Infant phonetic learning as perceptual space learning: A crosslinguistic evaluation of computational models

Yevgen Matusevych, School of Informatics and School of Philosophy, Psychology and Language Sciences, University of Edinburgh, United Kingdom, ymatusevych@gmail.com

Thomas Schatz, Institute for Language, Cognition and the Brain, Aix-Marseilles University, France

Herman Kamper, Department of Electrical and Electronic Engineering, Stellenbosch University, South Africa

Naomi H. Feldman, Department of Linguistics and the Institute for Advanced Computer Studies, University of Maryland, USA

Sharon Goldwater, School of Informatics, University of Edinburgh, United Kingdom

Keywords: phonetic learning, computational modeling, perceptual space, language acquisition

Abstract

In the first year of life, infants' speech perception becomes attuned to the sounds of their native language. This process of early phonetic learning has traditionally been framed as phonetic category acquisition. However, recent studies have hypothesized that the attunement may instead reflect a perceptual space learning process that does not involve categories. In this article, we explore the idea of perceptual space learning by implementing five different perceptual space learning models and testing them on three phonetic contrasts that have been tested in the infant speech perception literature. We reproduce and extend previous results showing that a perceptual space learning model that uses only distributional information about the acoustics of short time slices of speech can account for at least some cross-linguistic differences in infant perception. Moreover, we find that a second perceptual space learning model which benefits from word-level guidance performs equally well in capturing cross-linguistic differences in infant speech perception. These results provide support for the general idea of perceptual space learning as a theory of early phonetic learning, but suggest that more fine-grained data are needed to distinguish between different formal accounts. Finally, we provide testable empirical predictions of the two most promising models and show that these are not identical, making it possible to independently evaluate each model in experiments with infants in future research.

1 Introduction

Infants' speech perception changes in the first year of their life. For example, at the age of 6–8 months, English-learning and Japanese-learning infants are equally able to detect the difference between sounds [ɹ] (as in *rock*) and [l] (as in *lock*), whereas by the age of 10–12 months, the two groups diverge, showing attunement to the phonetic contrasts present in their input language (Kuhl et al., 2006). Similar results have been reported for many other languages, such as Catalan (Bosch & Sebastián-Gallés, 2003), Zulu (Best & McRoberts, 2003), Mandarin Chinese (Tsao, Liu, & Kuhl, 2006), French (Burns, Yoshida, Hill, & Werker, 2007), Hebrew (Segal, Hejli-Assi, & Kishon-Rabin, 2016), etc. This process of attunement is known as early phonetic learning, and a number of existing theoretical accounts have been proposed to explain such learning (e.g., Best, 1994; Feldman, Goldwater, Dupoux, & Schatz, 2021; Kuhl & Iverson, 1995; Werker &

Curtin, 2005). At the same time, the exact mechanisms underlying early phonetic learning are not fully understood. One way to study such mechanisms is to implement them as computational models and then evaluate on data from experiments with infants. A computational implementation requires that each mechanism is specified in sufficient detail and guarantees that it can be easily tested on new data sets as they become available, ensuring the method's scalability (Cruz Blandón, Cristia, & Räsänen, 2021). Here, we adopt computational modeling to answer the question: Which computationally implemented mechanisms of early phonetic learning, if any, can correctly predict the existing crosslinguistic data on infants' phone discrimination?

Until recently, no computational models could explain how the specific speech input to which infants are exposed leads to the observed changes in those infants' discrimination of phonetic contrasts. In a recent study, Schatz, Feldman, Goldwater, Cao, and Dupoux (2021) presented such a computational model, which correctly predicted the documented cross-linguistic difference in infants' discrimination of [ɹ] and [l] after learning from unsegmented speech. They explicitly simulated the learning process for Japanese and American English infants by (separately) training their model on unsegmented multispeaker speech recordings either in Japanese or in American English. They then measured the trained models' ability to discriminate [ɹ] and [l] with the machine ABX task, a flexible measure of discrimination that can be applied to model representations in essentially any format. In this task, the Japanese model showed a significantly higher discrimination error than the American English model, a crosslinguistic pattern observed in 10–12-month-old infants (Kuhl et al., 2006). By testing the model on one phonetic contrast, Schatz et al. (2021) showed the *feasibility* of predicting crosslinguistic differences in phone discrimination by a model that applies distributional learning mechanisms to unsegmented speech data. For this purpose, a single phonetic contrast provides sufficient evidence. At the same time, the success of Schatz et al.'s approach calls for a more rigorous testing of their model, to determine whether this is a plausible model of early phonetic learning. Our first goal is to test whether Schatz et al.'s model can correctly predict crosslinguistic differences in infants' discrimination of other phonetic contrasts in other languages.

Our second goal relates to the particular model choice in Schatz et al. (2021). To our knowledge, this is the only model proposed in the literature that learns from unsegmented speech and has been shown to correctly predict crosslinguistic discrimination patterns. To simulate a learner capable

of handling realistic input, they selected a cognitively plausible model for unsupervised learning from speech, proposed in the context of engineering applications. At the same time, the model implements a relatively simple unsupervised clustering algorithm. Many other cognitively plausible models have been recently proposed in the context of engineering applications (e.g., Chung, Hsu, Tang, & Glass, 2019; Kamper, 2019; Kamper, Elsner, Jansen, & Goldwater, 2015). These models implement various versions of perceptual space learning from the speech signal, i.e., a process of transforming the acoustic similarity space, leading to changes of the distances between speech sounds (Feldman et al., 2021). Although Schatz et al. (2021) showed that early phonetic learning can be modeled in terms of such perceptual space transformations, perceptual space learning is not a single unified theory, and we need to understand which mechanisms better explain the existing data. Therefore, we test which learning mechanisms, as implemented in specific computational models, lead to results qualitatively matching infants' behavioral data.

To address these two issues, in Study 1 we apply the model of Schatz et al. (2021) to three crosslinguistic phone discrimination tasks grounded in infant studies from different languages. We consider Schatz et al.'s (2021) original American English data in order to reproduce the reported findings and two other data sets (Mandarin Chinese and Catalan) in order to determine whether the findings generalize to other contrasts and languages. We find that the model can correctly predict the crosslinguistic pattern for the Mandarin Chinese contrast, but not for the Catalan contrast. In Study 2, we consider four other models developed in the speech technology community; these models are all state-of-the-art extensions of the well-known autoencoder neural network (Kramer, 1991) commonly used in modeling statistical language learning (e.g., Jones & Brandt, 2020; Mareschal & French, 2017; Plaut & Vande Velde, 2017). We evaluate these four algorithms on the same three data sets, to study whether any of the algorithms can correctly predict the discrimination patterns for all the three contrasts, potentially providing a better model of infant phonetic learning than the one proposed in Schatz et al. (2021). Doing so allows us to gain insight into the kinds of representations and learning mechanisms that infants are likely to employ.

We find that one model (Kamper, 2019) shows infant-like crosslinguistic discrimination patterns for the same two contrasts as Schatz et al.'s model — the American English [ɹ]–[l] and the Mandarin Chinese contrast — while three other models appear less successful as models of early phonetic learning. As a result, we have two models — Schatz et al. (2021) and Kamper (2019) — that

substantially differ in their learning mechanisms, but make qualitatively identical predictions on the three target contrasts. An important implication of this result is that the existing discrimination data sets allow us to rule out some of the models, but are equally compatible with more than one model of early phonetic learning. Ideally, we would like to have a large collection of infants' phone discrimination data that would include a wide variety of phonetic contrasts across languages from multiple experiments. However, such experiments can be costly to run, and therefore, it is important to carefully select phone contrasts that are likely to yield crosslinguistic differences in discrimination. To help identify such contrasts, we further use our two "best" computational models to make predictions about discrimination difficulty of various contrasts. These predictions can further guide experiments with infants.

The models that we use learn from unsegmented speech data and do not rely on symbolic representations of phones (see next section for a relevant discussion). To better understand the representations that emerge in our models, we provide additional analyses of these representations, focusing on how well they match known phonetic categories. The results suggest that the phonetic representation space of either model cannot be easily separated into areas that would correspond to meaningful adult-like phonetic categories in a given language, although the information about such categories may be easier to access in the representation space of Schatz et al.'s (2021) model than in the space of Kamper's (2019) model.

Our study contributes to understanding the mechanisms of early phonetic learning, and perceptual space learning in particular, by providing a systematic crosslinguistic evaluation of five relevant computational models implementing such mechanisms, by generating concrete predictions using two of these models, and by analyzing these models' representations. In the following section, we briefly introduce various accounts on early phonetic learning, describe phone discrimination tasks that these accounts build on, and provide an overview of relevant computational work.

2 Background

2.1 Accounts of early phonetic learning

Theoretical accounts of infant phonetic learning (e.g., Best, 1994; Feldman et al., 2021; Kuhl & Iverson, 1995; Werker & Curtin, 2005) explain how infants move from language-universal to language-specific phonetic perception. Traditionally, these accounts have assumed that infants learn phonetic categories, and the emergence of such language-specific categories affects infants' ability to perceive differences between some phones in other languages. Recently, such accounts have been challenged (Feldman et al., 2021; McMurray, 2022), and an alternative view was proposed (Feldman et al., 2021): infants start by learning perceptual spaces, while category learning comes later in life. Under this view, infants' discrimination ability changes due to the transformation of the acoustic similarity space. While this account may contradict some commonly made assumptions in the existing literature, Feldman et al. (2021) explain that the discrimination data alone cannot provide sufficient evidence in favor of *category* learning, making both accounts equally viable. As a result, it is unclear whether infants have acquired phonetic categories before they start learning words, an important question in language acquisition literature.

Formal computationally implemented models can help us evaluate the existing explanatory theories (Robinaugh, Haslbeck, Ryan, Fried, & Waldorp, 2021) and better understand which account of phonetic learning is more viable (Cruz Blandón et al., 2021). With regards to the category learning account, there are no models that show how phonetic categories are acquired from naturalistic input to infants. Instead, most existing modeling work has focused on learning phonetic categories from highly idealized stimuli. To give a few examples, Vallabha, McClelland, Pons, Werker, and Amano (2007) and Feldman, Griffiths, Goldwater, and Morgan (2013) trained their models on the pre-computed values of simple features of vowel tokens (first two formants and duration), while McMurray, Aslin, and Toscano (2009) only used voice onset time of stop consonants. Such manual selection of phones and features makes the models' learning task much easier than the one infants face. Natural speech that infants are exposed to is very noisy and is not aligned along the acoustic dimensions relevant in a given language (see Feldman et al., 2021, for a relevant discussion). For example, statistical learning of the vowel length contrast in Japanese from individual phone duration values is not a trivial task due to overlapping distributions

(Bion, Miyazawa, Kikuchi, & Mazuka, 2013). As a result, when the above-mentioned models are evaluated on the kinds of data that better resemble naturalistic speech data, their ability to learn is drastically reduced (Antetomaso et al., 2017). This is why it is essential to show how computational models learn from uncurated speech, as has been done by, e.g., Miyazawa, Kikuchi, and Mazuka (2010); Miyazawa, Miura, Kikuchi, and Mazuka (2011); Nixon and Tomaschek (2021); Schatz et al. (2021). At the same time, some of these studies only evaluate models in terms of how well their representations match adult phonetic categories (Miyazawa et al., 2010, 2011), without looking at infants' data. To test the viability of the perceptual space learning account, it is important to test the models on actual behavioral data from infants (i.e., from discrimination tasks), as has been done by Nixon and Tomaschek (2021) and Schatz et al. (2021), without making assumptions about the underlying representations. Nixon and Tomaschek (2021) evaluate their model on the data from one language, German. Therefore, the only study that considers cross-linguistic phone discrimination data is Schatz et al. (2021), who tested one model on one phonetic contrast. Their model implements a particular version of the perceptual space learning account, which we introduce in the next section. At the same time, there is a variety of mechanisms that fit various versions of this account in principle (Feldman et al., 2021). In this article, we consider additional models and additional phonetic contrasts to further evaluate the account of early phonetic learning without phonetic categories. If such models are able to successfully predict crosslinguistic differences in infants' phone discrimination, this can help us better understand which mechanisms the infants are more likely to rely on.

2.2 Sources of information available to the learner

Typically developing infants can listen to the low-level speech signal and naturally learn from it using distributional learning mechanisms, i.e., by tracking the statistical distribution of phonetic variation (Maye, Werker, & Gerken, 2002). At the same time, there is evidence that 6–8-month-old infants can segment and recognize some word forms in the input (Bortfeld, Morgan, Golinkoff, & Rathbun, 2005; Jusczyk & Aslin, 1995; Jusczyk, Houston, & Newsome, 1999), and it has been shown that such *top-down* guidance can aid the process of phonetic learning (Feldman et al., 2013; McMurray, Danelz, Rigler, & Seedorff, 2018). Both of these strategies, which we call *bottom-up*

and *top-down*, respectively, have been implemented in computational models. Those computational models implemented phonetic category learning theories and asked which sources of information could help models converge on the correct set of categories for the training language.

Similar questions arise in the context of perceptual space learning. The computational learner of Schatz et al. (2021), whose setup we follow in Study 1, uses a relatively simple algorithm — Dirichlet process Gaussian mixture model — that clusters short slices of speech (frames) in an unsupervised way. This algorithm implements a bottom-up distributional learner, which only relies on the low-level information about the acoustic spectrum, or the distribution of energy across various acoustic frequencies. At the same time, some of the existing neural network models in the speech engineering literature implement the top-down strategy by exposing the learner to acoustic words or word-size units (e.g., Kamper et al., 2015; Thiollière, Dunbar, Synnaeve, Versteegh, & Dupoux, 2015). In Study 2, we employ such top-down models to simulate early phonetic learning and compare them to the model of Schatz et al. (2021), and to other, more closely matched, bottom-up models, in terms of their ability to predict crosslinguistic patterns of phone discrimination observed in infants. Therefore, our two studies can additionally inform the discussion on the usefulness of bottom-up and top-down strategies for infant phonetic learning.

2.3 Infants' phone discrimination data

Infants cannot be directly tested on tasks that require explicit instruction, and the field predominantly relies on data from phone discrimination tasks (Best & McRoberts, 2003; Bosch & Sebastián-Gallés, 2003; Burns et al., 2007; Kuhl et al., 2006; Segal et al., 2016; Tsao et al., 2006; Werker & Tees, 1984, etc.), using paradigms such as conditioned head turn (Kuhl, 1979) or habituation (Best, McRoberts, & Sithole, 1988). While there are many studies testing a group of infants on native and non-native contrasts (see, e.g., an overview for vowels by Tsuji & Cristia, 2014), there are far fewer studies which test two groups of infants (native and non-native) on the same phone contrast. We adopt the latter setup and focus on data sets from three such experiments, based on the availability of corresponding speech corpora for training computational models:[1]

[1]In principle there are other suitable contrasts to test based on infant data, for example, Thompson [k']–[q'], Hindi [t]–[t̪] (Werker & Tees, 1984), Zulu [ɬ]–[ɮ], [kʰ]–[k'] and [p]–[ɓ] (Best & McRoberts, 2003), but suitable corpora for these languages were either unavailable or difficult to obtain.

