capabilities and exhibits a slower learning trajectory. Similarly to models which have not been pretrained, models pretrained on ambient sounds obtain a higher ABX accuracy with an increase in the quantity of speech. They also learn more language-specific representations as they receive more speech (the gap between the orange solid and dashed lines broadens with the number of training hours). Interestingly, after exposure to 500 hours of speech, pretrained models performed slightly worse than models that were not pretrained. Similarly, they learn representations that are less language-specific. Indeed, the relative difference in ABX error rate between native and non-native models is 16.5% in the no-pretraining condition versus 11.9% in the pretrained condition. We conducted two-way ANOVA analyses with factors nativeness and training language for each speech quantity (1h, 4h, 20h and 100h). In all settings the p-value was lower than $.0001$ indicating significant differences between the native and non-native models. While pretraining on ambient sounds initially steers the model in a favorable direction enabling it to discriminate speech sound contrasts effectively, this pre-exposure to non-speech sounds ends up hurting the performance of our model in processing speech sounds. Although it is hard to provide precise evidence, we hypothesize that, even after exposure to 500 hours of speech, some neurons are still dedicated to processing non-speech sounds.

Visualization of the initial sound discrimination capabilities

To better understand the initial sound discrimination capabilities induced previously through pretraining, we visualize in Figure 3 the phone representations in a two-dimensional space using the t-distributed Stochastic Neighbor Embedding (t-SNE) method – as done in de Seyssel et al. (2022) or Lavechin et al. (2022).

Panel **a)** shows the t-SNE projection of the phone representations of our two initial states: no pretraining versus pretrained on ambient sounds. Although no fine-grained separability between sonority categories was expected for the unpretrained model, we still observe some degree of separability between consonants and vowels. This aligns with the above-chance ABX accuracy of 62.3% obtained by this model (Figure 2). The model pretrained on ambient sounds show drastically different separability patterns. Here, we observe that phones are organized along a sonority continuum with a relatively good separability between the different categories, despite the model never receiving speech sounds during pretraining. Although results are only presented on our American English test sets, similar patterns are observed on our Japanese test set.

In panel **b)**, we specifically study the separability between the [ɹ]-[l] and [w]-[j] contrasts which will be the focus of the upcoming section. Our unpretrained model shows no separability for the [ɹ]-[l] or [w]-[j] contrast. However, this is not the case with our model pretrained on ambient sounds, which shows good separability for both contrasts.

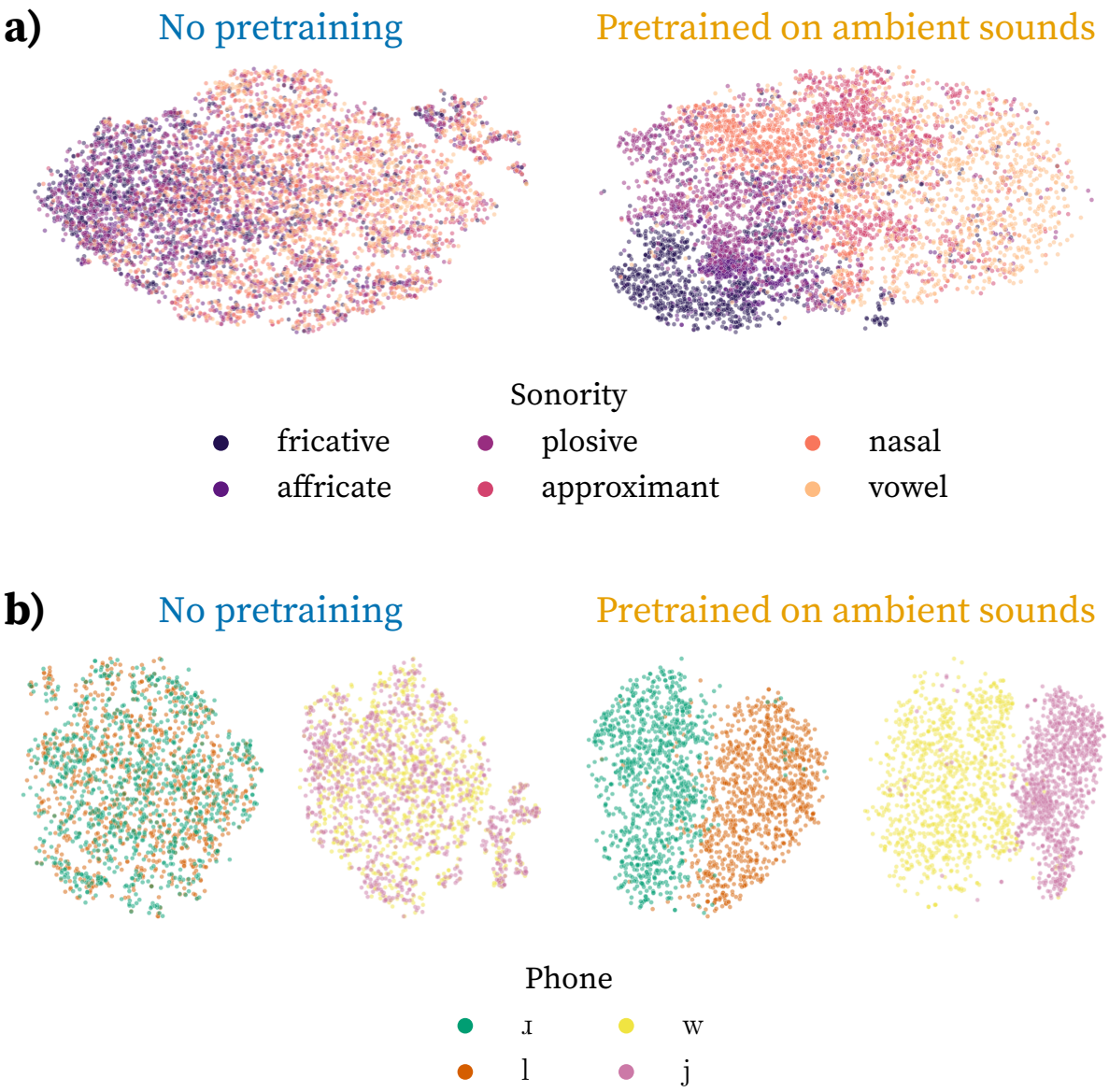

Figure 3. ***Visualization of the initial sound discrimination capabilities for our learner trained with no pretraining (in blue) or with pretraining on ambient sounds (in orange). Panel a) shows t-SNEs visualizations of the continuous representations (last layer) averaged within phones in the American English test set according to sonority. Panel b) shows the same information for the American English [ɹ]-[l] and [w]-[j] contrasts. Each point is the t-SNE projection of an individual phone's representation.***

These results demonstrate that inducing 'innate' speech sound discrimination capabilities is possible via pretraining on non-speech sounds. The first version of our model comes with no pretraining (random initialization) and shows limited initial speech

sound discriminability. This version corresponds to the initial state of most computational models of early phonetic learning but does not necessarily align with dominant theories of early phonetic acquisition in infants – it implements the perceptual learning theory. A second version of our model comes with pretraining and shows relatively good speech sound discriminability – it implements the attunement or universal theory.

Now that we have two different initial states at both ends of the nature-nurture continuum, an important question arises: Which better predicts the developmental trajectory observed in infants?