1. Kuhl et al. (2006) tested American English and Japanese infants on the English [ɹ]–[l] contrast. English- and Japanese-learning infants showed similar discrimination rates for synthesized [ɹa]–[la] stimuli at the age of 6–8 months, but English-learning infants showed higher rates at the age of 10–12 months. Similar findings are reported in Tsushima et al. (1994).
2. Tsao et al. (2006) tested Mandarin Chinese and American English infants on the Mandarin [ɕ]–[tɕʰ] contrast. Mandarin- and English-learning infants showed similar discrimination rates for synthesized [ɕi]–[tɕʰi] stimuli (commonly denoted in pinyin as *xi* and *qi*, respectively) at the age of 6–8 months, but Mandarin-learning infants were better at 10–12 months. This is also consistent with the results reported by Kuhl, Tsao, and Liu (2003).
3. Bosch and Sebastián-Gallés (2003) tested Catalan- and Spanish-learning infants on the Catalan [e]–[ɛ] contrast. At the age of 4 months both groups could discriminate between pseudowords [deði] and [dɛði] (in this case, stimuli recorded with human speakers), but at 8 months only the Catalan group showed successful discrimination. Similar results are reported by Albareda-Castellot, Pons, and Sebastián-Gallés (2011).

We use these three data sets to evaluate the computational models on their ability to correctly predict the described qualitative crosslinguistic patterns of phone discrimination. The next section provides methodological details of our simulation setup, data processing, and models.

3 Method

3.1 General setup of the simulations

We carry out two studies: Study 1 seeks to answer whether the model of Schatz et al. (2021) correctly predicts two other crosslinguistic data sets in addition to the English–Japanese result they report on. Study 2 tests four neural network models on the same three data sets. In each study we train computational models on unsegmented speech data from three data sets (i.e., language pairs). Each data set focuses on one phonetic contrast (such as American English [ɹ]–[l]) for which cross-linguistic phone discrimination data for infants exist. For each computational model, we train two different versions: a 'native' model, which simulates a learner of the language from which the contrast is drawn (American English, in this example), and a 'non-native' model, which

Table 1: Training and test conditions.

Data set	Test language	Training language	Listener type
1	EN	English (EN)	Native
		Japanese (JA)	Non-native
2	ZH	Mandarin (ZH)	Native
		English (EN)	Non-native
3	CA	Catalan (CA)	Native
		Spanish (ES)	Non-native

simulates a learner of another language that does not contain the relevant contrast (here, Japanese). Models are trained on corpora of natural speech. We then test each model by simulating a phone discrimination task, using real examples from the language where the contrast exists. To show an infant-like pattern, the 'native' trained version of the model should display better discrimination than the 'non-native' trained version of the model.

The training and test data sets are summarized in Table 1 and correspond to the experiments with infants described in Section 2.1. Data set 1 is designed to test models learning American English and Japanese on the English [ɹ]–[l] contrast, where English learners show better discrimination than Japanese learners. Data set 2 is designed to test models learning Mandarin Chinese and American English on the Mandarin [ɕ]–[tɕʰ] contrast, where Mandarin learners show better discrimination than English learners. Finally, data set 3 is for testing Catalan- and Spanish-learning models on the Catalan [e]–[ɛ] contrast, where Catalan learners show better discrimination than Spanish learners. In the experiments with infants, each phonetic contrast was tested in a particular phonetic context (e.g., [ɹa]–[la]). We report the results averaged over all phonetic contexts (to have sufficient test data for our models), but the results for the restricted contexts that were actually used in the experiments can be found in Supplementary Materials S2.

3.2 Simulating phone discrimination tasks

To test a model's ability to discriminate a phonetic contrast, similar to the tests carried out with infants such as conditioned head turn, we use the machine ABX task (Schatz et al., 2013).[2] In this task, A and X are two instances of the same phone (e.g., [l]), while B is an instance of a different

[2] https://github.com/bootphon/ABXpy

phone (e.g., [ɪ]). Note that while the order of item presentation (A followed by B or vice versa) is important in human experiments, it plays no role in our modeling setup. If A and X are closer to each other in a model's representation space than B and X, the model's prediction is correct, otherwise it is not. We use Kullback–Leibler divergence to measure distances in the representations for one of the models (DPGMM, see Table 2), and angular distance (which is similar to cosine distance, but not the same) for the other models. For models using frame-level representations (see Table 2), we align the frames in each pair of phones using dynamic time warping (Vintsyuk, 1968), and compute the distance as an average over the framewise distances. This method follows earlier studies (e.g., Jansen et al., 2013; Matusevych, Kamper, Schatz, Feldman, & Goldwater, 2021; Schatz, Bach, & Dupoux, 2018; Schatz et al., 2021).

A model is evaluated by considering the proportion of ABX triplets for which it makes correct predictions: 0% error rate corresponds to perfect discrimination, 50% to chance performance, and 100% means that for no triplets the discrimination was made correctly. Following Schatz et al. (2021), we sample ABX test triplets in such a way that all three phones — A, B, and X — appear in the same neighboring phonetic context and are uttered by the same speaker. This is a within-speaker version of the ABX task, which tests discrimination of phones produced by the same speaker and which we believe better matches the setup of most discrimination experiments with infants, compared to the across-speaker version. An alternative to using the machine ABX task would be to test the model on the exact test stimuli from the original experiments, but these are often synthesized. Using synthetic stimuli to test a computational model trained on natural speech could make the model perform poorly due to a confounding factor, the quality of mapping between synthetic stimuli and natural speech. Therefore, we only test whether a model can discriminate natural speech stimuli in a way similar to human infants.[3] Instead, we use the machine ABX task, which is a conceptual analogue of infant discrimination studies that is robust to the noise in the data given the large number of data points.

To test whether the difference between the ABX error rates in a given pair of simulated listeners (native vs. non-native) is significant, we fit mixed-effects regressions (using *lme4* package; Bates, Mächler, Bolker, & Walker, 2015) to the error rates of the two models in question. Each regression

[3]The Catalan test stimuli in Bosch and Sebastián-Gallés (2003) were recorded with human speakers, but we could not obtain the original recordings from the authors.

includes main effects of simulated listener type (native vs. non-native) and data subset (as shown in Table 3 in Section 3.4) and random intercepts, to account for the variation among data subsets, speakers and phonetic contexts. In the Mandarin data, tones often substantially change the pitch contour of the vowel, and we consider vowels (but not consonants) to be different if they come from syllables with different tones: e.g., in syllables [ɕi˥] and [ɕi˧˥] the target consonant [ɕ] is considered to be the same phone, but the right context is different: [i˥] vs. [i˧˥]. Significance for the effect of simulated listener type is then determined using two-tailed ANOVA tests (with Satterthwaite degrees of freedom approximation; Kuznetsova, Brockhoff, & Christensen, 2017) on the predicted values of the regressions.[4]

3.3 Computational models

In total we consider five models: the one used in Schatz et al. (2021) in Study 1, and four neural network models inspired by existing work in unsupervised speech representation learning in Study 2. All these models show high performance in low-resource speech technology applications, making them a good starting point for modeling unsupervised infant learning. In addition, at test time all the models provide a way to compute distances between speech sequences (in this case, phones) of any duration. The models differ along two dimensions, as summarized in Table 2. Three of the models learn representations at the level of speech frames (i.e., 25-millisecond-long chunks of speech commonly used in automatic speech recognition), while two learn to encode word-sized units of variable length as vector representations of fixed length (i.e., *acoustic word embeddings*, analogous to the semantic word embeddings often used as vector representations of word meaning). This distinction also corresponds to how the models deal with the time dimension: for the first three models, temporal information is encoded only weakly, by including first- and second-order derivatives of the acoustic spectrum in the representations of individual frames (see Section 3.4 below), while for the other two models the order of frame presentation is important too. In addition, three models are strictly unsupervised (i.e., bottom-up learners), while two others rely on top-down guidance from known word forms. In all cases, we use existing implementations developed for

[4]Note that this approach assumes independent data samples, which is not the case for our subsets of the Catalan corpus (as explained below in Section 3.4). This could potentially lead to underestimating variance and overestimating statistical significance.

Table 2: Models of early phonetic learning used in our studies. The DPGMM was used by Schatz et al. (2021), the other models have not been tested before in this capacity.

Model	Representation type	Top-down guidance
Dirichlet process Gaussian mixture model (DPGMM)	Frames	No
Autoencoder (AE)	Frames	No
Correspondence autoencoder (CAE)	Frames	Yes
Autoencoding recurrent neural network (AE-RNN)	Word-sized	No
Correspondence-autoencoding recurrent neural network (CAE-RNN)	Word-sized	Yes

processing speech without supervision or with a weak teaching signal from the word level (Kamper, 2019; Kamper et al., 2015; Schatz et al., 2021), we adopt the previously used training options, and we do not retune hyperparameters. A description of each model is provided in the sections for the individual studies below.

3.4 Input to the models

To prepare input to the models from unsegmented speech data, we follow a standard approach in speech processing: we divide the speech data into 25-millisecond-long frames (sampled every 10 milliseconds) and extract mel-frequency cepstral coefficients (MFCCs), together with their first- and second-order time derivatives, from each frame using Kaldi (Povey et al., 2011). The frequency range is set to be within the standard Kaldi values of 20 and 7,800 Hz (i.e., close to 0 and 8,000 Hz, respectively, where the latter value is the Nyquist frequency equal to half of our sampling rate of 16,000 Hz). The MFCCs encode the auditory spectrum for each frame, while the first- and second-order derivatives encode the change of this spectrum over time. The three types of features are concatenated, resulting in a vector of $13 \times 3 = 39$ features. Representing speech using its auditory spectrum is grounded in human auditory processing and is different from traditional accounts of phonetic learning, which assume phonetic feature detectors (see Schatz et al., 2021, for further discussion).

Additionally, for testing the models on the ABX discrimination of individual phones, as described in Section 3.2, we need to extract series of frames (i.e., 'chunks' of speech) corresponding to the target phones. To identify which frames correspond to which phone, we use phone alignments,

Table 3: Corpus samples used in the simulations. WSJ refers to the Wall Street Journal CSR corpus (Paul & Baker, 1992), GP to the Globalphone, a multilingual text and speech database (Schultz, 2002), Buckeye to the Buckeye corpus of conversational speech (Pitt et al., 2005), CSJ to the corpus of spontaneous Japanese (Maekawa, 2003), AIShell to the open-source Mandarin speech corpus (Bu et al., 2017), and Glissando to the corpus for multidisciplinary prosodic studies in Spanish and Catalan (Garrido et al., 2013). For the training data set 2, we used two different samples (2B and 2D) from the English WSJ corpus, to match the data available in the respective samples (2A and 2C) from the Mandarin AIShell and GP corpora. For the data set 3, we used two different training/test splits (3A/3E and 3C/3F) from the same Catalan corpus. For data set 1, all training/test combinations originated from different corpora and therefore were considered (i.e., 1A and 1B tested on 1E, 1A and 1B tested on 1F, etc.). For the other two data sets, to ensure that no data from the same speakers appeared both in the training and in the test data, some training/test combinations were excluded from the analyses: 2C and 2D were not tested on 2F; 3A and 3B were not tested on 3F; 3C and 3D were not tested on 3E. Rd and Sp stand for read and spontaneous speech registers, respectively.

(a) Training data.

Data set	Language	Sample	Corpus	Register	Amount of data (hh:mm)	No. of spk.
1	EN	A	WSJ	Rd	19:30	96
	JA	B	GP	Rd	19:33	96
	EN	C	Buckeye	Sp	9:13	20
	JA	D	CSJ	Sp	9:11	20
2	ZH	A	AIShell	Rd	58:59	166
	EN	B	WSJ	Rd	58:49	166
	ZH	C	GP	Rd	11:51	48
	EN	D	WSJ	Rd	11:49	48
3	CA	A	Glissando	Rd+Sp	7:41	26
	ES	B	Glissando	Rd+Sp	7:41	26
	CA	C	Glissando	Rd+Sp	7:02	17
	ES	D	Glissando	Rd+Sp	7:03	17

(b) Test data.

Data set	Language	Sample	Corpus	Register	Amount of data (hh:mm)	No. of spk.
1	EN	E	WSJ	Rd	9:39	47
		F	Buckeye	Sp	9:01	20
2	ZH	E	AIShell	Rd	58:45	165
		F	GP	Rd	11:51	48
3	CA	E	Glissando	Rd+Sp	1:15	2
		F	Glissando	Sp	2:19	11

i.e., labels that map series of frames to their corresponding phones. We obtain such alignments using the Montreal Forced Aligner (McAuliffe, Socolof, Mihuc, Wagner, & Sonderegger, 2017); lists of phones that we used are provided in Supplementary Materials S1.[5]

Ideally, we would use transcribed high-quality speech recordings of infant-directed speech. Such recordings may be available for English but are difficult to obtain for other languages. Because we needed matching samples (in terms of the register, number of speakers, etc.) to train the native and the non-native version of the model (e.g., English–Japanese, Mandarin–English), we used various other corpora in our simulations. In each case we train and test models on two different subsets of speech data per language, in order to ensure that the results for each model are robust across various data sets. Ideally, each subset should come from a different corpus, and the corpora should represent two different speech registers: spontaneous and read, which was the approach taken by Schatz et al. (2021). In practice, our choices are limited to the available speech corpora, so that for the Mandarin–English simulations (data set 2) we use corpora of read speech only, and for the Catalan–Spanish simulations (data set 3) all our data come from the same bilingual corpus (see Table 3). In all cases, we ensure that no data from the same speakers appear both in the training and the test data. To further reduce potential variability across corpora, we sample the audio signal in each corpus at 16 kHz and balance the speakers' gender within each corpus sample.

We now turn to our two studies that follow this methodological setup.

4 Study 1: Testing the DPGMM on other phonetic contrasts

In this study, we run the simulations described above using the model of Schatz et al. (2021). Training and testing the model on data set 1 here effectively reproduces their study, while data sets 2–3 enable us to test whether their model can correctly predict the discrimination patterns found in Mandarin- vs. English-learning and Catalan- vs. Spanish-learning infants. In other words, this study reproduces and extends the work of Schatz et al. (2021) to new phonetic contrasts.

[5]For English and Japanese in data set 1, we obtained the existing alignments (Schatz et al., 2021) generated with Kaldi (Povey et al., 2011). For the Catalan data, the transcription quality in the original corpus turned out to be low (confirmed in consultation with a native Catalan speaker), and we replaced word transcriptions with standard transcriptions for words available in Wiktionary (approximately 11.6% of the word types): `http://wiktionary.org`. This, however, did not change the models' qualitative patterns compared to our preliminary simulations with the data aligned using the original transcriptions.