Individual trajectories for the [ɹ]-[l] and [w]-[j] pairs

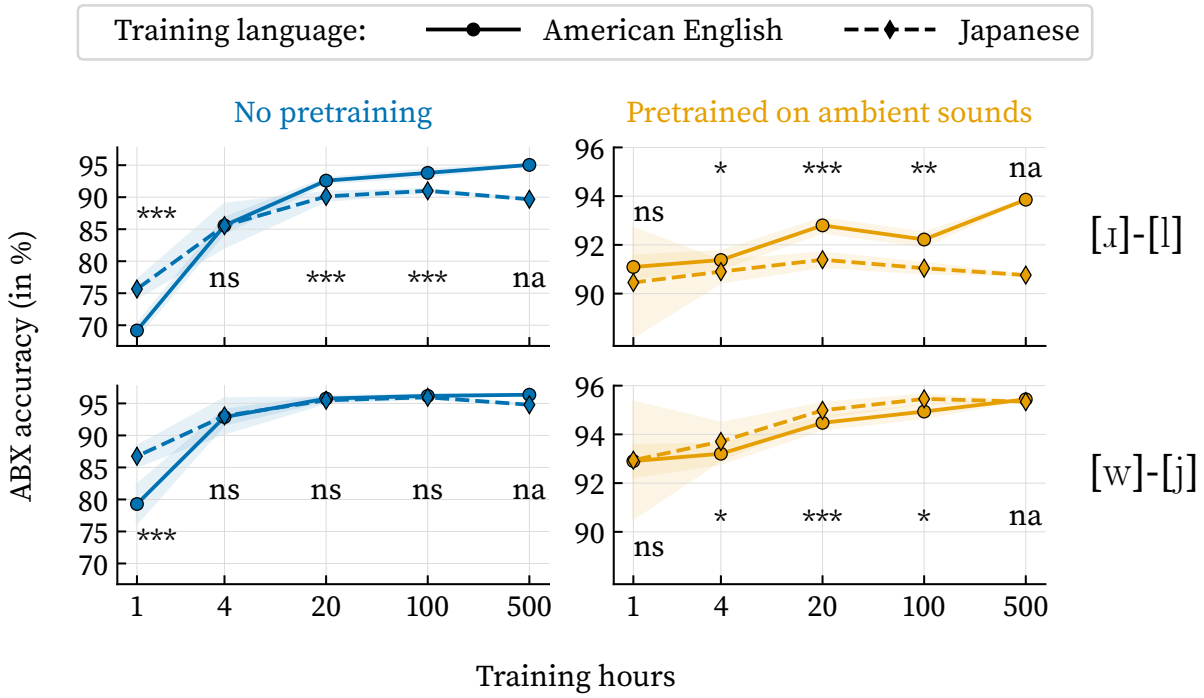

Figure 4. ***Comparison of the learning trajectory exhibited by our model without pretraining (in blue) or with pretraining on ambient sounds (in orange) on the [ɹ]-[l] and [w]-[j] pairs. Models are trained on either American English (solid line) or Japanese (dashed line). [ɹ]-[l] and [w]-[j] occurrences are extracted from our American English evaluation sets. Shaded areas represent +/- the standard deviation computed across mutually exclusive training sets whose number depends on the quantity of data. Significance scores are obtained with a one-way ANOVA with factor training language (na: not applicable, ns: not significant, * p<.05, **, p<.001, *** p<.0001).***

To investigate this question, we study the learning trajectories exhibited by our models on the American English [ɹ]-[l] pair, contrastive in American English but not in Japanese, and the [w]-[j] control pair, contrastive in both languages. Figure 4 shows the ABX

accuracy obtained on these contrasts for our American English model (solid line) or our Japanese model (dashed line), in the no pretraining condition (in blue) or the pretrained condition (in orange).

Let us first focus on the no pretraining condition. The American English model better discriminates the [ɹ]-[l] contrast than the non-native Japanese model. We also observe that the gap between the two models increases with the quantity of speech. On the contrary, on the [w]-[j] contrast, our native American English and our non-native Japanese models develop similar discrimination performances. These results closely replicate those of Li et al. (2020) and Schatz et al. (2021) with a different model and correspond, at least to some extent, to what has been observed in infants, namely that 10-12 month-old American English- and Japanese-learning infants show a similar discrimination performance on the [w]-[j] contrast, but American English-learning infants show better discrimination on the [ɹ]-[l] contrast.

Looking more closely at how the trajectories exhibited by our models fare with those observed in infants, we observe notable differences. While the American English model succeeds in reproducing the facilitation trajectory observed in American English infants on the [ɹ]-[l] contrast, this is not the case with the Japanese model. Indeed, our unpretrained Japanese model also follows an inductive trajectory, while Kuhl et al. (2006) and Tsushima et al. (1994) reported a loss trajectory in Japanese-learning infants between 6-8 and 10-12 months for this specific contrast. Although there is less data available on the [w]-[j] contrast, prevailing theories predict either a facilitation or a maintenance trajectory compatible with the trajectories exhibited by our unpretrained model.

We now turn to a similar analysis of the trajectories exhibited by our models pretrained on ambient sounds in orange. In this condition, our native American English model replicates the facilitation trajectory observed in American English-learning infants on the [ɹ]-[l] contrast. On this same contrast, our non-native Japanese model now exhibits a maintenance trajectory with constant performance regardless of the quantity of speech available in the training set. While this maintenance trajectory still does not perfectly match what has been observed in infants (i.e., a loss trajectory), the match is closer than in the no pretraining condition. Indeed, Kuhl et al. (2006) report a low difference in discrimination accuracy between the 6-8 month-old group and the 10-12 month-old Japanese group. Furthermore, the effect of age was found not significant for the Japanese group. Therefore, we interpret Kuhl's results as compatible with the maintenance trajectory exhibited by our Japanese model. Our interpretation of the learning trajectories exhibited by our model concerning the [w]-[j] contrast in relation to the infant literature is similar to that presented for the no pretraining condition and will not be repeated here. An interesting observation, however, is that performance on the [w]-[j] contrast still improves after 500 hours of speech, contrary to the no pretraining condition in which performance flattens after 20 hours of speech.

Experiment 2: inducing initial speech sound discrimination capabilities through multilingual pretraining

Experiment 1 showed that it is possible to induce innate speech sound discrimination capabilities by pretraining on ambient sounds. During the developmental phase, our pretrained model exposed to Japanese exhibits a maintenance trajectory on the [ɹ]-[l] contrast, more closely resembling infant behavioral data that suggest a loss trajectory (Figure 4). In the present experiment, we ask whether it is possible to induce higher initial speech sound discrimination capabilities – and perhaps obtain a loss trajectory on the [ɹ]-[l] contrast – with a different pretraining strategy: pretraining on a set of typologically diverse languages.

Methods

We use the same training sets, learner, evaluation sets, and evaluation protocol as used in Experiment 1. The only difference is that we pretrain on a multilingual corpus derived from VoxPopuli (Wang et al., 2021), a large-scale multilingual speech corpus containing recordings of European Parliament events. We remove the Germanic languages from the 23 languages present in the dataset to prevent the model trained on English from being positively biased. This procedure resulted in selecting 18 typologically diverse languages[5]. To ensure consistency with Experiment 1, our pretraining set is made of 500 hours of speech uniformly sampled across languages, resulting in approximately 28 hours per language.

Results and discussion

Initial sound discrimination capabilities and developmental trajectories

Panel **a)** of Figure 5 suggests that pretraining on multilingual is sensibly similar to pretraining on ambient sounds in terms of initial speech sound discrimination capabilities (μ_{JP} = 93.5 %, μ_{EN} = 92.1%). Contrary to our initial hypothesis, training on mulitilingual speech does not yield higher speech sound discrimination capabilities compared to pretraining on ambient sounds.

During the developmental phase, the learning trajectories obtained in the pretrained condition are similar than those obtained in Experiment 1. Two-way ANOVAs also resulted in p-values lower than .0001 for all speech quantities indicating significant differences between the native and non-native models. A notable difference compared to Experiment 1 is that pretraining on multilingual speech does not hurt the performance obtained after 500 hours of exposure, contrary to what was observed when pretraining on ambient sounds, as shown in panel **b)** of Figure 5.

[5]Bulgarian, Czech, Croatian, Estonian, Finnish, French, Greek, Hungarian, Italian, Latvian, Lithuanian, Maltese, Polish, Portuguese, Romanian, Slovak, Slovene and Spanish.