4.1 Model description

The model is a Dirichlet process Gaussian mixture model (DPGMM; Chen, Leung, Xie, Ma, & Li, 2015) based on a commonly used Gaussian mixture clustering algorithm. Specifically, the DPGMM is a probabilistic generative model that takes individual speech frames as input and groups them into 'soft' clusters (Gaussian components), i.e., each frame can be assigned to multiple clusters with various probabilities. As described above in Section 3.4, each frame is represented as a vector that includes MFCC features and their time derivatives. A frame is considered to have been generated by a mixture of Gaussian probability distributions. Based on how likely each frame is to originate from each distribution, the model updates the mixture weights and the parameters of these distributions. More specifically, the model maximizes the likelihood of the data sample using Bayesian inference, specifically parallel Markov chain Monte Carlo (MCMC) sampling, following Chang and Fisher III (2013); Schatz et al. (2021). The model is non-parametric: i.e., the number of clusters is not specified in advance, but is derived from the data. It learns in a fully unsupervised bottom-up manner. The result of the learning process is a mixture of Gaussian components (clusters) fitted to the training data in the vector space of MFCC features and their derivatives.

For testing, we first represent each test frame into the same vector space as the training frames: each frame is encoded using the same MFCC feature extraction process as during the training phase (including the extraction of derivatives). We then can compute the probability of a test frame given each component. The model's output for each frame is, therefore, a vector of posterior probabilities with a size equal to the number of Gaussian components in the model, or a *posteriorgram*. A sequence of speech frames (in our case, a phone) is encoded in the model's representation space as a sequence of posteriorgrams. In this space, distances between phones (in our case, KL divergence) can be computed to perform the ABX discrimination task as described above in Section 3.2.

We use the DPGMM implementation[6] based on Chang and Fisher III (2013) and Schatz et al. (2021) and parameter settings from Schatz et al.: the model is initialized with 10 clusters and is trained for 1500 iterations, at which point there are between 348–1535 clusters in our simulations. This high number is consistent with Schatz et al. (2021), who found that the learned clusters are much more fine-grained than phonetic categories. The exact number of clusters depends on the size of the training data (i.e., larger data sets yield more clusters), the language (e.g., more clusters in

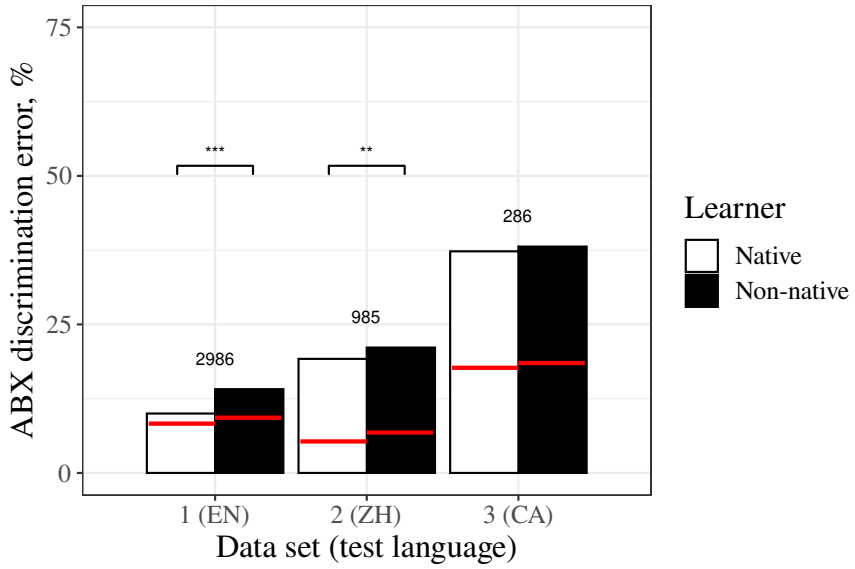


Figure 1: ABX error rates of the native and non-native DPGMM models in the three discrimination tasks (EN [ɹ]–[l], ZH [ɕ]–[tɕʰ] and CA [e]–[ɛ]). The number of data pairs (i.e., different speaker–phonetic context combinations) in each test set is shown on top of each bar. Red lines indicate model's error rates averaged over all consonant (for EN and ZH) or all vowel (for CA) contrasts, with the number of different contrast–speaker–context combinations varying between approximately 10K for CA to 53K for ZH to 260K for EN. To match the infant pattern of discrimination, the native model in each pair must show significantly lower error rates than the non-native model. The number of asterisks denotes significance level: *** corresponds to $p < .001$, and ** to $p < .01$.

models trained on Japanese vs. English), and potentially other factors that contribute to the amount of variation in the training data. We refer to Appendix and the two above-mentioned studies for a more detailed model description.

4.2 Results

The model's ABX error rates across languages are shown in Figure 1, together with the average performance of each model across all consonant (for English and Mandarin Chinese) or vowel (for Catalan) contrasts (red lines in the figure). In this figure, results are averaged over multiple ABX triplets, speakers, neighboring phonetic contexts, and subsets of the corpus, but the mixed-effects models fitted to the data take into account all of these variables (as previously discussed in Section 3.2). The reported patterns are consistent over the two corpus subsets. In what follows, we compare the performance of each simulated 'native' listener to its corresponding simulated 'non-

[6] https://github.com/Thomas-Schatz/perceptual-tuning-pnas

native' listener. Note that comparing the absolute error rates *across* the data sets (i.e., language pairs) is not meaningful, as the amount of training data and number of speakers differed depending on the data set, and interpreting the magnitude of the *differences* between native vs. non-native model across the data sets may not always be straightforward[7]. In data set 1, where the models are tested on the English [ɹ]–[l] contrast, the model correctly predicts the discrimination pattern observed in infants: the error rate of the simulated native listener is significantly lower compared to the simulated non-native listener (10.0% vs. 14.1%). This reproduces the result of Schatz et al. (2021). On this contrast, both native and non-native models show the discrimination error comparable to the average error rates on all English consonant contrasts (red lines in the figure, 8.3% and 9.3% for the native and non-native model, respectively). In other words, the [ɹ]–[l] contrast is only somewhat more difficult to discriminate than an average English consonant contrast.

In data set 2, with the Mandarin [ɕ]–[tɕʰ] contrast, the model also correctly predicts the infants' discrimination pattern: the error rates are significantly lower in the simulated native than non-native listener (19.2% vs. 21.1%). This difference of 1.9% is smaller than on the English contrast in data set 1 (4.1%), possibly because this is a generally difficult contrast to learn for the model. Indeed, the discrimination error on this contrast is noticeably higher than the average error over all Mandarin consonant contrasts (red lines, 5.3% and 6.8% for native and non-native model, respectively). In other words, the [ɕ]–[tɕʰ] discrimination is difficult for the model. This may be due to the kinds of phones in this contrast: one of them, [ɕ], is a fricative; the other, [tɕʰ], is an affricate, which sounds a bit like a combination of a short [t] followed by a 'breathy' version of [ɕ]. An example of a similarly sounding fricative–affricate distinction in English is the difference between the first phones in *cheap* vs. *sheep*. As a result, one of the phones, [ɕ], is almost a 'subchunk' of the other phone, [tɕʰ], a distinction potentially difficult to learn for our models. Nevertheless, the native model shows lower error rate than the non-native model.

In data set 3, the model predicts no significant difference for the Catalan [e]–[ɛ] contrast. In general, this contrast is more difficult for both native and non-native model (32.5% and 40.3% error, respectively) than an average Catalan vowel contrast (17.7% and 18.5%), and we discuss possible reasons for that below.

[7]This is because the discrimination error rates are expressed in percentages, and the true size of the target effect is not necessarily a linear function of the difference between percentages: i.e., the 5% difference in 50 – 45% vs. 6 – 1% likely corresponds to different effect sizes.

4.3 Summary

We have reproduced the result of Schatz et al. (2021) and also shown that their DPGMM model can correctly predict the cross-linguistic differences in infants' phone discrimination on another contrast from Mandarin Chinese. At the same time, the model struggles with predicting the infants' data for the Catalan contrast. On the one hand, this may be because of the smaller size or potentially lower quality of the Spanish–Catalan data set. On the other hand, there is a chance that the DPGMM model is simply not a good model of phonetic learning. For example, in the domain of phonetic category learning, bottom-up distributional models have been shown to perform poorly when trained on uncurated data (e.g., Antetomaso et al., 2017), and it has been argued that infants can use top-down guidance (i.e., word-level information) to constrain phonetic learning (Feldman et al., 2013; McMurray et al., 2018; Swingley, 2009). In the speech engineering literature, word-level information has been integrated into some of the neural network models, and in the next section, we test two such models, together with two corresponding models without top-down guidance, to see whether they can correctly predict the crosslinguistic patterns for the three target contrasts.

5 Study 2: Testing other models

In this section, we train and test four neural network models on the same three data sets as before. These models have been proposed in speech technology research, in particular in low-resource settings where transcribed data may not be available, and showed high performance in word and phone discrimination tasks (Kamper, 2019; Kamper et al., 2015; Matusevych et al., 2021; Renshaw, Kamper, Jansen, & Goldwater, 2015). Figure 2 schematically shows the difference between the models' architectures and input data. We consider two different versions of the models with top-down supervision (see the two panels on the right in Figure 2), which differ from each other in the kinds of representations they learn: frames vs. word-sized units. To be able to tell how much the word-level information is contributing, we also consider the corresponding versions of these models that have the same architectures but are trained *without* top-down guidance (see the two panels on the left in Figure 2). It is worth noting that these simpler versions have also been proposed as representation learning models in speech technology research in their own right, so it is possible that they will show good performance even without the word-level information. Below we provide

a brief description of each model, while formal definitions and parameter settings are summarized in Appendix.

5.1 Models' description

All four models are based on the idea of auto-associative learning: they are provided with an input and an identical (or a very similar) output and try to reconstruct the output from the input. While autoassociators do not implement a biologically plausible learning mechanism at the algorithmic level, they are commonly used in cognitive science for unsupervised representation learning. They use gradient descent to slowly adjust neural connections between layers in order to minimize a given reconstruction loss. We are primarily interested in the models with top-down supervision, but we first introduce one of the corresponding baseline models (i.e., without top-down supervision) to ease the understanding.

A basic version of the auto-associative learning mechanism is implemented in an **autoencoder** (AE; top left panel in Figure 2), a classic unsupervised feedforward neural network popularized by Kramer (1991). In our case, this is a 'stacked' autoencoder consisting of input, output, and multiple fully connected layers, see Appendix. While it is common to introduce a 'bottleneck' layer in the middle, which forces the model to compress the information, we use the implementation of Kamper et al. (2015)[8], who found no benefit of using such a layer in a word discrimination task. The model reconstructs the input frame X (the orange vectors in Figure 2, top left panel) and learns by minimizing the mean squared error between the original and the reconstructed frame. This is a fully unsupervised model that does not use any top-down guidance. The learning results in the emergence of latent representations in the model's hidden layers. Following Kamper et al.'s (2015) approach, we use the second-last layer for encoding our test data: for a given speech frame, we compute the vector encoding of the frame in that layer. A phone is then represented as a sequence of such vectors, and the distances (in this case, angular distances computed with dynamic time warping, as described in Section 3.2) between phones are computed for running the ABX discrimination task as described in Section 3.2.

In Figure 2 we see how the basic stacked AE can be extended along two orthogonal dimensions. First, the model can be trained in a slightly different manner, so that it reconstructs not the same

[8] https://github.com/kamperh/speech_correspondence

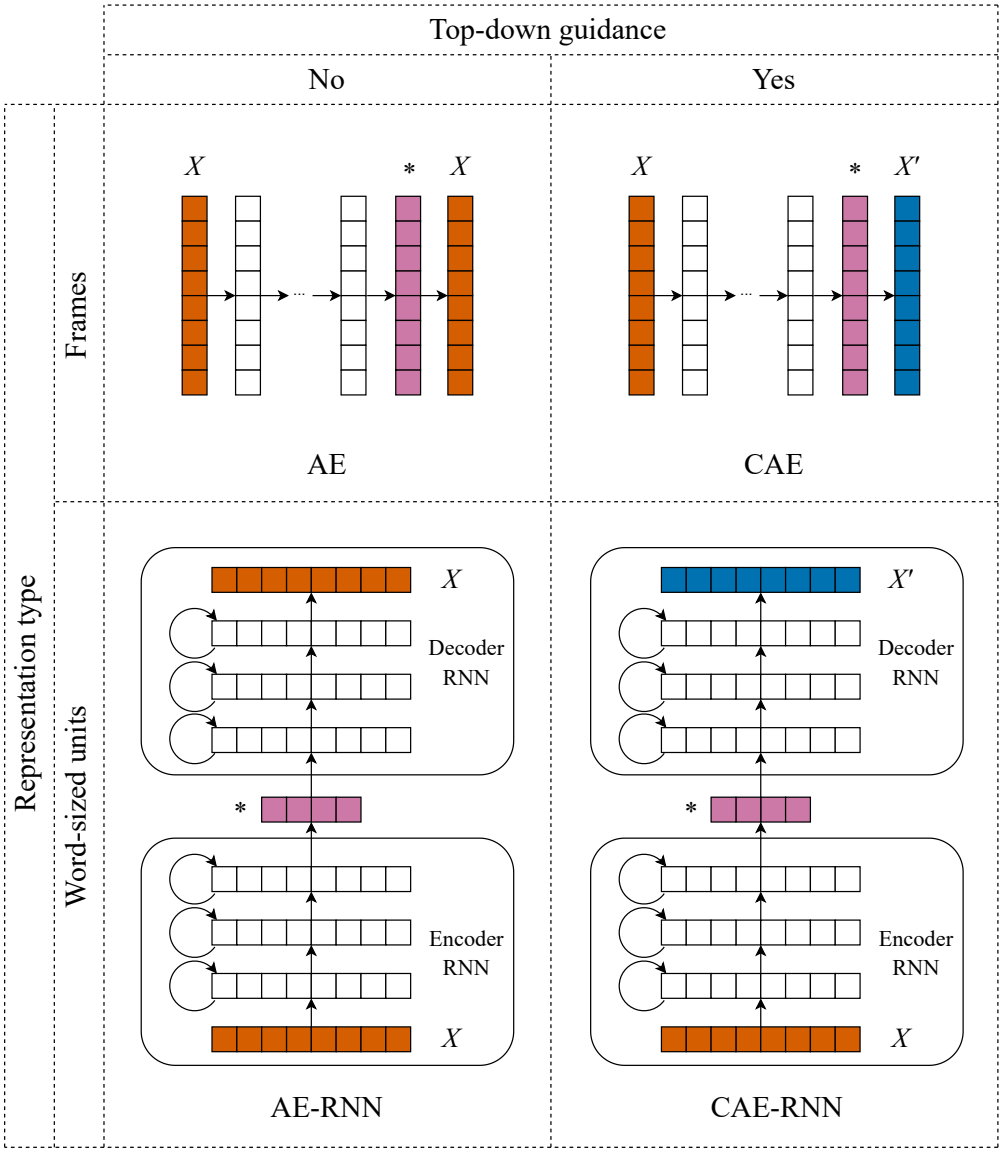


Figure 2: Neural network models used in Study 2. The AE and the CAE learn frame-level representations, while the AE-RNN and the CAE-RNN are recurrent models and learn word-sized representations. The AE and the AE-RNN are strictly auto-associative and reconstruct the input unit X itself (orange), while the CAE and the CAE-RNN reconstruct a different instance X' (blue) of the same type as the input unit X (orange). The layers from which we extract the models' representations are shown in pink and marked with an asterisk (*).

speech sequence (i.e., a spoken word), but another similar sequence (a different instance of the same word). This is the idea behind the **correspondence autoencoder** (CAE; Kamper et al., 2015, see top right panel in Figure 2). Instead of trying to encode and reconstruct each input frame to itself, as is done in the AE, it is given a pair of corresponding frames from two instances of the same word. The model tries to reconstruct a particular frame X' in one instance of a word from the aligned frame X in the other instance, one frame at a time (cf. top left and top right panels in Figure 2, where the blue color indicates a frame X' that is different from the orange frame X). Note that the two acoustic instances of the same word would normally have different duration (i.e., different number of frames), and the correspondence between frames across the two instances is established using dynamic time warping (Vintsyuk, 1968). In this article, we obtain the pairs of word instances in a supervised way using forced alignment (a by-product of phone-level alignments described in Section 3.4), though unsupervised alternatives are possible (Kamper, 2019). Because the model learns by reconstructing a speech signal into a different version of that signal, the encoded representation must focus on linguistically meaningful information and abstract away from other variation between the aligned frames. Importantly for us, the top-down guidance in the form of weak word-level supervision can aid the process of early phonetic learning, as we mentioned earlier. We use the same codebase as for the AE. Again, at test time each frame is encoded in the model's second-last layer, and the angular distances are computed in the resulting representation space.

Second, one can change the basic architecture of the AE model by turning it into a recurrent model capable of encoding sequential information. The bottom left panel in Figure 2 shows such a model, **autoencoding recurrent neural network** (AE-RNN; Chung, Wu, Shen, Lee, & Lee, 2016). It is a sequence-to-sequence autoencoder, a type of AE in which both the encoder and the decoder are recurrent neural networks (RNNs). RNNs are commonly used in language modeling (see Linzen, 2019, for an overview), as they can process an input sequence as a whole. In our case, the model is given a random word-sized chunk of speech, X, although not necessarily a real word, one frame at a time, encodes it into a vector of fixed dimensionality, and then uses this vector to reconstruct the same chunk sequentially, frame-by-frame (see bottom left panel in Figure 2, where the orange vectors represent the model's input/output frames). Here, we consider the model's middle, or acoustic embedding layer, in which speech sequences are represented as fixed-dimensional vectors. We encode the test phones into this embedding space and compute

angular distances between them to run the ABX task. We use the implementation by Kamper (2019).[9]

Finally, introducing both these changes at the same time yields our final model, a **correspondence-autoencoding recurrent neural network** (CAE-RNN; Kamper, 2019, see bottom right panel in Figure 2). This is similar to the AE-RNN, but instead of training on random chunks of speech, it is trained on pairs of instances of the same word (X and X' in the bottom right panel in Figure 2; note the two colors indicating different instances of the same word) — i.e., like the CAE, it also relies on weak top-down supervision (cf. top right and bottom right panels), and in our study the pairs of word instances were obtained in a supervised way using forced alignment (as for the CAE). We use the same codebase as for the AE-RNN and the same approach to compute the distances between phones.

Based on the differences between the four models, we can look for two patterns in the models' ability to predict the infants' data. As we pointed out in Section 3.3, because the AE and the AE-RNN do not require supervision from the word level, they are *bottom-up* models (cf. Table 2), whereas the CAE and the CAE-RNN receive additional *top-down* guidance. If the two latter, but not the former, models show infant-like discrimination patterns, it can be seen as additional computational evidence that top-down strategies can potentially be beneficial for phonetic learning (Feldman et al., 2013; Swingley, 2009). Also, the AE and the CAE learn frame-level representations and represent test phones as sequences of vectors, whereas the AE-RNN and the CAE-RNN encode word-sized units and represent each test phone as a fixed-dimensional vector. If the two latter, but not the former, models show infant-like discrimination patterns, this means that the holistic processing of longer units (word-sized during training, phone-sized during testing) may be beneficial for simulating phonetic learning. Note, however, that we only test specific implementations of the four models and do not change their architecture or retune hyperparameters, so the patterns we observe may not generalize to all instantiations of these models, and should be interpreted only as preliminary evidence. Most importantly, however, we are interested to know whether any of the models described above would show infant-like discrimination ability equal or better than the DPGMM model in Study 1.

[9] https://github.com/kamperh/recipe_bucktsong_awe_py3

5.2 Results

Figure 3 shows the ABX error rates of the four models. In data set 1 (the English [ɹ]–[l] contrast), two models — the CAE and the CAE-RNN — correctly predict the discrimination pattern observed in infants: the error rate of the simulated native listener is significantly lower compared to the simulated non-native listener (CAE: 5.8 vs. 8.2%, CAE-RNN: 13.4 vs. 19.2%). The AE shows no significant difference between the two types of simulated learners (6.1% for both), and the AE-RNN predicts a significant difference in the wrong direction (i.e., lower error rate in the non-native listener: 19.7 vs. 16.2%). As in Study 1, all models' error rates are comparable to the average error rates on English consonant contrasts (red lines in the figure). This suggests that the English [ɹ]–[l] contrast is generally easy to discriminate.

In data set 2 (the Mandarin [ɕ]–[tɕʰ] contrast), a different set of two models — the AE-RNN and the CAE-RNN — correctly predict the infants' discrimination pattern (AE-RNN: 26.6 vs. 30.0%; CAE-RNN: 25.9 vs. 30.5% for native vs. non-native listener). Two models — the AE and the CAE — predict no significant difference between the simulated native and the non-native listener (AE: 18.1 for both; CAE: 16.1 vs. 16.0). As in Study 1, over all models, the discrimination error on this contrast is noticeably higher than the average error over all Mandarin consonant contrasts, suggesting that the [ɕ]–[tɕʰ] discrimination is difficult (relative to other Mandarin consonant contrasts) for all models. Note, however, that on this contrast the CAE-RNN shows a difference of 4.6% between the native vs. non-native listener, which is larger compared to the DPGMM model in Study 1 (1.9%). This may indicate that the CAE-RNN makes more robust predictions on this contrast than the DPGMM, but the evidence is weak.

In data set 3, no model predicts a significant difference for the Catalan [e]–[ɛ] contrast (AE: 33.3 vs. 33.4%; CAE: 32.5 vs. 32.6%; AE-RNN: 40.3 vs. 39.5%; CAE-RNN: 39.6 vs. 37.2%). Note that the models' average error rates on Catalan are generally high, suggesting that the models could benefit from additional training data. At the same time, speaker idiosyncrasies in the test data are unlikely to affect the results, as we observe no meaningful differences in the discrimination error across the two test samples (consisting of data from 2 vs. 11 speakers, see Table 3). Thus, all models struggle to discriminate the Catalan [e]–[ɛ] contrast, as well as to reproduce empirically observed cross-linguistic differences in its discrimination.

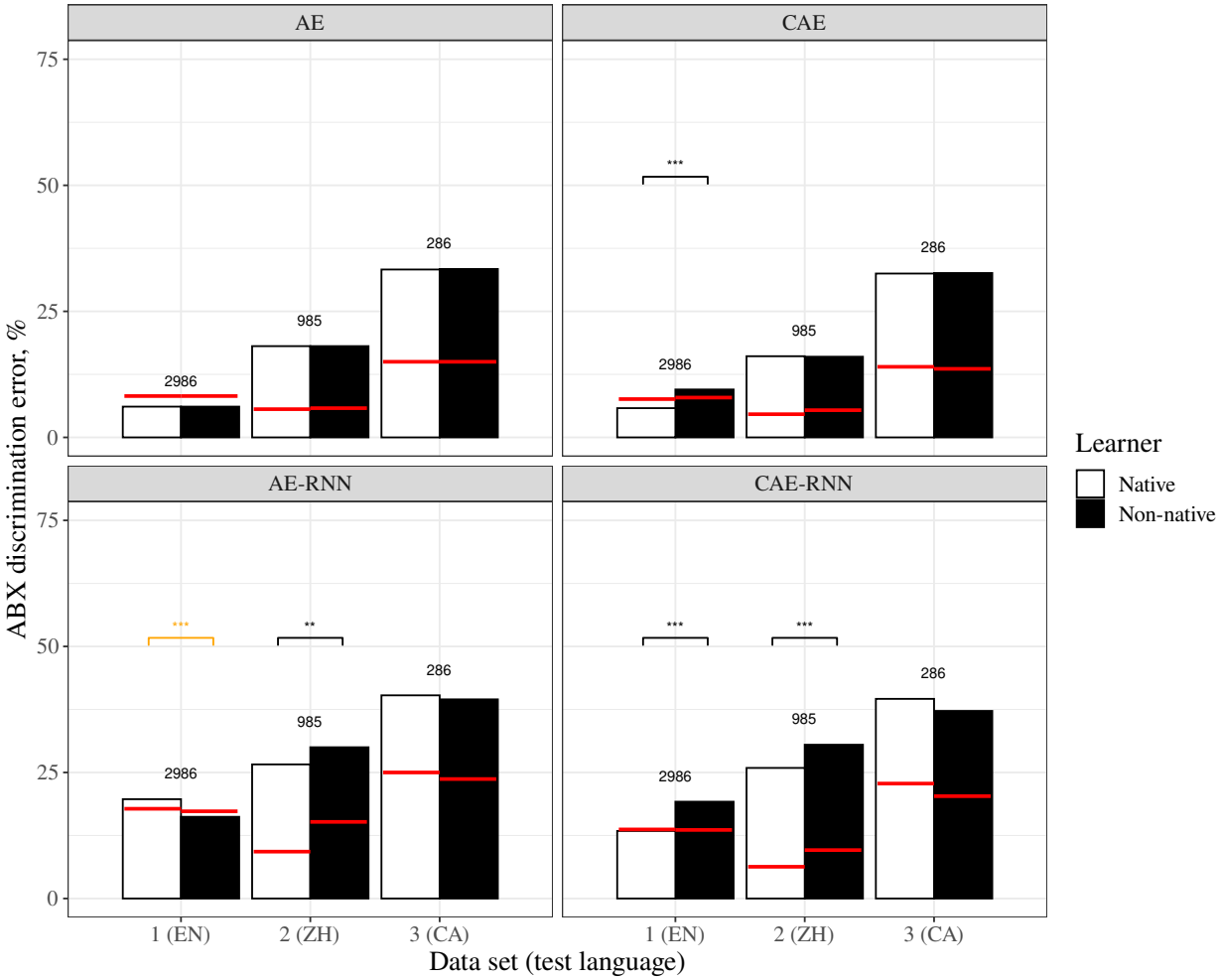


Figure 3: ABX error rates of the native and non-native neural network models in the three discrimination tasks (EN [ɹ]–[l], ZH [ɕ]–[tɕʰ] and CA [e]–[ɛ]). The number of data pairs (i.e., different speaker–phonetic context combinations) in each test set is shown on top of each bar. Red lines indicate models' error rates averaged over all consonant (for EN and ZH) or all vowel (for CA) contrasts. To match the infant pattern of discrimination, the native model in each pair must show significantly lower error rates than the non-native model: out of 5 total patterns with a significant difference, 4 are in the predicted direction (black brackets) and 1 is in the wrong direction (orange bracket). The number of asterisks denotes significance level: *** corresponds to $p < .001$, and ** to $p < .01$.

At the end of Study 1, we considered a possibility that the DPGMM model did not make correct predictions on the Catalan data because it might not be a good model of phonetic learning. If that was true, the results from this study could suggest that *none* of our models make good models of early phonetic learning. But given their correct predictions on the English and the Mandarin contrasts, it is worth considering other explanations of the models' incorrect predictions on the Catalan contrast and overall high discrimination error rates. To determine possible reasons why this contrast was particularly difficult for the models, we carried out additional analyses. First, we looked at various subsets of the test data controlled for the neighboring phonetic context, yet still did not find infant-like discrimination patterns in any model (see Supplementary Materials S2). Second, we looked at the duration of the target [e] and [ɛ] vowels in the test data, which revealed that some instances were very short, compared to the lab stimuli used with infants. Because we could not obtain the original stimuli of Bosch and Sebastián-Gallés (2003) from the authors, we instead filtered out very short (< 80 milliseconds) phones from the test data. This reduced the overall error rates, but still yielded similar performance between the 'native' and 'non-native' models. Third, we asked whether the target contrast can be learned from the training data at all. As an upper-bound baseline, we trained and tested a supervised phoneme recognizer model (see Appendix) on the same data, and the error rates for the target contrast were still high, although somewhat lower than for our models, 31.5 ± 2.8%. This suggests that either the target contrast is very difficult to learn from this data set, or that the test data is noisy. At the same time, a Spanish phoneme recognizer model that we trained on the same Spanish data showed significantly higher error rates on the Catalan contrast, 41.4 ± 0.2%, suggesting that a supervised model can correctly predict the infants' phone discrimination pattern.

To summarize, none of the models could capture all three crosslinguistic discrimination patterns. At the same time, the CAE-RNN, just like the DPGMM in Study 1, correctly predicts two patterns out of three. The CAE and the AE-RNN only predict one pattern each, while the AE makes no correct predictions. On the one hand, because the CAE-RNN and the DPGMM make equally good predictions, comparing just these two models does not let us conclude which of their mechanisms better explain infants' phone discrimination. On the other hand, recall that our four neural network models are matched on two dimensions, and we can look more closely at how the models' predictions differ along those dimensions. If we first compare the two models with the top-down strategies (the

CAE and the CAE-RNN) to the two corresponding models without such strategies (the AE and the AE-RNN), we can see that the top-down strategies can be beneficial for phonetic learning. Second, if we compare the two models which process word-sized units holistically (the AE-RNN and the CAE-RNN) to those that do not (the AE and the CAE), we find that the holistic processing can potentially benefit the models' predictions as well. These results, however, should be interpreted with caution, as there is no guarantee that similar patterns would replicate in other classes of models: i.e., designing a mechanism to provide the DPGMM with word-sized units would not necessarily improve its predictions.