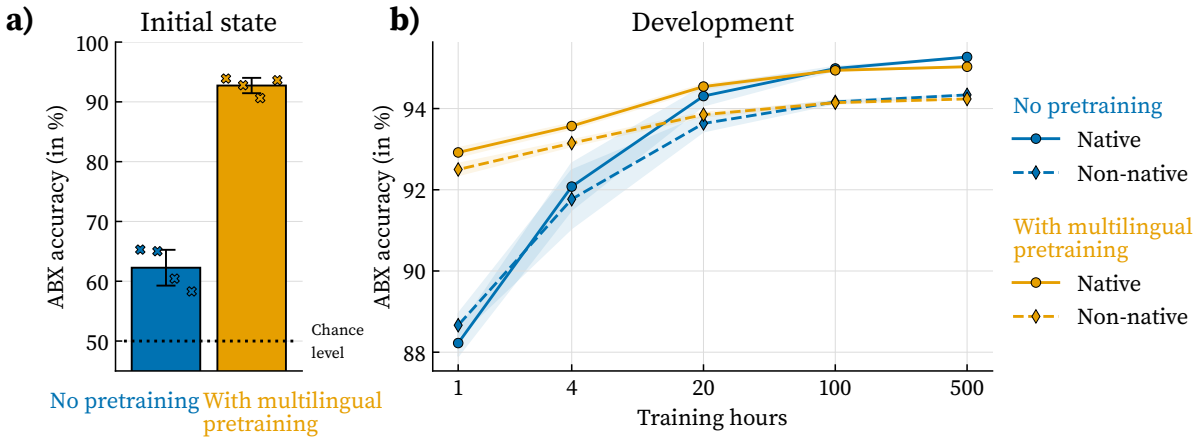

Figure 5. ***Comparison of our learner trained with no pretraining (in blue) or with multilingual pretraining (in orange) for native (same training and test language; solid line) and non-native (different training and test languages; dashed line) models as a function of the quantity of speech available in the training set (development). Shaded areas represent +/- the standard deviation computed across mutually exclusive training sets whose number depends on the quantity of data.***

Individual trajectories for the [ɹ]-[l] and [w]-[j] pairs

Again, the learning trajectories on the [ɹ]-[l] and [w]-[j] contrasts in Figure 6 are sensibly similar to those observed in Experiment 1. However, in the pretrained condition, the Japanese model seems to follow a facilitation trajectory on the [ɹ]-[l] contrast, contrary to the maintenance trajectory observed in Experiment 1. This is due to the lower ABX accuracy on the [ɹ]-[l] contrast obtained by the initial state pretrained on multilingual speech (85.8% on Buckeye, 91.7% on WSJ) compared to the initial state pretrained on ambient sounds (87.8% on Buckeye, 93.9% on WSJ).

In the context of this study, pretraining on 500 hours of multilingual speech does not seem to present any advantage as compared to pretraining on ambient sounds. Admittedly, training on larger quantities of multilingual speech – and perhaps a higher number of languages – may yield a model that starts with higher initial speech sound discrimination capabilities, as was the initial goal of this experiment.

In a concluding experiment (see Experiment 3 in Appendix), we show that our model reproduces the trajectories observed in infants: facilitation on the [ɹ]-[l] contrast and maintenance on the [w]-[j] contrast for American English-learning infants; loss on the [ɹ]-[l] contrast and maintenance on the [w]-[j] contrast for Japanese-learning infants. This is achieved through cross-lingual pretraining, where models are pre-trained on either American English or Japanese and then further trained on the language to which they have not been exposed. This protocol assumes higher non-native sound discrimination

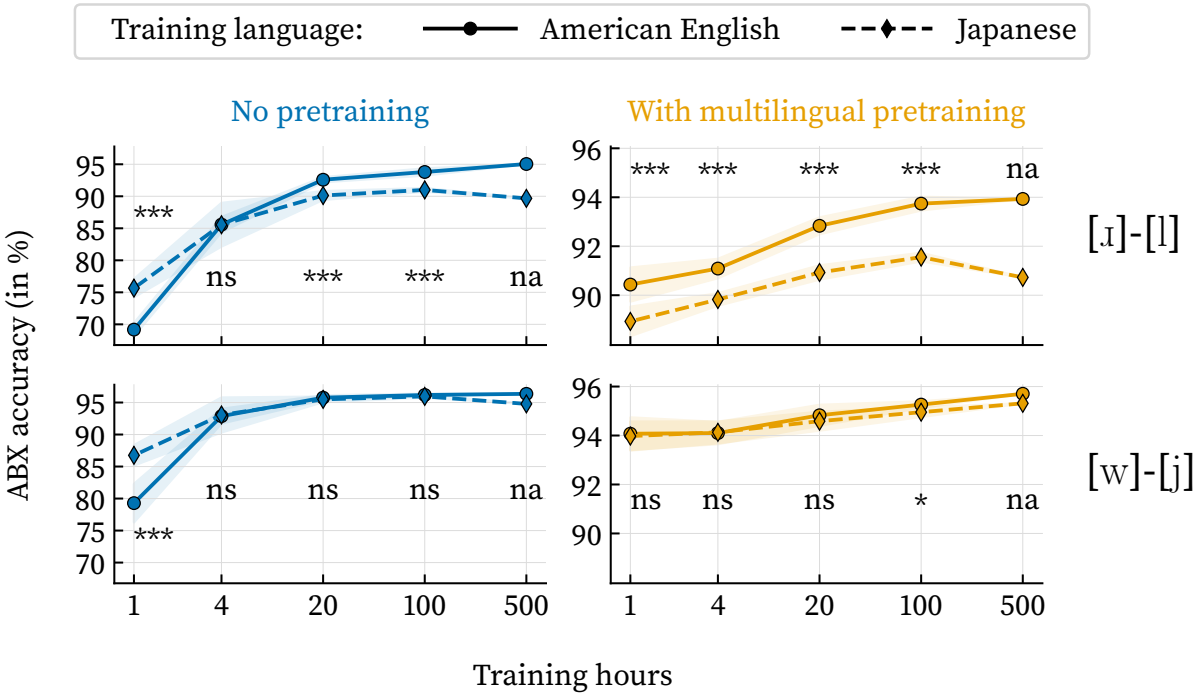

Figure 6. ***Comparison of the learning trajectory exhibited by our model without pretraining (in blue) or with multilingual pretraining (in orange) on the [ɹ]-[l] and [w]-[j] pairs. Models are trained on either American English (solid line) or Japanese (dashed line). [ɹ]-[l] and [w]-[j] occurrences are extracted from our American English evaluation sets. Shaded areas represent +/- the standard deviation computed across mutually exclusive training sets whose number depends on the quantity of data. Significance scores are obtained with a one-way ANOVA with factor training language (na: not applicable, ns: not significant, * p<.05, **, p<.001, *** p<.0001).***

capabilities than in Experiment 1 or 2. While this final experiment may not have direct relevance from an evolutionary perspective, it achieves to demonstrate that our model's performance can maintain, improve, or deteriorate depending on the interaction between innate and environmental factors.

General discussion

What is the respective contribution of innate factors and experience in child language acquisition? Without bringing indisputable evidence to the question, we proposed a novel method to build computational models of early phonetic learning that start with initial sound discrimination capabilities. Conducting two experiments, we showed that the model endowed with initial capabilities could demonstrate maintenance, facilitation, or loss trajectories, as opposed to the standard model, which learns from scratch

and mostly exhibits facilitation trajectories. Here, we reflect on the implications of our findings for the existing literature on modeling infant phonetic learning. We first return to the idea of language-universal capabilities in newborns. We then propose other approaches to induce such capabilities in computational models. Finally, we reflect on how our work can be extended in a more systematic approach to the study of infant phonetic learning.