The two models that perform best in our two studies — the DPGMM and the CAE-RNN — use very different learning algorithms and representation formats, effectively presenting two alternative hypotheses about early phonetic learning, yet they make qualitatively identical predictions regarding the crosslinguistic discrimination of three phone contrasts. Both hypotheses have been argued for in the literature on early phonetic learning. The DPGMM embodies a purely distributional bottom-up learning account (see Schatz et al., 2021, for a detailed account), while the CAE-RNN brings in the top-down guidance from the word level (Swingley, 2009). Therefore, our results on the English and the Mandarin contrast are compatible with both theories. But because these are contrasting theories in the acquisition literature, the field could benefit from a method that could distinguish between them computationally. This is why in the next section we use the two models — the DPGMM and the CAE-RNN — to make predictions about the difficulty of discrimination of specific phone contrasts which have not yet been tested with infants.

6 Models' predictions for other phone contrasts

In this section, we identify phone contrasts for which the two models — the DPGMM and the CAE-RNN — make different crosslinguistic predictions in the discrimination tasks. We follow the general method of Schatz et al. (2021) for deriving models' predictions. We only focus on data set 1 — that is, English and Japanese models tested on English phone contrasts.

We first split our English and Japanese training corpora into 10 parts and train the DPGMM and the CAE-RNN on $1/10^{\text{th}}$ of the data. As in our Study 1 and 2 above, for each model we compute ABX discrimination scores for all English contrasts in the native (English) and the non-native

Table 4: Contrasts for which the DPGMM and/or the CAE-RNN predict robust crosslinguistic differences in the discrimination difficulty. All contrasts are predicted to be easier to discriminate for the English learner than for the Japanese learner.

DPGMM		CAE-RNN	
Contrast	Mean difference	Contrast	Mean difference
[n]–[ɹ]	4.9	[f]–[z]	6.9
[d]–[ɹ]	4.9	[ʌ]–[ɔʊ]	5.8
[ɝ]–[l]	4.9	[f]–[s]	5.5
[ɝ]–[ɹ]	4.9	[l]–[ɹ]	4.8
[h]–[ɹ]	4.6	[m]–[ɹ]	4.5
[ʌ]–[ɝ]	4.4	[ɹ]–[w]	4.5
[m]–[ɹ]	4.5	[ʌ]–[ɑʊ]	4.3
[ð]–[ɹ]	3.8	[ɑ]–[ʌ]	3.0
[l]–[ɹ]	3.7		
[ɹ]–[v]	3.4		
[ɹ]–[t]	2.6		

(Japanese) versions of that model. We then compute the differences between the ABX scores of the native vs. non-native version on each English phone contrast, and look for the contrasts with robust crosslinguistic discrimination differences. A difference is considered robust when it is (1) statistically significant from zero across the ten data subsets for each corpus, (2) in the same direction across the two training/test data registers. These two criteria are based on the method of Schatz et al. (2021), but are somewhat relaxed compared to theirs, to ensure we obtain robust predictions from both models. To give an example, the [ɹ]–[l] contrast should be more difficult to discriminate for the Japanese model than for the English model when trained and tested on read speech as well as on spontaneous speech, and this difference in the discrimination difficulty should be statistically significant, in order for the difference to be considered robust.

We report the contrasts with such robust differences in Table 4. Note that both models predict that some contrasts are easier to discriminate for the native (English) than non-native (Japanese) learner, and no contrasts are predicted to have robust differences in the opposite direction. For the DPGMM, we see that the robust differences are only detected for contrasts involving the rhotic consonant [ɹ] or the rhotacized vowel [ɝ]. This is less so for the CAE-RNN: while its predictions also include three contrasts with [ɹ] (but not [ɝ]), there are also contrasts with fricative sounds

([f]–[z] and [f]–[s]), as well as vowel contrasts, all of which involve [ʌ] (none of these contrasts are phonemic in Japanese). First, the models' predictions can be directly used to inform future research on early phonetic learning, as infants can be directly tested on the contrasts for which our models predict crosslinguistic differences, and the results of those tests can be further used to understand which model is a better model of early phonetic learning. In particular, if Japanese-learning infants would show significantly worse discrimination rates than English-learning infants on contrasts with the rhotacized vowel [ɝ] (but not on fricative consonant contrasts or the vowel contrasts with [ʌ]), this would speak in favor of the DPGMM model. The reverse pattern — i.e., difficulties with fricatives and [ʌ], but not [ɝ] — would speak in favor of the CAE-RNN model. Note, however, that the absence of a particular phonetic contrast in Table 4 does not mean that the model predicts *no* difference for that contrast, but rather no *robust* difference. In other words, if Japanese-learning infants show more difficulties on, e.g., [ʌ]–[ɔʊ] contrast (which only appears among the CAE-RNN predictions) than English-learning infants, this does not necessarily rule out the DPGMM model.

Second, these results show that the two models do not make identical predictions on various phone contrasts. This is a positive result, because, as we mentioned above, experiments with infants can help us distinguish between the two models.

At the computational level, an analysis of the models' representations can help us understand what leads to different predictions. Below, we focus on comparing the models' representations to adult-like phonetic categories. Recall that the models we are using simulate perceptual space learning, a framework that challenges the existing accounts which attribute infants' behavior to phonetic *category* learning. This is why we are interested to know whether the representations that the two models learn contain readily accessible information about adult-like phonetic categories. We address this question in the next section.

7 Analyzing models' representations

We have considered five models that implement various versions of perceptual space learning. Because these models do not rely on symbolic representations of phones, it would be helpful to know what kind of representations they learn. Here, we analyze the representations that emerge in the DPGMM and the CAE-RNN, the two models that correctly predicted two out of three

crosslinguistic patterns of phone discrimination in our two studies. Representation analysis methods are commonly used to study not only neural networks, but also phone encoding in human brain (e.g., Evans & Davis, 2015; Levy & Wilson, 2020; Reh, Hensch, & Werker, 2021).

The DPGMM representations have been earlier studied by Schatz et al. (2021), who showed that the individual Gaussian components in the model do not correspond to phonetic categories: the number of components was an order of magnitude larger than the number of phonetic categories in English, and each component was on average activated only for under 20 milliseconds, much shorter than the average duration of a phoneme. It is an important result that the models of perceptual space learning, such as the DPGMM, do not use explicit phonetic categories yet make correct predictions about infants' phone discrimination. However, it is possible that categories may still be found implicitly in the model's posteriorgram space. Here, we look for structure in the representation spaces of both the DPGMM and CAE-RNN, by running unsupervised clustering and supervised classification of phone instances encoded in the two models' representation spaces, as explained below in the respective sections.

For both clustering and classification, we consider the DPGMM and the CAE-RNN model trained on a specific corpus (e.g., WSJ) and use a test sample from a *different* corpus in the same language (here, Buckeye for English). From the test corpus, we sample 100 acoustic realizations of each phone and remove phones which appear fewer than 100 times in the corpus. We consider the model's representations for that sample and compute pairwise distances (KL divergence for the DPGMM and angular distance for the CAE-RNN, for consistency with our ABX discrimination simulations) between all phone instances. These distances are then used in both clustering and classification tasks as described below. Note that this is different from the comparison of crosslinguistic discrimination patterns in Study 1–2. First, here we only consider models trained and tested on the same language: English, Japanese, or Mandarin. Second, we only consider models trained and tested on samples from different corpora, to exclude the possibility that our models learn some corpus-specific properties (e.g., register or channel effects) that could make clustering and classification easier. For this reason, we do not consider Spanish and Catalan models here, for which we only have a single corpus (as shown in Table 3 above). At the same time, we do provide the relevant results for these two languages in the Supplementary Materials S3.

As an upper-bound baseline, we additionally train and test a supervised phoneme recognizer

(see Appendix) on the same data sets to analyze its representations using an analogous setup.

7.1 Clustering

Cluster quality. First, we examine whether phone instances in each model's representation space naturally group into clusters that align with adult-like phonetic categories. We split the data described above in Section 7 into 10 equal parts (10 acoustic realizations per phone) for robust estimation and, using the distance matrices between phone instances, we run an agglomerative hierarchical clustering algorithm, with the number of clusters equal to the true number of different phones in the sample. Following some of the existing studies that compare unsupervised clusters to phonetic categories (Frank, Feldman, & Goldwater, 2014; Shain & Elsner, 2019), we use three information-theoretic measures ranging from 0 to 1 to evaluate the cluster quality: homogeneity (H), completeness (C), and V-measure (V). Homogeneity is the highest when each cluster only has phones of the same type, completeness is maximized when all phones of the same type are put in the same cluster. V-measure is a harmonic mean of the two (Rosenberg & Hirschberg, 2007).

High clustering performance would mean that the phonetic categories are easily separable in a model's representation space. Our results (see Figure 4) for the two target models of perceptual space learning are lower compared to the supervised model (which we consider to be an approximate upper-bound baseline). In other words, the clustering performance is far from perfect in all cases and is not as good as the performance of the supervised model, suggesting that adult-like phonetic categories may not be the most natural clusters in the models' learned phonetic spaces. On average, the clustering performance for the DPGMM-based representations is somewhat higher than for the CAE-RNN-based representations (except for the Japanese and the Mandarin models tested on GlobalPhone). A mixed-effects linear regression model fitted to the data showed that the described differences are statistically significant on all three measures: i.e., on average the supervised model has the highest cluster quality, followed by the DPGMM, followed by the CAE-RNN (except that the difference between the DPGMM and the CAE-RNN using C measure was not significant).

Cluster-to-phone mappings. For illustration purposes, we also show a confusion matrix between the true phonetic category labels and the unsupervised clusters, computed on the full data set with 100 acoustic realizations per phone. Note that there is no immediate correspondence

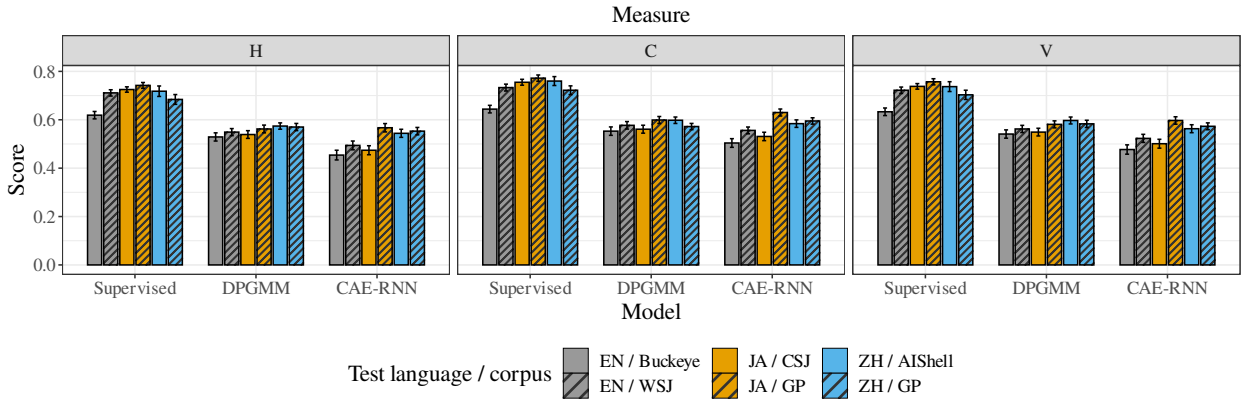
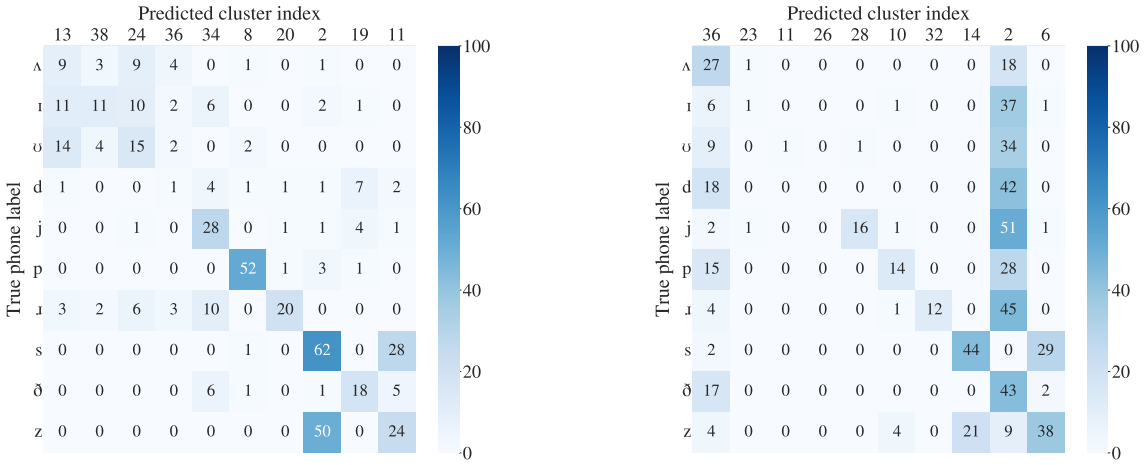


Figure 4: Quality of the unsupervised clusters, found by clustering the representations in the two most successful models (DPGMM and CAE-RNN) and the supervised baseline in English, Japanese, and Mandarin. Plots display homogeneity (left), completeness (middle), and V-measure (right) averaged over the 10 data splits, error bars show standard error of the mean.

between a set of true phonetic category labels and a set of unsupervised clusters, and we need to find the best assignment between true labels and unlabeled clusters. One classic method commonly used for this purpose is the Hungarian algorithm, a global combinatorial optimization method which solves the assignment problem in polynomial time (Kuhn, 1955). We present the resulting confusion matrix for English (trained on WSJ, tested on Buckeye). Figure 5 shows fragments of the matrices, while full matrices can be found in Supplementary Materials S4. Comparing the diagonal values in Figure 5a vs. 5b, we can see that the alignment between the true and the predicted phone labels is better for the DPGMM-based representations, where confusions are observed between similar sounds (e.g., vowels or fricative consonants). By contrast, the CAE-RNN-based representations have at least two clusters (with indices 36 and 2) whose boundaries cross many phones with very different acoustic characteristics (e.g., vowels [ʌ] and [ʊ], but also consonants [p] and [ð]). This difference is also noticeable in Figure 6, which shows that the label assignment to clusters is more accurate for the DPGMM than the CAE-RNN representations. This is consistent with the results in terms of the H, C, and V measures above.

Summary. Overall, our analyses show that the clusters computed on top of the CAE-RNN representations have lower quality than the clusters computed on the DPGMM posteriorgrams: they are less homogeneous, less complete and align worse with true phonetic categories. Recall, however, that we used different distance measures for the two models, following our main simulations using ABX discrimination task — KL divergence for the DPGMM vs. angular distance for the CAE-

Figure 5: Fragments of confusion matrices between the true phone labels vs. best matching clusters on English Buckeye data. The clusters are obtained using unsupervised agglomerative clustering on each model's representations of 100 instances for each English phone, and the matching is done by solving the assignment problems between true categories and predicted clusters using the Hungarian Algorithm. See Supplementary Materials S4 for full matrices.