The idea of universal speech perception capabilities at the initial state is prevailing in current theories of language acquisition. In Kuhl's (2004) words, infants have an "initial universal ability to distinguish between phonetic units". Werker and Curtin (2005) write about "language-general" and "language-specific" perception. In our view, testing these theories should not only consist in collecting relevant data in infants but also in implementing them (de Seyssel et al., 2023; Dupoux, 2018). In that regard, our approach has two advantages. First, it encourages us to transform verbally-expressed ideas into implementable algorithms. Second, it offers us ways to test and compare our verbal theories.

In Experiment 1, we implemented the idea of a language-universal perceptual space by pretraining on ambient sounds. We found that the learning trajectories taken by the model during the developmental phase better fit the observed data in infants, providing evidence in favor of attunement and universal theories. In Experiment 2, we proposed a second strategy that consists of pretraining on multilingual speech data. Admittedly, one could devise different strategies – that should be equally evaluated in terms of their fit with observed data in infants. For instance, one might pretrain at a larger scale both in terms of quantity of speech data and number of languages. This could be done by training on the more than 7,000 languages being spoken worldwide[6] before comparing the learning trajectories taken by the model when trained on a single language with those observed in infants (hypothesizing rather strong initial capabilities). On the contrary, one could devise strategies to build models that assume poorer initial capabilities by training on a different source of data or by lowering the amount of data available in the pretraining set. Importantly, our goal is not to provide an explanation of the evolutionary transition from a primitive amphibian auditory system to the human auditory system. In that regard, the pretraining strategy has no other function than to hypothesize some degree of initial capabilities, offering us a rather vast ground for exploration.

In contrast to existing modeling studies (Lavechin et al., 2024; Li et al., 2020; Matusevych et al., 2023; Schatz et al., 2021), our approach goes beyond evaluating models solely based on measures of native advantage (i.e., better discrimination score for the model for whom the contrast is native). It includes assessing their fit to developmental trajectories observed in infants. This work focused on the [ɹ]-[l] pair contrastive in American

[6] https://www.ethnologue.com

English but not in Japanese and the [w]-[j] pair contrastive in both languages. However, there is available data in Zulu, Tigrinya, Taa, Nuu-Chah-Nulth, Spanish, Hindi, Czech, Nthlakampx, and Mandarin (Best et al., 2016). A more systematic approach would involve building a training set for each of these languages and studying the speech sound discrimination patterns developed by computational models. Successful models, capturing a significant proportion of the variance of the available empirical record, can then be used to obtain predictions on contrasts that have yet to be studied. These predicted trajectories can subsequently be validated or refuted through new sound discrimination experiments with infants. Alternatively, instead of focusing on data from individual studies, one could compare the learning outcomes developed by computational models against robust data from meta-analyses as proposed by Cruz Blandón et al. (2023). We strongly believe that such a systematic dialogue between experimental and modeling studies is essential to foster theory-building in psychological sciences (Frank et al., 2017).

Conclusion

Even though current AI language models have been considered as supporting empiricist views of language learning, these models offer a much larger range of theoretical options. By decomposing model training in a (potentially long) evolutionary phase and a (potentially short) developmental phase, they can implement either extreme versions of empiricism, or extreme versions of nativism, with a whole range of intermediary cases. In our work, we conducted two experiments that demonstrated the possibility of inducing initial sound discrimination capabilities in our computational model of early phonetic learning. Contrary to the randomly initialized model, the models pre-equipped with discrimination capabilities showed learning trajectories more closely resembling those observed in infants. Further research is needed to establish in a more quantitative fashion what model of the initial state would fit best the observed learning trajectories in infants.

References

Ambridge, B., & Lieven, E. V. (2011). *Child language acquisition: Contrasting theoretical approaches*. Cambridge University Press. https://doi.org/10.1017/CBO9780511975073

Antetomaso, S., Miyazawa, K., Feldman, N., Elsner, M., Hitczenko, K., & Mazuka, R. (2017). Modeling phonetic category learning from natural acoustic data. *Proceedings of the annual Boston University Conference on Language Development*. https://par.nsf.gov/biblio/10057880

Aslin, R., & Pisoni, D. (1980). Some developmental processes in speech perception. In G. H. Yeni-Komshian, J. F. Kavanagh, & C. A. Ferguson (Eds.), *Child phonology: Vol 2, Perception* (pp. 67–96). New York: Academic Press.

Aslin, R., Werker, J. F., & Morgan, J. L. (2002). Innate phonetic boundaries revisited (l). *The Journal of the Acoustical Society of America*, *112*(4), 1257–1260. https://doi.org/10.1121/1.1501904

Bates, E., Elman, J., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996, October). *Rethinking innateness: A connectionist perspective on development*. The MIT Press. https://doi.org/10.7551/mitpress/5929.001.0001

Best, C. T., Goldstein, L. M., Nam, H., & Tyler, M. D. (2016). Articulating what infants attune to in native speech. *Ecological Psychology*, *28*(4), 216–261. https://doi.org/10.1080/10407413.2016.1230372

Best, C. T., McRoberts, G. W., LaFleur, R., & Silver-Isenstadt, J. (1995). Divergent developmental patterns for infants' perception of two nonnative consonant contrasts. *Infant behavior and development*, *18*(3), 339–350. https://doi.org/10.1016/0163-6383(95)90022-5

Best, C. T., & Strange, W. (1992). Effects of phonological and phonetic factors on cross-language perception of approximants. *Journal of phonetics*, *20*(3), 305–330. https://doi.org/10.1016/S0095-4470(19)30637-0

Bredin, H., Yin, R., Coria, J. M., Gelly, G., Korshunov, P., Lavechin, M., Fustes, D., Titeux, H., Bouaziz, W., & Gill, M.-P. (2020). Pyannote.audio: Neural building blocks for speaker diarization. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7124–7128. https://doi.org/10.1109/ICASSP40776.2020.9052974

Chomsky, N. (1957). Syntactic structures. Mouton de Gruyter. https://doi.org/10.1515/9783112316009

Cristia, A. (2023). A systematic review suggests marked differences in the prevalence of infant-directed vocalization across groups of populations. *Developmental Science*, *26*(1), e13265. https://doi.org/10.1111/desc.13265

Cristia, A., Dupoux, E., Gurven, M., & Stieglitz, J. (2019). Child-directed speech is infrequent in a forager-farmer population: A time allocation study. *Child Development*, *90*(3), 759–773. https://doi.org/10.1111/cdev.12974

Cruz Blandón, M. A., Cristia, A., & Räsänen, O. (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

de Seyssel, M., Lavechin, M., Adi, Y., Dupoux, E., & Wisniewski, G. (2022). Probing phoneme, language and speaker information in unsupervised speech representations. *Proc. Interspeech 2022*, 1402–1406. https://doi.org/10.21437/Interspeech.2022-373

de Seyssel, M., Lavechin, M., & Dupoux, E. (2023). Realistic and broad-scope learning simulations: first results and challenges. *Journal of Child Language*, 1–24. https://doi.org/10.1017/S0305000923000272

Dunbar, E., Bernard, M., Hamilakis, N., Nguyen, T. A., de Seyssel, M., Rozé, P., Rivière, M., Kharitonov, E., & Dupoux, E. (2021). The Zero Resource Speech Challenge 2021: Spoken language modelling. *Proc. Interspeech 2021*, 1574–1578. https://doi.org/10.21437/Interspeech.2021-1755

Dupoux, E. (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

Eilers, R. E., & Minifie, F. D. (1975). Fricative discrimination in early infancy. *Journal of speech and Hearing Research*, *18*(1), 158–167. https://doi.org/10.1044/jshr.1801.158

Eimas, P. D., Siqueland, E. R., Jusczyk, P., & Vigorito, J. (1971). Speech perception in infants. *Science, 171*(3968), 303–306. https://doi.org/10.1126/science.171.3968.303

Frank, M. C., Bergelson, E., Bergmann, C., Cristia, A., Floccia, C., Gervain, J., Hamlin, J. K., Hannon, E. E., Kline, M., Levelt, C., et al. (2017). A collaborative approach to infant research: Promoting reproducibility, best practices, and theory-building. *Infancy*, *22*(4), 421–435. https://doi.org/10.1111/infa.12182

Frommolt, K.-H., Bardeli, R., Kurth, F., & Clausen, M. (2006). The animal sound archive at the Humboldt-University of Berlin: Current activities in conservation and improving access for bioacoustic research. *Advances in Bioacoustics 2*, 139–144. https://www.ibac.info/advances-in-bioacoustics-ii#aib10

GBIF.org. (2023). GBIF occurrence download. https://doi.org/10.15468/dl.dmckt3

Gemmeke, J. F., Ellis, D. P. W., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M., & Ritter, M. (2017). Audio set: An ontology and human-labeled dataset for audio events. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 776–780. https://doi.org/10.1109/ICASSP.2017.7952261

Gilkerson, J., Richards, J. A., Warren, S. F., Montgomery, J. K., Greenwood, C. R., Kimbrough Oller, D., Hansen, J. H., & Paul, T. D. (2017). Mapping the early language environment using all-day recordings and automated analysis. *American journal of speech-language pathology, 26*(2), 248–265. https://doi.org/10.1044/2016_AJSLP-15-0169

Hauser, M. D., Chomsky, N., & Fitch, W. T. (2002). The faculty of language: what is it, who has it, and how did it evolve? *Science, 298*(5598), 1569–1579. https://doi.org/10.1126/science.298.5598.1569

Huang, Y., & Rao, R. P. (2011). Predictive coding. *Wiley Interdisciplinary Reviews: Cognitive Science, 2*(5), 580–593. https://doi.org/10.1002/wcs.142

Hueber, T., Tatulli, E., Girin, L., & Schwartz, J.-L. (2020). Evaluating the potential gain of auditory and audiovisual speech-predictive coding using deep learning. *Neural Computation, 32*(3), 596–625. https://doi.org/10.1162/neco_a_01264

Joanisse, M. F., & McClelland, J. L. (2015). Connectionist perspectives on language learning, representation and processing. *Wiley Interdisciplinary Reviews: Cognitive Science, 6*(3), 235–247. https://doi.org/10.1002/wcs.1340

Kahn, J., Rivière, M., Zheng, W., Kharitonov, E., Xu, Q., Mazaré, P., Karadayi, J., Liptchinsky, V., Collobert, R., Fuegen, C., Likhomanenko, T., Synnaeve, G., Joulin, A., Mohamed, A., & Dupoux, E. (2020). Libri-light: A benchmark for ASR with limited or no supervision. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7669–7673. https://doi.org/10.1109/ICASSP40776.2020.9052942

Kearns, J. (2014). LibriVox: Free public domain audiobooks. *Reference Reviews, 28*(1), 7–8. https://doi.org/10.1108/RR-08-2013-0197

Kharitonov, E., Rivière, M., Synnaeve, G., Wolf, L., Mazaré, P.-E., Douze, M., & Dupoux, E. (2021). Data augmenting contrastive learning of speech representations in the time domain. *2021 IEEE Spoken Language Technology Workshop (SLT)*, 215–222. https://doi.org/10.1109/SLT48900.2021.9383605

Kuhl, P. K. (2004). Early language acquisition: cracking the speech code. *Nature reviews neuroscience, 5*(11), 831–843. https://doi.org/10.1038/nrn1533

Kuhl, P. K., Conboy, B. T., Coffey-Corina, S., Padden, D., Rivera-Gaxiola, M., & Nelson, T. (2008). Phonetic learning as a pathway to language: new data and native language magnet theory expanded (NLM-e). *Philosophical Transactions of the Royal Society B: Biological Sciences, 363*(1493), 979–1000. https://doi.org/10.1098/rstb.2007.2154

Kuhl, P. K., & Iverson, P. (1995). Linguistic experience and the "perceptual magnet effect". *Speech perception and linguistic experience: Issues in cross-language research*, 121–154.

Kuhl, P. K., Stevens, E., Hayashi, A., Deguchi, T., Kiritani, S., & Iverson, P. (2006). Infants show a facilitation effect for native language phonetic perception between 6 and 12 months. *Developmental science, 9*(2), F13–F21. https://doi.org/10.1111/j.1467-7687.2006.00468.x

Kuhl, P. K., Tsao, F.-M., & Liu, H.-M. (2003). Foreign-language experience in infancy: Effects of short-term exposure and social interaction on phonetic learning. *Proceedings of the National Academy of Sciences, 100*(15), 9096–9101. https://doi.org/10.1073/pnas.1532872100

Lavechin, M., De Seyssel, M., Titeux, H., Bredin, H., Wisniewski, G., Cristia, A., & Dupoux, E. (2022). Can statistical learning bootstrap early language acquisition? a modeling investigation. https://doi.org/10.31234/osf.io/rx94d

Lavechin, M., de Seyssel, M., Métais, M., Metze, F., Mohamed, A., Bredin, H., Dupoux, E., & Cristia, A. (2024). Modeling early phonetic acquisition from child-centered audio data. *Cognition, 245*, 105734. https://doi.org/10.1016/j.cognition.2024.105734

Lavechin, M., Sy, Y., Titeux, H., Blandón, M. A. C., Räsänen, O., Bredin, H., Dupoux, E., & Cristia, A. (2023). BabySLM: language-acquisition-friendly benchmark of self-supervised spoken language models. *Proc. INTERSPEECH 2023*, 4588–4592. https://doi.org/10.21437/Interspeech.2023-978

Li, R., Schatz, T., Matusevych, Y., Goldwater, S., & Feldman, N. H. (2020). Input matters in the modeling of early phonetic learning. *Proceedings of the Annual Conference of the Cognitive Science Society*. https://par.nsf.gov/biblio/10176646

Maekawa, K. (2003). Corpus of spontaneous japanese: its design and evaluation. *Proc. ISCA/IEEE Workshop on Spontaneous Speech Processing and Recognition*, paper MMO2. https://www.isca-speech.org/archive/sspr_2003/maekawa03_sspr.html

Matusevych, Y., Schatz, T., Kamper, H., Feldman, N. H., & Goldwater, S. (2023). Infant Phonetic Learning as Perceptual Space Learning: A Crosslinguistic Evaluation of Computational Models. *Cognitive Science, 47*(7), e13314. https://doi.org/10.1111/cogs.13314

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82*(3), B101–B111. https://doi.org/10.1016/s0010-0277(01)00157-3

McMurray, B., Danelz, A., Rigler, H., & Seedorff, M. (2018). Speech categorization develops slowly through adolescence. *Developmental psychology*, *54*(8), 1472. https://doi.org/10.1037/dev0000542

Miyawaki, K., Jenkins, J. J., Strange, W., Liberman, A. M., Verbrugge, R., & Fujimura, O. (1975). An effect of linguistic experience: The discrimination of [r] and [l] by native speakers of japanese and english. *Perception & Psychophysics*, *18*(5), 331–340.

Miyazawa, K., Kikuchi, H., & Mazuka, R. (2010). Unsupervised learning of vowels from continuous speech based on self-organized phoneme acquisition model. *Interspeech*. https://doi.org/10.21437/Interspeech.2010-757

Nguyen, T. A., de Seyssel, M., Rozé, P., Rivière, M., Kharitonov, E., Baevski, A., Dunbar, E., & Dupoux, E. (2020). The Zero Resource Speech Benchmark 2021: Metrics and baselines for unsupervised spoken language modeling. *NeurIPS Workshop on Self-Supervised Learning for Speech and Audio Processing*. https://arxiv.org/abs/2011.11588

Nittrouer, S. (2001). Challenging the notion of innate phonetic boundaries. *The Journal of the Acoustical Society of America*, *110*(3), 1598–1605. https://doi.org/10.1121/1.1379078

Oord, A. van den, Li, Y., & Vinyals, O. (2018). *Representation Learning with Contrastive Predictive Coding*. arXiv: 1807.03748 [cs, stat].