(a) DPGMM

(b) CAE-RNN

RNN — and the latter may be less suitable for clustering. To see whether this is the case, we applied the same clustering algorithm to the CAE-RNN representations using the Euclidean distance measure. The clusters that include very different phones (such as [ʌ] and [ð] in Figure 5b) no longer occurred, but the overall cluster quality was not substantially higher.[10] The CAE-RNN confusion matrix computed with Euclidean distances is provided in Supplementary Materials S5 for reference.

7.2 Classification

The goal of our unsupervised clustering analysis was to test whether each model's representations of phone instances naturally group into classes corresponding to adult-like phonetic categories. Here, we test if — and to what extent — such phonetic categories can be inferred from each model's representations by an algorithm explicitly trained on this task. For this, we train a linear

[10]To ensure that our main ABX results in Study 2 were not affected by the choice of the distance measure for the CAE-RNN model, we ran the discrimination experiments with this model using the Euclidean distance measure (KL divergence could not be used, because it requires each representation to be a valid probability distribution), and the main patterns of results for out target phone contrasts in all languages did not change; see Supplementary Materials S5.

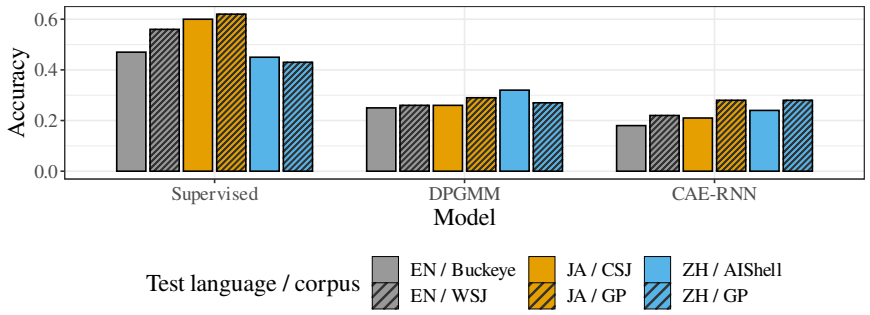


Figure 6: Accuracy of the mapping between true phone labels and the best matching clusters in English, Japanese, and Mandarin, computed using the Hungarian Algorithm.

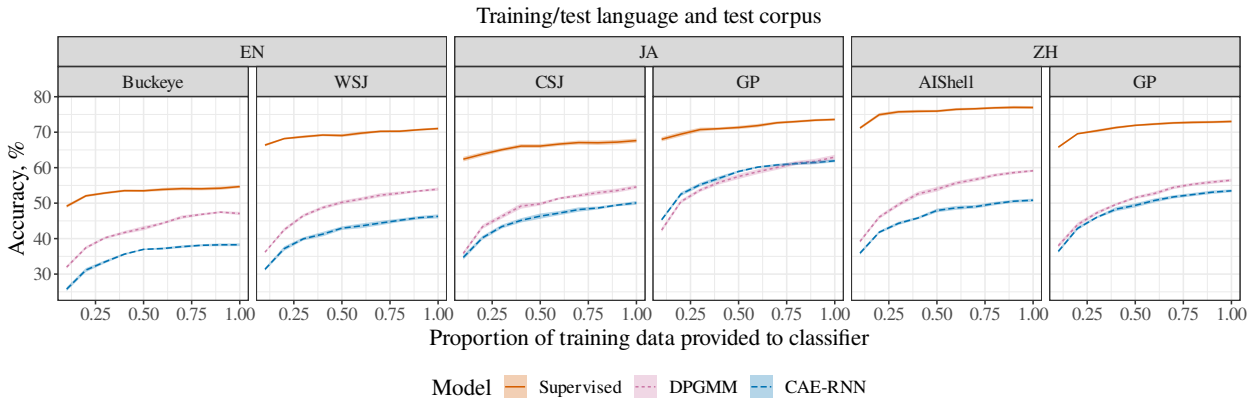


Figure 7: Accuracy of the k-NN phone classifier trained and tested on the representations of phone instances in the DPGMM and the CAE-RNN model for English, Japanese, and Mandarin. Bands show the standard error of the mean for 10 random training–test data splits.

k-nearest neighbors classifier (we use $k = 10$) using the precomputed distances between phone representations in each model described above. For each distance matrix, we use an 80–20% split (for each phone) for training and testing, respectively. We always use the full test data (i.e., 20%) for evaluating the classifier, but variable amounts of data for training, between 10 and 100% of the all training data instances, in increments of 10%, and look at how much data is needed for achieving high classification accuracy. The less data is needed, the more readily available the phone categories are in the model's representations.

Figure 7 shows that the classifier trained on the representations of the supervised model, unsurprisingly, achieves the highest accuracy on all corpora (the solid orange lines), between 54.6 and 77.0%. Moreover, the lines are not steep: the classifier trained on the supervised models' data achieves near-ceiling accuracy already after seeing only 20% of the data. This is not the pattern

we see for the classifier trained on the representations of the two perceptual space learning models: classification accuracy keeps increasing with more training data, suggesting that the information about phonetic categories is not readily available in the representations of the DPGMM and CAE-RNN.

Next, DPGMM representations (dotted pink lines) nearly always result in higher classification accuracy than the CAE-RNN representations (dashed blue lines). The only exception is the Japanese GlobalPhone corpus, where the two models yield nearly identical classification accuracy (the dotted pink and the dashed blue line closely follow each other). This pattern of results suggests that the information about the phones is more readily available in the DPGMM than CAE-RNN representations. To summarize, our clustering and classification results are consistent in providing evidence that the DPGMM encodes more information about phonetic categories than the CAE-RNN.

8 Discussion

8.1 Phone discrimination in computational models

Using computational modeling on realistic input, we compared possible models of early phonetic learning in their ability to predict the changes in discrimination empirically observed in infants. In the first study, we tested Schatz et al.'s (2021) DPGMM model on three phone contrasts from different languages, using a phone discrimination task. We first reproduced their result for the crosslinguistic discrimination of the English [ɹ]–[l] contrast, and then found that their model also shows the infant-like pattern of discrimination for the Mandarin [ɕ]–[tɕʰ] contrast. This means that their earlier result was not specific to a particular English contrast. In the second study, we tested four neural network models. One of these models, the CAE-RNN, also made correct predictions on the same two contrasts as the DPGMM. Although no model predicted the correct pattern for the Catalan vowel contrast [e]–[ɛ], the fact that two of them make correct predictions on the other two contrasts is promising. This result supports the idea that models learning perceptual spaces directly from unsegmented natural speech can correctly predict some of the infant phone discrimination data (Feldman et al., 2021; Schatz et al., 2021). Based on the results of this study, the DPGMM

and the CAE-RNN show some promise as models of early phonetic learning. Their lack of correct predictions on the Catalan contrast may have to do with the amount of noise in the training and/or test data, because a supervised phoneme-recognizer trained and tested on the same data also showed high discrimination error. This suggests that our results on Catalan might be seen as an absence of evidence in favor or against any of the models. However, the noise in the data cannot explain why the supervised model could correctly predict the differences in the discrimination of the Catalan contrast for the Catalan- vs. Spanish-learning infants, and ideally our models should be evaluated on different suitable corpora, once they become available. Although we have been able to show that only two out of the five tested models are likely candidates for modeling infant speech perception, the existing infants' phone discrimination data may simply not be sufficient for distinguishing between the two more successful models.

8.2 Mechanisms of early phonetic learning

The second focus of our studies was the distinction between bottom-up vs. top-down learning mechanisms. In particular, the DPGMM, the AE, and the AE-RNN are purely unsupervised models that learn from frame-level data. In contrast, the CAE and the CAE-RNN use weak top-down guidance from the word level (although the word forms could also be detected in an unsupervised way, see Kamper, 2019). The DPGMM and the CAE-RNN were equally successful in predicting the infants' crosslinguistic phone discrimination patterns. At the same time, these two models represent very different algorithms. If we compare models that use the same architectures — AE vs. CAE and AE-RNN vs. CAE-RNN — we can conclude that, everything else being equal, the top-down guidance can help in making predictions that are more similar to infants' data on phone discrimination across languages. At the same time, recall that the pairs of acoustic words instances for the CAE and the CAE-RNN models in our case were obtained using supervised alignment methods. Training these models on a noisier set of word pairs obtained using fully unsupervised word discovery methods, as in Kamper (2019), could potentially diminish the benefits of the top-down guidance.

An orthogonal distinction we have made relates to the representation unit the models are trained on: frames vs. word-sized units. Recall that at test time this corresponds, respectively, to represent-

ing *phones* either as sequences of frames or as holistic units. Again, if we compare the models with similar architectures that differ primarily in their representation type (AE vs. AE-RNN and CAE vs. CAE-RNN), we can conclude that processing longer units holistically in the recurrent models results in better predictions. Therefore, such holistic processing of sequences is advantageous for our autoencoder-based models of infant speech perception. There may be at least two mechanisms providing this advantage. First, at test time, the recurrent models integrate the information from the whole series of frames (corresponding to a phone) into a holistic representation, and comparing such integral representations to each other in an ABX task may be a better model of infants' discrimination behavior than the alternative, i.e., comparing sequences of individual 25-ms-long frames to each other. Second, representation spaces in the recurrent models are built based on sequences longer than a frame or even a phone, and this is compatible with some early theories of phonetic learning (Jusczyk, 1992, 1993), which argue that infants store representations for speech sequences that are longer than a phone.

Because we have only tested four neural network models and did not explore whether their behavior changes depending on the exact architecture and hyperparameters, our results regarding the two distinctions above only present inconclusive evidence and may not be generalizable. Relatedly, while we cannot provide a fair comparison of the neural networks to the DPGMM model, we can speculate that a version of the DPGMM model that could better integrate the information over time could be a better model of early phonetic learning.

8.3 Models' predictions for future testing with infants

To test whether the DPGMM and the CAE-RNN make similar predictions on all phone contrasts, we used the general method proposed by Schatz et al. (2021) to derive English contrasts for which each of the two models predicts robust differences in discrimination by the Japanese vs. English learner. Such contrasts can inform future experimental studies with infants: it is costly to run experiments with infants, and our models' predictions can inform future testing with Japanese- and English-learning infants. While there are some contrasts, such as [l]–[ɹ] and [m]–[ɹ], for which both our models — the DPGMM and the CAE-RNN — predict robust crosslinguistic differences in discrimination rates, our analysis shows that the models do not always make identical predictions

about discriminability, and some contrasts are predicted to yield such differences only by one of the models. Because of this disagreement between the models, the outcomes of future studies with infants can help one decide whether the DPGMM or the CAE-RNN is a better model of early phonetic learning. In this respect, specific recommendations include testing Japanese- and English-learning infants on two groups of contrasts: (1) contrasts that include the rhotacized vowel [ɝ] (predicted to yield a crosslinguistic difference by the DPGMM model), and (2) vowel contrasts that include [ʌ] and some fricative consonant contrasts with [f] (predicted by the CAE-RNN model). If Japanese-learning infants find it more difficult (compared to English-learning infants) to discriminate contrasts from group (1) but not (2), this would speak in favor of the DPGMM model; an inverse pattern would lend support to the CAE-RNN model. If Japanese-learning infants find both groups of contrasts challenging, this would speak in favor of both our models, while a lack of difference between Japanese- and English-learning infants on either group of contrasts would suggest that either our models are not sufficiently detailed models of early phonetic learning, or that the mechanism of deriving robust predictions should be improved. To summarize, consistent discrimination results in infant experiments would speak in favor of one or the other model, or against both, thus making it possible to falsify one or both models on the ground of their ability to generate the effects of interest rather than correctly predict those already observed (Palminteri, Wyart, & Koechlin, 2017).

8.4 Phonetic category information in models' representations

To investigate to what extent the information about adult-like phonetic categories is readily available in the representations of our two 'best' (in terms of predicting infants' patterns) models, and whether the two models are similar in the amount of such information they encode, we carried out an unsupervised clustering and a supervised classification analyses on the phone representations in the DPGMM and the CAE-RNN models. The results across the two types of analyses were consistent in that the DPGMM representations are organized in such a way that the information necessary to discriminate between adult-like phonetic categories can be derived more easily, compared to the CAE-RNN representations. In their study, Schatz et al. (2021) analyzed the representations of the DPGMM model and showed that those were not similar to adult-like phonetic categories in any

meaningful sense. Our result shows that the representations of the CAE-RNN are even less similar to phonetic categories, compared to the DPGMM representations. At the same time, note that a model encoding more information about adult-like phonetic categories is not necessarily a better model of infant phonetic learning, and vice versa.

Irrespective of the exact differences between the DPGMM and the CAE-RNN, it is not a trivial finding that the models encode non-negligible amount of information about phonetic categories in their representations. This finding is interesting for the existing accounts of perceptual space learning, because these accounts argue that categories are learned later in life by carving up the acquired perceptual space (Feldman et al., 2021). The models of perceptual space learning only simulate the early part of this complex learning process, before infants acquire adult-like phonetic categories. Indeed, our classification analyses show that the information about categories may not be readily available in the models' representation spaces. At the same time, our clustering analyses show that even after training computational models of perceptual space learning on small amounts of data (compared to what an infant hears by her first birthday), their perceptual spaces can already be carved up into clusters that vaguely resemble phonetic categories, potentially mimicking infants' very first steps towards adult-like categories. Finally, note that our clustering and classification results should be interpreted with caution, as they only present indirect evidence for or against the emergence of actual phonetic categories in the models' representations. Even if phonetic categories could be separated nearly perfectly in a model's representations, it does not mean that the model uses top-down categorical information during the learning or at test time (see Feldman et al., 2021, for a relevant discussion).

8.5 Future directions

In this work, we have only tested our models of phonetic learning on a particular kind of infants' phone discrimination data, where two groups of infants — native and non-native learners of a certain language — are tested on a given phone contrast from that language. In other kinds of experiments, infants from a single group were tested only on a native or a non-native phonetic contrast (e.g., see an overview for vowel contrasts in Tsuji & Cristia, 2014), and in the future the field can benefit from carrying out a meta-analytic evaluation of how well our and similar models

predict this kind of data, as in the framework proposed by Cruz Blandón et al. (2021).

Moreover, our work has only focused on particular implementations of five computational models. The field of speech engineering has recently seen a large increase in the number of available models for unsupervised and weakly supervised learning (see reviews by Aldarmaki, Ullah, Ram, & Zaki, 2022; Mohamed et al., 2022). Some of these models have been evaluated against adults' categorical phone representations (e.g., Cruz Blandón & Räsänen, 2020) or against behavioral data (e.g., Millet & Dunbar, 2022), and future work could evaluate them on infants' phone discrimination data of the kind used in our work.