Paul, D. B., & Baker, J. M. (1992). The design for the Wall Street Journal-based CSR corpus. *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*. https://aclanthology.org/H92-1073

Pierrehumbert, J. B. (2003). Phonetic diversity, statistical learning, and acquisition of phonology. *Language and speech*, *46*(2-3), 115–154. https://doi.org/10.1177/00238309030460020501

Pitt, M. A., Johnson, K., Hume, E., Kiesling, S., & Raymond, W. (2005). The Buckeye corpus of conversational speech: labeling conventions and a test of transcriber reliability. *Speech Communication*, *45*(1), 89–95. https://doi.org/10.1016/j.specom.2004.09.001

Polka, L., Colantonio, C., & Sundara, M. (2001). A cross-language comparison of/d/–/ð/perception: evidence for a new developmental pattern. *The Journal of the Acoustical Society of America*, *109*(5), 2190–2201. https://doi.org/10.1121/1.1362689

Räsänen, O. (2012). Computational modeling of phonetic and lexical learning in early language acquisition: Existing models and future directions. *Speech Communication, 54*(9), 975–997. https://doi.org/10.1016/j.specom.2012.05.001

Räsänen, O., Nagamine, T., & Mesgarani, N. (2016). Analyzing distributional learning of phonemic categories in unsupervised deep neural networks. *CogSci... Annual Conference of the Cognitive Science Society. Cognitive Science Society (US). Conference, 2016*, 1757. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5775908

Rivière, M., Joulin, A., Mazaré, P.-E., & Dupoux, E. (2020). Unsupervised pretraining transfers well across languages. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7414–7418. https://doi.org/10.1109/ICASSP40776.2020.9054548

Rowland, C. (2013). *Understanding child language acquisition*. Routledge. https://doi.org/10.4324/9780203776025

Saffran, J. R., & Kirkham, N. Z. (2018). Infant statistical learning. *Annual Review of Psychology*, *69*(1), 181–203. https://doi.org/10.1146/annurev-psych-122216-011805

Schatz, T., Feldman, N. H., Goldwater, S., Cao, X.-N., & Dupoux, E. (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

Schatz, T., Peddinti, V., Bach, F., Jansen, A., Hermansky, H., & Dupoux, E. (2013). Evaluating speech features with the Minimal-Pair ABX task: Analysis of the classical MFC/PLP pipeline. *Proc. Interspeech 2013*, 1781–1785. https://doi.org/10.21437/Interspeech.2013-441

Schultz, T. (2002). Globalphone: a multilingual speech and text database developed at karlsruhe university. *Proc. 7th International Conference on Spoken Language Processing (ICSLP 2002)*, 345–348. https://doi.org/10.21437/ICSLP.2002-151

Singh, L., Rajendra, S. J., & Mazuka, R. (2022). Diversity and representation in studies of infant perceptual narrowing. *Child Development Perspectives, 16*(4), 191–199. https://doi.org/10.1111/cdep.12468

Steels, L., & De Boer, B. (2008). Embodiment and self-organization of human categories: A case study for speech. *Body, Language and Mind, 1*, 411–430. https://doi.org/10.1515/9783110207507.3.411

Swingley, D. (2009). Contributions of infant word learning to language development. *Philosophical Transactions of the Royal Society B: Biological Sciences, 364*(1536), 3617–3632. https://doi.org/10.1098/rstb.2009.0107

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). *LLaMA: Open and efficient foundation language models*. arXiv: 2302.13971 [cs.CL].

Tsao, F.-M., Liu, H.-M., & Kuhl, P. K. (2006). Perception of native and non-native affricate-fricative contrasts: Cross-language tests on adults and infants. *The Journal of the Acoustical Society of America, 120*(4), 2285–2294. https://doi.org/10.1121/1.2338290

Tsuji, S., & Cristia, A. (2014). Perceptual attunement in vowels: A meta-analysis. *Developmental psychobiology, 56*(2), 179–191. https://doi.org/10.1002/dev.21179

Tsushima, T., Takizawa, O., Sasaki, M., Shiraki, S., Nishi, K., Kohno, M., Menyuk, P., & Best, C. (1994). Discrimination of english /r-l/ and /w-y/ by japanese infants at 6-12 months: language-specific developmental changes in speech perception abilities. *Proc. 3rd International Conference on Spoken Language Processing (ICSLP 1994)*, 1695–1698. https://doi.org/10.21437/ICSLP.1994-438

Vallabha, G. K., McClelland, J. L., Pons, F., Werker, J. F., & Amano, S. (2007). Unsupervised learning of vowel categories from infant-directed speech. *Proceedings of the National Academy of Sciences, 104*(33), 13273–13278. https://doi.org/10.1073/pnas.0705369104

Wang, C., Riviere, M., Lee, A., Wu, A., Talnikar, C., Haziza, D., Williamson, M., Pino, J., & Dupoux, E. (2021). VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 993–1003. https://doi.org/10.18653/v1/2021.acl-long.80

Warstadt, A., Mueller, A., Choshen, L., Wilcox, E., Zhuang, C., Ciro, J., Mosquera, R., Paranjabe, B., Williams, A., Linzen, T., & Cotterell, R. (2023). Findings of the BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora. *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, 1–34. https://doi.org/10.18653/v1/2023.conll-babylm.1

Werker, J. F., & Curtin, S. (2005). PRIMIR: A developmental framework of infant speech processing. *Language learning and development, 1*(2), 197–234. https://doi.org/10.1207/s15473341lld0102_4

Werker, J. F., Gilbert, J. H., Humphrey, K., & Tees, R. C. (1981). Developmental aspects of cross-language speech perception. *Child development*, 349–355. https://doi.org/10.2307/1129249

Yamada, R. A., & Tohkura, Y. (1992). The effects of experimental variables on the perception of american english/r/and/l/by japanese listeners. *Perception & psychophysics*, *52*, 376–392. https://doi.org/10.3758/BF03206698

Data, Code and Materials Availability Statement

The source code of all models, experimentation scripts, and data processing scripts are available at https://github.com/mxmpl/initial-phonetic-learning. This repository also contains links to download the pretraining datasets of ambient sounds, the multilingual and English training sets, model checkpoints, and comprehensive results. Audioset and the Animal Sound Archive occurrence data are made available under a CC BY 4.0 license. The Animal Sound Archive audio files are licensed under CC BY-SA 4.0 and CC BY-NC-SA 4.0. The VoxPopuli and LibriVox recordings are in the public domain. The Editor granted an exemption to materials sharing for the following datasets, on the grounds that they are subject to copyright: CSJ, GPJ, WSJ, and Buckeye.

Authorship and Contributorship Statement

M.P., T.S., E.D., M.L. designed research; M. P. performed research; M.P, M.L. wrote the manuscript with contributions from T.S. and E.D. All authors approved the final version of the manuscript and agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Acknowledgements

This work was granted access to the HPC resources of IDRIS under the allocation 2023-AD011014368 made by GENCI and was supported in part by the Agence Nationale pour la Recherche (ANR-17-EURE-0017 Frontcog, ANR10-IDEX-0001-02 PSL*, ANR19-P3IA-0001 PRAIRIE 3IA Institute) and grants from CIFAR (Learning in Machines and Brains) and Meta AI Research (Research Grant). M. P. acknowledges Ph.D. funding from Agence de l'Innovation de Défense. T.S. work, carried out within the Institute of Convergence ILCB, was supported by grants from France 2030 (ANR-16-CONV-0002) and the Excellence Initiative of Aix-Marseille University (A*MIDEX).

Appendices

Contrastive predictive coding (implementation details)

Training a neural network in an unsupervised manner often requires designing a pretext task that will force the model to learn high-level representations of the input data. The pretext task in a Contrastive Predictive Coding algorithm is forward modeling. where the model is trained to predict the future states of a sequence based on its past context. During training, the model receives a positive example drawn from the near future up to 120 ms, and multiple negative examples not drawn from the near future. Given the past context of a sequence, the model has to come to reliably choose the positive sample over the negative ones.

In more technical terms, a non-linear encoder denoted as g_{enc} maps the observations x_t at time t to a latent representation $z_t = g_{\text{enc}}(x_t)$. The context-dependent representation c_t is then built by an autoregressive model, g_{ar}, which aggregates the latent representations: $c_t = g_{\text{ar}}(z_1, ..., z_t)$. Given the past context c_t, a predictor g_{pred} is asked to predict future representations z_{t+k} for $k \in \{1, ..., K\}$. The model is trained to maximize the categorical cross-entropy to correctly identify a positive future sample x_{t+k} from a set of unrelated negative samples. Formally, at step t, the loss function $\mathcal{L}_t$ for the pretext task is defined as follows:

$$\mathcal{L}_t = -\frac{1}{K}\sum_{k=1}^{K}\log\left[\frac{\exp(g_{\text{pred}}(c_t)_k^\top z_{t+k})}{\sum_{n\in\mathcal{N}_t}\exp(g_{\text{pred}}(c_t)_k^\top g_{\text{enc}}(n))}\right] \tag{1}$$

with $\mathcal{N}_t$ the set of negatives samples. The model is asked to predict up to $K = 12$ time steps in the future, equivalent to 120ms. The encoder g_{enc} comprises 5 one-dimensional convolutional layers with kernel sizes (10, 8, 4, 4, 4) and strides (5, 4, 2, 2, 2) and returns a 256-dimensional vector every 10 milliseconds. The auto-regressive model g_{ar} is a 2-layer long short-term memory network of dimension 256. The predictor g_{pred} is a single multi-head transformer layer with $K = 12$ heads, each predicting at time step $k \in \{1, ..., 12\}$. All models are trained for 100 epochs and the best epoch is selected according to the validation accuracy. For each independent dataset, 5 % of the data is used as the validation set. The other hyperparameters follow Kharitonov et al. (2021).

When training on speech, the negative samples $\mathcal{N}_t$ are drawn from within the same speaker. On the other hand, when training on ambient sounds, as there is no notion of speaker in this particular dataset, the negative samples are drawn from within the sequence. This is the only difference in the training process between the two approaches.

t-SNE visualization

To compute the t-SNE visualizations in panel **a)** of Figure 3, we first extract the audio representations of the American English Buckeye corpus. For each phone, we average the representations over time to get a single vector representation. Next, we apply the t-SNE method to reduce the 256-dimensional space into a 2-dimensional space. For the sake of clarity, only 1,000 randomly sampled phones for each category are displayed. For panel **b)** we apply the t-SNE method only on the representations of [ɹ]-[l] or [w]-[j] occurrences. Similarly, we display only 1,000 randomly sampled representations for each phonetic category.

Evaluated phonetic inventory

Table 1 shows the American English and Japanese phonetic inventory used in the ABX sound discrimination task and the t-SNE visualization.

Sonority	American English	Japanese
Fricative	f, v, θ, ð, s, z, ʃ, ʒ, h	ɸ, s, sː, z, ɕ, ɕː, ʑ, h
Affricate	ʤ, ʧ	t͡s, t͡sː, t͡ɕ, t͡ɕː
Plosive	p, b, d, t, k, g	p, pː, b, d, t, tː, k, kː, g
Approximant	w, j, ɹ, l	w, j, r
Nasal	m, n, ŋ	m, n, ɴ
Vowel	ɪ, iː, ɛ, ʌ, ɝ, æ, ɑː, ɔː, ʊ, uː, eɪ, aɪ, aʊ, ɔɪ, oʊ	ä, äː, e, eː, i, iː, o, oː, ɯ, ɯː

Table 1. ***Evaluated phonetic inventory in American English and Japanese in the International Phonetic Alphabet (IPA) standard (same as Schatz et al., 2021).***

Additional experiment: inducing initial speech sound discrimination capabilities through cross-lingual pretraining

Experiment 1 showed that it was possible to induce initial speech sound discrimination capabilities in our learner through pretraining on ambient sounds. Despite a better match between the learning trajectory exhibited by our learner and the observed data in infants, we could only simulate a maintenance trajectory on the [ɹ]-[l] contrast for the Japanese model. Experiment 2 showed that pretraining on multilingual speech did not yield higher initial speech sound discrimination capabilities than pretraining on ambient sounds.

In the present experiment, we ask whether the Japanese model can exhibit a loss trajectory on the [ɹ]-[l] contrast. We likely need to hypothesize even higher speech sound discrimination capabilities to do so. In this experiment, this is achieved through cross-lingual pretraining. Namely: we first pre-train models on either American English or

Japanese and train them on the language they have not been exposed to. This experimental protocol is akin to assuming near-perfect sound discrimination capabilities of American English contrasts by Japanese infants and near-perfect sound discrimination capabilities of Japanese contrasts by American English infants.

Arguably, such a protocol lacks ecological validity as: 1) it assumes different initial states for our American English and Japanese models; 2) it assumes near-perfect discrimination of English sounds for our Japanese model; and 3) near-perfect discrimination of Japanese sounds for our English model. However, this Experiment serves as proof that, being gifted with high enough initial sound discrimination capabilities, our Japanese model can follow a loss trajectory on the [ɹ]-[l] contrast, while maintaining a maintenance trajectory on the [w]-[j] contrast, similar to what is observed in infants (which was not shown in Experiment 1 and 2).

Methods

We use the exact same training sets, learner, evaluation sets, and evaluation protocol used in Experiment 1. The only difference is that we pretrain cross-linguistically instead of pretraining on ambient sounds. Our approach involves two distinct initial states for English and Japanese models. They are derived from the models trained on 500 hours of speech in Experiment 1. The two initial states consist of the model's weights after exposure to either 500 hours of American English or Japanese. We then train the English models starting from the Japanese weights and the Japanese models starting from the English weights.

Results and discussion

Figure 7 shows the trajectories taken by models with a cross-lingual pretraining compared to those without any pretraining. Two-way ANOVAs with factors nativeness and training language resulted in $p<.0001$ for each data quantity, indicating significant differences between the native and non-native models. We observe a slight negative native advantage when training on as little as 1 hour of speech. This arises from the fact that the initial state of the Japanese models, composed of weights from a model trained on 500 hours of English, performs slightly better on English than the initial state of the English models (and vice-versa). This negative native advantage reverses after exposure to 4 hours of speech.

After training on 500 hours of speech, pretrained models show identical ABX accuracies to those obtained by non-pretrained models. In other words, pre-exposure to another language does not harm or benefit the final performance obtained by the models. The cross-lingual pretraining yields different learning trajectories than those observed in Experiment 1. Here, we observe a loss trajectory for the non-native pretrained model (decreasing orange dashed line).