Another research direction worth exploring is using more ecological data for training the models. While our study is a step forward compared to work in which models are trained on idealized laboratory stimuli, there is still a gap between the type of data we used and naturalistic input that infants are exposed to. In particular, the performance of a computational model in the machine ABX discrimination task varies depending on the type of input to the model (read speech vs. child-centered recordings, see Lavechin et al., 2023) and the exact composition of the input (in terms of the number of speakers and their gender, see Li, Schatz, Matusevych, Goldwater, & Feldman, 2020). We have used a combination of read and spontaneous adult speech corpora thanks to their availability, and Schatz et al. (2021) showed that qualitative patterns were consistent for both speech registers on their data set. At the same time, once data sets of child-directed recordings for the target languages become available, our models can be trained and tested on such data for their more ecologically valid evaluation. Moreover, training the models on more naturalistic data would also be the first step towards measuring the fit between models' error rate and infants' behavior, rather than evaluating it qualitatively, as in this work.

9 Conclusion

It has recently been proposed that infants' phone discrimination data can be explained in terms of perceptual space learning, without use of phonetic categories (Feldman et al., 2021; Schatz et al., 2021). Here, we have evaluated five computational models of early phonetic learning from naturalistic speech data on three phonetic contrasts, for which infants' phone discrimination data is available. We have found that the generative probabilistic model of Schatz et al. (2021),

DPGMM, and a neural network with weak top-down supervision from the word level, CAE-RNN, can correctly predict qualitative patterns of phone discrimination exhibited by infants on two out of three phonetic contrasts. While Schatz et al. (2021) found that their model's representations did not resemble phonetic categories, we have observed that the representations of the other model, CAE-RNN, resemble phonetic categories to an even lesser extent. Thus, our findings extend the previous proof-in-principle that perceptual space learning is a viable account of early phonetic learning and contribute to the growing body of work which argues against the unconditional assumption of phonetic category learning in infancy (Feldman et al., 2021; McMurray, 2022). Finally, three other models that we tested appeared to be less successful at predicting infants' discrimination data. This result suggests that the existing data from infants can help us distinguish between some (but not all) formal algorithms of phonetic learning and calls for collection of more fine-grained data. Such data would help us test whether the purely bottom-up distributional learning account, as in the DPGMM, or the account with weak top-down guidance from the word level, as in the CAE-RNN, makes a better theory of early phonetic learning. The fact that the two models make different predictions on other phonetic contrasts suggests that more data would help in resolving this issue, and we have provided concrete suggestions about which contrasts may be promising for future data collection with infants.

References

Albareda-Castellot, B., Pons, F., & Sebastián-Gallés, N. (2011). The acquisition of phonetic categories in bilingual infants: new data from an anticipatory eye movement paradigm. *Developmental Science*, *14*, 395–401.

Aldarmaki, H., Ullah, A., Ram, S., & Zaki, N. (2022). Unsupervised automatic speech recognition: A review. *Speech Communication*, *139*, 76–91.

Antetomaso, S., Miyazawa, K., Feldman, N., Elsner, M., Hitczenko, K., & Mazuka, R. (2017). Modeling phonetic category learning from natural acoustic data. In M. LaMendola & J. Scott (Eds.), *Proceedings of the 41st Annual Boston University Conference on Language Development* (pp. 32–45). Somerville, MA: Cascadilla Press.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1–48.

Best, C. T. (1994). The emergence of native-language phonological influences in infants: A perceptual assimilation model. In J. C. Goodman & H. C. Nusbaum (Eds.), *The development of speech perception: The transition from speech sounds to spoken words* (pp. 167–224). Cambridge: MA: The MIT Press.

Best, C. T., & McRoberts, G. W. (2003). Infant perception of non-native consonant contrasts that adults assimilate in different ways. *Language and Speech*, *46*, 183–216.

Best, C. T., McRoberts, G. W., & Sithole, N. M. (1988). Examination of perceptual reorganization for nonnative speech contrasts: Zulu click discrimination by english-speaking adults and infants. *Journal of Experimental Psychology: Human Perception and Performance*, *14*, 345–360.

Bion, R. A., Miyazawa, K., Kikuchi, H., & Mazuka, R. (2013). Learning phonemic vowel length from naturalistic recordings of Japanese infant-directed speech. *PLoS ONE*, *8*, e51594.

Bortfeld, H., Morgan, J. L., Golinkoff, R. M., & Rathbun, K. (2005). Mommy and me: Familiar names help launch babies into speech-stream segmentation. *Psychological Science*, *16*, 298–304.

Bosch, L., & Sebastián-Gallés, N. (2003). Simultaneous bilingualism and the perception of a language-specific vowel contrast in the first year of life. *Language and Speech*, *46*, 217–243.

Bu, H., Du, J., Na, X., Wu, B., & Zheng, H. (2017). AISHELL-1: An open-source Mandarin speech corpus and a speech recognition baseline. In C.-w. Joo and Jihwan Kim (Ed.), *Proceedings of the 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment* (pp. 1–5). Piscataway, NJ: The Institute of Electrical and Electronics Engineers Signal Processing Society.

Burns, T. C., Yoshida, K. A., Hill, K., & Werker, J. F. (2007). The development of phonetic representation in bilingual and monolingual infants. *Applied Psycholinguistics*, *28*, 455–474.

Chang, J., & Fisher III, J. W. (2013). Parallel sampling of DP mixture models using sub-cluster splits. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Proceedings of the 26th International Conference on Neural Information Processing*

Systems - Volume 1 (pp. 620–628). Piscataway, NJ: The Institute of Electrical and Electronics Engineers Signal Processing Society.

Chen, H., Leung, C.-C., Xie, L., Ma, B., & Li, H. (2015). Parallel inference of Dirichlet process Gaussian mixture models for unsupervised acoustic modeling: A feasibility study. In S. Möller, H. Ney, B. Möbius, E. Nöth, & S. Steidl (Eds.), *Proceedings of the 16th Annual Conference of the International Speech Communication Association* (pp. 3189–3193). Piscataway, NJ: The Institute of Electrical and Electronics Engineers Signal Processing Society.

Chung, Y.-A., Hsu, W.-N., Tang, H., & Glass, J. R. (2019). An unsupervised autoregressive model for speech representation learning. In G. Kubin & Z. Kačič (Eds.), *Proceedings of the 20th Annual Conference of the International Speech Communication Association* (pp. 146–150). Piscataway, NJ: The Institute of Electrical and Electronics Engineers Signal Processing Society.

Chung, Y.-A., Wu, C.-C., Shen, C.-H., Lee, H.-Y., & Lee, L.-S. (2016). Audio word2vec: Unsupervised learning of audio segment representations using sequence-to-sequence autoencoder. In N. Morgan (Ed.), *Proceedings of the 17th Annual Conference of the International Speech Communication Association* (pp. 765–769). Piscataway, NJ: The Institute of Electrical and Electronics Engineers Signal Processing Society.

Cruz Blandón, M. A., Cristia, A., & Räsänen, O. (2021). *Evaluation of computational models of infant language development against robust empirical data from meta-analyses: What, why, and how?* PsyArXiv. Retrieved from `https://psyarxiv.com/yjz5a/download`

Cruz Blandón, M. A., & Räsänen, O. (2020). *Analysis of predictive coding models for phonemic representation learning in small datasets.* Paper presented at Self-supervision in Audio and Speech (ICML 2020). Retrieved from `https://openreview.net/forum?id=cnLz5ckGs1y`

Evans, S., & Davis, M. H. (2015). Hierarchical organization of auditory and motor representations in speech perception: evidence from searchlight similarity analysis. *Cerebral Cortex*, *25*, 4772–4788.

Feldman, N. H., Goldwater, S., Dupoux, E., & Schatz, T. (2021). Do infants really learn phonetic categories? *Open Mind*, *5*, 113–131.

Feldman, N. H., Griffiths, T. L., Goldwater, S., & Morgan, J. L. (2013). A role for the developing lexicon in phonetic category acquisition. *Psychological Review*, *120*, 751–778.

Frank, S., Feldman, N., & Goldwater, S. (2014). Weak semantic context helps phonetic learning in a model of infant language acquisition. In K. Toutanova & H. Wu (Eds.), *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1073–1083). Stroudsburg, PA: Association for Computational Linguistics.

Garrido, J. M., Escudero, D., Aguilar, L., Cardeñoso, V., Rodero, E., de-la Mota, C., . . . Bonafonte, A. (2013). Glissando: a corpus for multidisciplinary prosodic studies in Spanish and Catalan. *Language Resources and Evaluation*, *47*, 945–971.

Jansen, A., Dupoux, E., Goldwater, S., Johnson, M., Khudanpur, S., Church, K., . . . Thomas, S. (2013). A summary of the 2012 JHU CLSP workshop on zero resource speech technologies and models of early language acquisition. In M. Adams & V. Zhao (Eds.), *Proceedings of 2013 IEEE International Conference on Acoustics, Speech, and Signal Processing* (pp. 8111–8115). Piscataway, NJ: The Institute of Electrical and Electronics Engineers Signal Processing Society.

Jones, S. D., & Brandt, S. (2020). Density and distinctiveness in early word learning: Evidence from neural network simulations. *Cognitive Science*, *44*, e12812.

Jusczyk, P. W. (1992). Developing phonological categories from the speech signal. In C. A. Ferguson, L. Menn, & C. Stoel-Gammon (Eds.), *Phonological development: Models, research, implications* (pp. 17–64). Timonium, MD: York Press.

Jusczyk, P. W. (1993). From general to language-specific capacities: The WRAPSA model of how speech perception develops. *Journal of Phonetics*, *21*, 3–28.

Jusczyk, P. W., & Aslin, R. N. (1995). Infants' detection of the sound patterns of words in fluent speech. *Cognitive Psychology*, *29*, 1–23.

Jusczyk, P. W., Houston, D. M., & Newsome, M. (1999). The beginnings of word segmentation in English-learning infants. *Cognitive Psychology*, *39*, 159–207.

Kamper, H. (2019). Truly unsupervised acoustic word embeddings using weak top-down constraints in encoder-decoder models. In L. Mihaylova, W. Wang, & S. Elliot (Eds.), *Proceedings of the 2019 IEEE International Conference on Acoustics, Speech, and Signal Processing* (pp. 6535–6539). Piscataway, NJ: The Institute of Electrical and Electronics Engineers Signal

Processing Society.

Kamper, H., Elsner, M., Jansen, A., & Goldwater, S. (2015). Unsupervised neural network based feature extraction using weak top-down constraints. In L. Hanlen (Ed.), *Proceedings of 2015 IEEE International Conference on Acoustics, Speech, and Signal Processing* (pp. 5818–5822). Piscataway, NJ: The Institute of Electrical and Electronics Engineers Signal Processing Society.

Kingma, D. P., & Ba, J. (2015). *Adam: A method for stochastic optimization.* Paper presented at the 3rd International Conference on Learning Representations. Retrieved from `https://arxiv.org/pdf/1412.6980.pdf`

Kramer, M. A. (1991). Nonlinear principal component analysis using autoassociative neural networks. *AIChE Journal*, *37*, 233–243.

Kuhl, P. K. (1979). Speech perception in early infancy: Perceptual constancy for spectrally dissimilar vowel categories. *The Journal of the Acoustical Society of America*, *66*, 1668–1679.

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

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*, F13–F21.

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*, 9096–9101.

Kuhn, H. W. (1955). The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, *2*, 83–97.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. (2017). lmerTest package: tests in linear mixed effects models. *Journal of Statistical Software*, *82*, 1–26.

Lavechin, M., de Seyssel, M., Titeux, H., Bredin, H., Wisniewski, G., Cristia, A., & Dupoux, E. (2023). *Statistical learning bootstraps early language acquisition.* PsyArXiv. Retrieved from `https://psyarxiv.com/hav58/download`

Levy, D. F., & Wilson, S. M. (2020). Categorical encoding of vowels in primary auditory cortex. *Cerebral Cortex*, *30*, 618–627.

Li, R., Schatz, T., Matusevych, Y., Goldwater, S., & Feldman, N. H. (2020). Input matters in the modeling of early phonetic learning. In S. Denison, M. Mack, Y. Xu, & B. Armstrong (Eds.), *Proceedings of the 42nd Annual Conference of the Cognitive Science Society* (pp. 578–584). Austin, TX: Cognitive Science Society.

Linzen, T. (2019). What can linguistics and deep learning contribute to each other? Response to Pater. *Language*, *95*, e99–e108.

Maekawa, K. (2003). *Corpus of Spontaneous Japanese: Its design and evaluation.* Paper presented at the ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition. Retrieved from `https://www.isca-speech.org/archive_open/archive_papers/sspr2003/sspr_mmo2.pdf`

Mareschal, D., & French, R. M. (2017). TRACX2: a connectionist autoencoder using graded chunks to model infant visual statistical learning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*, 20160057.

Matusevych, Y., Kamper, H., Schatz, T., Feldman, N., & Goldwater, S. (2021). A phonetic model of non-native spoken word processing. In J. Tiedemann & R. Tsarfaty (Eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume* (pp. 1480–1490). Stroudsburg, PA: Association for Computational Linguistics.

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82*, B101–B111.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal Forced Aligner: Trainable text-speech alignment using Kaldi. In F. Lacerda (Ed.), *Proceedings of the 18th Annual Conference of the International Speech Communication Association* (pp. 498–502). Piscataway, NJ: The Institute of Electrical and Electronics Engineers Signal Processing Society.

McMurray, B. (2022). The acquisition of speech categories: Beyond perceptual narrowing, beyond unsupervised learning and beyond infancy. *Language, Cognition and Neuroscience*, 1–27.

McMurray, B., Aslin, R. N., & Toscano, J. C. (2009). Statistical learning of phonetic categories: insights from a computational approach. *Developmental Science*, *12*, 369–378.

McMurray, B., Danelz, A., Rigler, H., & Seedorff, M. (2018). Speech categorization develops slowly through adolescence. *Developmental Psychology*, *54*, 1472–1491.

Millet, J., & Dunbar, E. (2022). Do self-supervised speech models develop human-like perception biases? In D. Croce, R. Cotterell, & J. Zhang (Eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 7591–7605). Stroudsburg, PA: Association for Computational Linguistics.

Miyazawa, K., Kikuchi, H., & Mazuka, R. (2010). Unsupervised learning of vowels from continuous speech based on self-organized phoneme acquisition model. In T. Kobayashi, K. Hirose, & S. Nakamura (Eds.), *Proceedings of the 11th Annual Conference of the International Speech Communication Association* (pp. 2914–2917). Piscataway, NJ: The Institute of Electrical and Electronics Engineers Signal Processing Society.

Miyazawa, K., Miura, H., Kikuchi, H., & Mazuka, R. (2011). The multi timescale phoneme acquisition model of the self-organizing based on the dynamic features. In P. Cosi & R. De Mori (Eds.), *Proceedings of the 12th Annual Conference of the International Speech Communication Association* (pp. 749–752). Piscataway, NJ: The Institute of Electrical and Electronics Engineers Signal Processing Society.

Mohamed, A., Lee, H.-y., Borgholt, L., Havtorn, J. D., Edin, J., Igel, C., . . . Watanabe, S. (2022). Self-supervised speech representation learning: A review. *IEEE Journal of Selected Topics in Signal Processing*, *16*, 1179–1210.