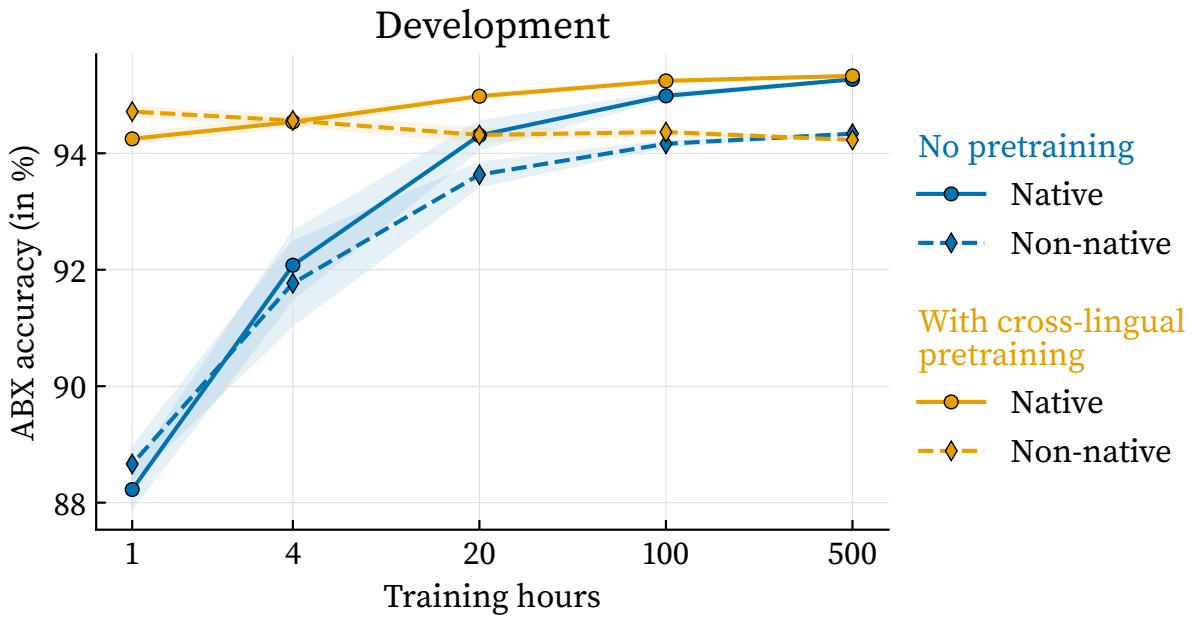

Figure 7. ***Comparison of our learner trained with no pretraining (in blue) or with cross-lingual pretraining (in orange) for native (same training and test language; solid line) and non-native (different training and test languages; dashed line) models as a function of the quantity of speech available in the training set (development). Shaded areas represent +/- the standard deviation computed across mutually exclusive training sets whose number depends on the quantity of data. Significance scores are obtained with a one-way ANOVA with factor training language (na: not applicable, ns: not significant, * p<.05, **, p<.001, *** p<.0001).***

Individual trajectories for the [ɹ]-[l] and [w]-[j] pairs

Figure 8 shows the trajectories on the [ɹ]-[l] and [w]-[j] contrasts for models that have not been pretrained (in blue) or pretrained cross-linguistically (in orange).

We will not repeat our interpretations of the trajectories taken by the unpretrained models (left column), which are the same results as those reported in Figure 4 and are left only for comparison.

In the pretrained condition, the American English model better discriminates the [ɹ]-[l] contrast than the non-native Japanese model. The American English model also successfully reproduces the facilitation trajectory observed in infants (Kuhl et al., 2006), similar to what has been observed when pretraining on ambient sounds in Experiment 1. Unlike what has been observed in Experiment 1, the Japanese model follows a loss trajectory, i.e., the performance on the [ɹ]-[l] contrast worsens as the quantity of speech increases.

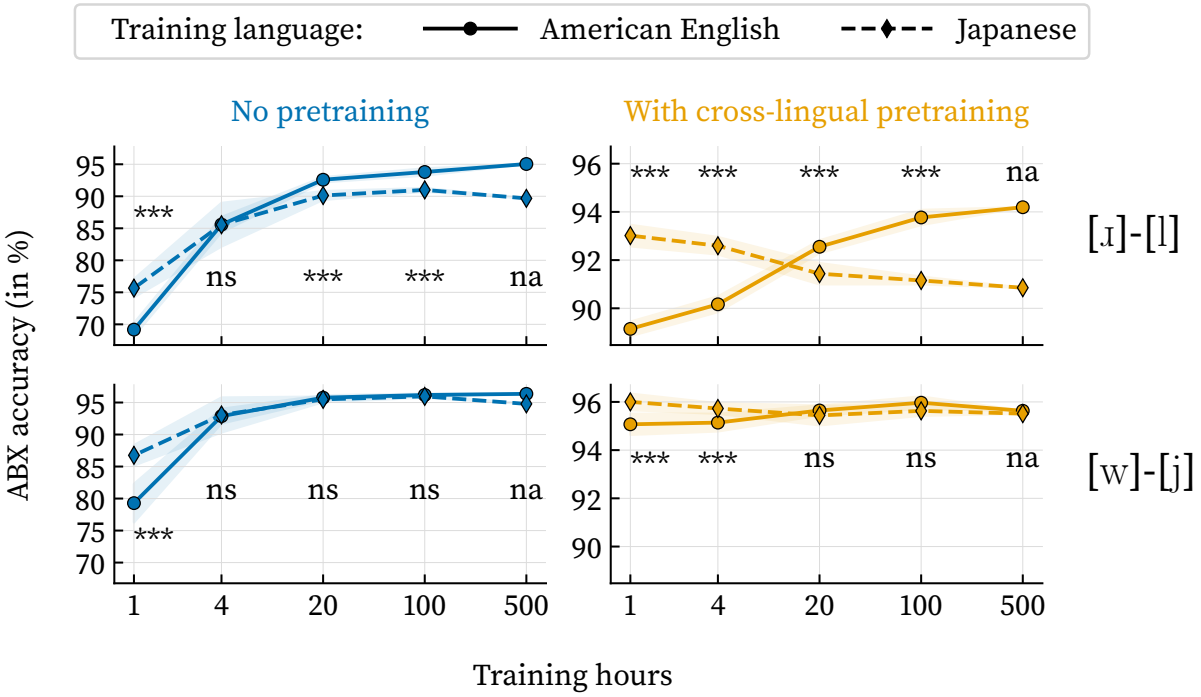

Figure 8. ***Comparison of the learning trajectory exhibited by our model without pretraining (in blue) or with cross-lingual pretraining (in orange) on the [ɹ]-[l] and [w]-[j] pairs. Models are trained on either American English (solid line) or Japanese (dashed line). [ɹ]-[l] and [w]-[j] occurrences are extracted from our American English evaluation sets. Shaded areas represent +/- the standard deviation computed across mutually exclusive training sets whose number depends on the quantity of data.***

On the [w]-[j] contrast, we now observe a maintenance trajectory instead of a facilitation trajectory in Experiment 1. In other words, after the cross-lingual pre-exposure phase, the discrimination performance obtained by our models has already converged on the [w]-[j] contrast. Performance does not benefit from further exposure to speech.

ABX sound discrimination accuracy

To enable comparisons, we provide the ABX accuracy obtained by models from Experiment 1, 2 and the additional experiment of the present study in Table 2.

Exp. #	Initial state	Training language	ABX accuracy in Japanese (CSJ / GPJ)	ABX accuracy in English (Buckeye / WSJ)
	MFCCs	–	90.8 / 91.2	87.5 / 93.4
1	No pretraining	–	65.0 / 65.3	60.4 / 58.3
1	Ambient sounds	–	92.7 / 93.5	89.5 / 94.0
2	Multilingual	–	92.8 / 93.9	90.6 / 93.6
1, 2	No pretraining	JP	96.1 / 95.5	92.0 / 94.7
1, 2	No pretraining	AE	95.2 / 95.5	93.0 / 96.4
1	Ambient sounds	JP	95.6 / 95.4	91.8 / 94.7
1	Ambient sounds	AE	94.3 / 94.9	92.1 / 95.5
2	Multilingual	JP	96.0 / 95.8	92.1 / 95.1
2	Multilingual	AE	94.6 / 95.2	92.6 / 95.8
3	Cross-lingual	JP	96.3 / 96.0	92.3 / 95.2
3	Cross-lingual	AE	94.5 / 94.9	92.7 / 96.3

Table 2. ***ABX accuracy (in %) on American English and Japanese evaluation sets.***

License