Nixon, J. S., & Tomaschek, F. (2021). Prediction and error in early infant speech learning: A speech acquisition model. *Cognition*, *212*, 104697.

Palminteri, S., Wyart, V., & Koechlin, E. (2017). The importance of falsification in computational cognitive modeling. *Trends in Cognitive Sciences*, *21*(6), 425–433.

Paul, D. B., & Baker, J. M. (1992). The design for the Wall Street Journal-based CSR corpus. In M. P. Marcus (Ed.), *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992* (pp. 357–362). San Mateo, CA: Morgan Kaufmann.

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*, 89–95.

Plaut, D. C., & Vande Velde, A. K. (2017). Statistical learning of parts and wholes: A neural network approach. *Journal of Experimental Psychology: General*, *146*, 318–336.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., ... Veselý, K. (2011). *The Kaldi speech recognition toolkit.* Retrieved from `https://infoscience.epfl.ch/record/192584/files/Povey_ASRU2011_2011.pdf`

Reh, R. K., Hensch, T. K., & Werker, J. F. (2021). Distributional learning of speech sound categories is gated by sensitive periods. *Cognition*, *213*, 104653.

Renshaw, D., Kamper, H., Jansen, A., & Goldwater, S. (2015). A comparison of neural network methods for unsupervised representation learning on the zero resource speech challenge. In B. Yegnanarayana (Ed.), *Proceedings of the 16th Annual Conference of the International Speech Communication Association* (pp. 3199–3203). Piscataway, NJ: The Institute of Electrical and Electronics Engineers Signal Processing Society.

Robinaugh, D. J., Haslbeck, J. M., Ryan, O., Fried, E. I., & Waldorp, L. J. (2021). Invisible hands and fine calipers: A call to use formal theory as a toolkit for theory construction. *Perspectives on Psychological Science*, *16*, 725–743.

Rosenberg, A., & Hirschberg, J. (2007). V-measure: A conditional entropy-based external cluster evaluation measure. In J. Eisner (Ed.), *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (pp. 410–420). Stroudsburg, PA: Association for Computational Linguistics.

Schatz, T., Bach, F., & Dupoux, E. (2018). Evaluating automatic speech recognition systems as quantitative models of cross-lingual phonetic category perception. *The Journal of the Acoustical Society of America*, *143*, EL372–EL378.

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*, e2001844118.

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. In F. Bimbot, C. Fougeron, & F. Pellegrino (Eds.), *Proceedings of the 14th Annual Conference of the International Speech Communication Association* (pp. 1781–1785). Piscataway, NJ: The Institute of Electrical and Electronics Engineers Signal Processing Society.

Schultz, T. (2002). GlobalPhone: a multilingual speech and text database developed at Karlsruhe University. In J. Hansen (Ed.), *Proceedings of the 7th International Conference on Spoken Language Processing* (pp. 345–348). Piscataway, NJ: The Institute of Electrical and Electronics Engineers Signal Processing Society.

Segal, O., Hejli-Assi, S., & Kishon-Rabin, L. (2016). The effect of listening experience on the discrimination of /ba/ and /pa/ in Hebrew-learning and Arabic-learning infants. *Infant Behavior and Development*, *42*, 86–99.

Shain, C., & Elsner, M. (2019). Measuring the perceptual availability of phonological features during language acquisition using unsupervised binary stochastic autoencoders. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 69–85). Stroudsburg, PA: Association for Computational Linguistics.

Swingley, D. (2009). Contributions of infant word learning to language development. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*, 3617–3632.

Thiollière, R., Dunbar, E., Synnaeve, G., Versteegh, M., & Dupoux, E. (2015). A hybrid dynamic time warping-deep neural network architecture for unsupervised acoustic modeling. In B. Yegnanarayana (Ed.), *Proceedings of the 16th Annual Conference of the International Speech Communication Association.* Piscataway, NJ: The Institute of Electrical and Electronics Engineers Signal Processing Society.

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*, 2285–2294.

Tsuji, S., & Cristia, A. (2014). Perceptual attunement in vowels: A meta-analysis. *Developmental Psychobiology*, *56*, 179–191.

Tsushima, T., Takizawa, O., Sasaki, M., Shiraki, S., Nishi, K., Kohno, M., . . . 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. In K. Shirai (Ed.), *Proceedings of the 3rd International Conference on Spoken Language Processing* (pp. 1695–1698). Piscataway, NJ: The Institute of Electrical and Electronics Engineers Signal Processing Society.

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*, 13273–13278.

Vintsyuk, T. K. (1968). Speech discrimination by dynamic programming. *Cybernetics*, *4*, 52–57.

Werker, J. F., & Curtin, S. (2005). PRIMIR: A developmental framework of infant speech processing. *Language Learning and Development*, *1*, 197–234.

Werker, J. F., & Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development*, *7*, 49–63.

Zeiler, M. D. (2012). *ADADELTA: An adaptive learning rate method.* ArXiv. Retrieved from `https://arxiv.org/pdf/1212.5701.pdf`

Appendix: Formal models

Here we provide formal definitions and the hyperparameters values of the DPGMM (Study 1), the four neural network models (Study 2), as well as the supervised baseline.

Dirichlet process Gaussian mixture model

The Dirichlet process Gaussian mixture model (DPGMM) is a generative probabilistic model, a GMM with a non-parametric Dirichlet process prior. It learns by maximizing the likelihood of its K components (K changes during learning) given the input speech frame X:

$$\mathcal{L}(\Theta|X) = \sum_{i=1}^{|K|} \pi_i \, p(X|\mu_i, \Sigma_i) \tag{1}$$

where Θ is a set of the model's parameters, and μ_i and Σ_i are the parameters of component i: mean and covariance, respectively. The mixture weights π_i are generated through the stick-breaking process (a particular version of the Dirichlet process), while μ_i and Σ_i of component i are sampled from the normal-inverse Wishart distribution. The inference is done using a parallel Markov chain Monte Carlo sampler (Chang & Fisher III, 2013). More details can be found in Chen et al. (2015); Schatz et al. (2021).

Neural network models

Autoencoder (AE) is a classic auto-associative model which uses reconstruction loss (here: mean squared error) between the feature representation X of an input acoustic frame and its output representation $\bar{X}$:

$$\ell(X) = ||X - \bar{X}||^2 \tag{2}$$

We follow Kamper et al. (2015) and use a stacked version with 8 hidden layers (7×100 and 1×39 units). The model is pretrained for 5 epochs per layer plus 5 epochs of final fine-tuning, without early stopping, with Adadelta optimization with adaptive learning rate (Zeiler, 2012) and decay 0.95. At test time, the second-to-last hidden layer is used to encode individual frames from the test data into the model's representation space.

Correspondence autoencoder (CAE) only differs from the AE in that it is trained on pairs of acoustic frames: a feature representation X of the acoustic frame from one instance of a given word type and a feature representation X' of the aligned (using dynamic time warping) frame from another instance of the same word type. The loss is then computed between X' and the reconstructed version $\bar{X}$ of the representation X:

$$\ell(X, X') = ||X' - \bar{X}||^2 \tag{3}$$

Following Kamper et al. (2015), we initialize the CAE using the AE (with parameters as described above), and then train the CAE with the same architecture for 120 epochs.

Autoencoding recurrent neural network (AE-RNN) includes an encoder RNN and a decoder RNN. The encoder reads an input sequence and updates its hidden state. The final state of the encoder is then transformed into an *acoustic embedding* and passed to the decoder, which uses it to generate an output sequence. Each input sequence consists of a sequence of MFCC feature vectors, $X = (\vec{x}_1, \ldots, \vec{x}_T)$, where T is the sequence length. The loss for a single training item is:

$$\ell(X) = \sum_{t=1}^{T} ||\vec{x}_t - \vec{f}_t(X)||^2 \tag{4}$$

where X is the input sequence, and $f_t(X)$ is the t^{th} decoder output conditioned on the embedding $\mathbf{z}$, which is obtained by transforming the encoder's final state (i.e., reducing the number of dimensions). At test time, a phone sequence is encoded into the model's fixed-dimensional acoustic embedding space. We use the parameters of Kamper (2019): 3 hidden layers (400 gated recurrent units each) in both the decoder and the encoder, embedding dimensionality of 130, 15 epochs training without early stopping using Adam optimization (Kingma & Ba, 2015) with a learning rate of 0.001.

Correspondence-autoencoding recurrent neural network (CAE-RNN) uses an identical architecture, but each input consists of word pairs X and X', which are different acoustic realizations of the same word type. The loss for a single training pair in this case is:

$$\ell(X, X') = \sum_{t=1}^{T'} ||\vec{x}'_t - \vec{f}_t(X)||^2 \tag{5}$$

where X is the input and X' the target output sequence, and $f_t(X)$ is the t^{th} decoder output conditioned on the embedding $\mathbf{z}$. Following Kamper (2019), we use the AE-RNN to initialize the parameters of the CAE-RNN and then train it (with parameters analogous to those of the AE-RNN) for 3 epochs.

Supervised baseline

Our supervised baseline is a phoneme recognizer, with the same architecture and settings described in Schatz et al. (2021). We use a standard training recipe commonly used in speech recognition. It is adapted from the Wall Street Journal corpus recipe available in Kaldi. Specifically, each phoneme recognizer is a combination of an acoustic hidden Markov model Gaussian mixture model (HMM-GMM) and a phoneme-level bigram language model trained using the Kaldi toolkit (Povey et al., 2011). The acoustic model is a probabilistic generative model, where each phoneme is represented as a set of variants conditioned on their position within a word as well as the neighboring phonetic context (i.e., preceding and following phonemes). Furthermore, each variant is modeled as a standard tri-state left-to-right HMM. Each acoustic model is trained using speaker-adaptive training (SAT) through feature maximum likelihood linear regression (fMLLR). The test data is then processed using the acoustic model and the language model together, with the acoustic scale parameter for decoding set to 0.1 (i.e., the probabilities from the language model are weighted higher than the probabilities from the acoustic model).

Supplementary materials

S1 Lists of phones

Table 5 provides lists of phones which we used for each corpus. These lists were largely based on the existing transcriptions of the corpora.

S2 Results controlled for phonetic context

Here, we present extended results for Study 1 and Study 2. While the respective studies only report the models' ABX discrimination error rates for the target phone contrasts in all neighboring phonetic contexts, here we present results for neighboring phonetic contexts that better resemble the experimental setup in the original experiments with infants. Figure S1 shows the error rates for the DPGMM model from Study 1, and Figure S2 shows the error rates for the four neural network models from Study 2. Note that the number of data pairs (i.e., different speaker–phonetic context combinations) in some conditions is very low (see the numbers above each pair of bars in the figures: e.g., for the target Catalan contrast in the 'Right only' condition we only have 2 data pairs), so that the results are likely not to be robust.

Table 5: Lists of phones used in each corpus sample.

Language	Corpus	List of phones
EN	WSJ, Buckeye	ɑː, æ, ʌ, ɔː, aʊ, aɪ, b, ʧ, d, ð, ɛ, ɝ, eɪ, f, ɡ, h, ɪ, iː, ʤ, k, l, m, n, ŋ, oʊ, ɔɪ, p, ɹ, s, ʃ, t, θ, ʊ, uː, v, w, j, z, ʒ
JA	GP	ɸ, ɴ, tɕː, kː, pː, sː, ɕː, tː, ä, äː, b, ʦ, tɕ, d, e, eː, ɡ, h, i, iː, k, m, n, o, oː, p, r, s, ɕ, t, ɯ, ɯː, w, j, z, ʑ
JA	CSJ	ɸ, ɴ, ʦː, tɕː, kː, pː, sː, ɕː, tː, ä, äː, b, ʦ, tɕ, d, e, eː, ɡ, h, i, iː, k, m, n, o, oː, p, r, s, ɕ, t, ɯ, ɯː, w, j, z, ʑ
MN	AIShell	ä˥, ä˧, ä↗, ä˩, ä˦, aɪ˥, aɪ˧, aɪ↗, aɪ˩, aɪ˦, ɑʊ˥, ɑʊ˧, ɑʊ↗, ɑʊ˩, ɑʊ˦, p, tsʰ, tʂʰ, t, ə˥, ə˧, ə↗, ə˩, ə˦, eɪ˥, eɪ˧, eɪ↗, eɪ˩, eɪ˦, f, k, x, i˥, i˧, i↗, i˩, i˦, iä˥, iä˧, iä↗, iä˩, iä˦, iɑʊ˥, iɑʊ˧, iɑʊ↗, iɑʊ˩, iɑʊ˦, iɛ˥, iɛ˧, iɛ↗, iɛ˩, iɛ˦, iʊ˥, iʊ˧, iʊ↗, iʊ˩, iʊ˦, ioʊ˥, ioʊ˧, ioʊ↗, ioʊ˩, ioʊ˦, iu˥, iu˧, iu↗, iu˩, iu˦, ʨ, kʰ, l, m, n, ŋ, ɤ˥, ɤ˧, ɤ↗, ɤ˩, ɤ˦, oʊ˥, oʊ˧, oʊ↗, oʊ˩, oʊ˦, pʰ, ʨʰ, ɻ̩, s, ʂ, tʰ, u˥, u˧, u↗, u˩, u˦, uä˥, uä˧, uä↗, uä˩, uä˦, uaɪ˥, uaɪ˧, uaɪ↗, uaɪ˩, uaɪ˦, uə˥, uə˧, uə↗, uə˩, uə˦, ueɪ˥, ueɪ˧, ueɪ↗, ueɪ˩, ueɪ˦, uɪ˥, uɪ˧, uɪ↗, uɪ˩, uɪ˦, uɔ˥, uɔ˧, uɔ↗, uɔ˩, uɔ˦, y˥, y˧, y↗, y˩, y˦, yɛ˥, yɛ˧, yɛ↗, yɛ˩, yɛ˦, yœ˥, yœ˧, yœ↗, yœ˩, yœ˦, ɕ, j, ʦ, ʈʂ
MN	GP	ä˥, ä˧, ä↗, ä˩, ä˦, aɪ˥, aɪ˧, aɪ↗, aɪ˩, aɪ˦, ɑʊ˥, ɑʊ˧, ɑʊ↗, ɑʊ˩, ɑʊ˦, p, tsʰ, tʂʰ, t, ə˥, ə˧, ə↗, ə˩, ə˦, eɪ˥, eɪ˧, eɪ↗, eɪ˩, eɪ˦, f, k, x, i˥, i˧, i↗, i˩, i˦, iä˥, iä˧, iä↗, iä˩, iä˦, iɑʊ˥, iɑʊ˧, iɑʊ↗, iɑʊ˩, iɑʊ˦, iɛ˥, iɛ˧, iɛ↗, iɛ˩, iɛ˦, ɨ˥, ɨ˧, ɨ↗, ɨ˩, ɨ˦, iʊ˥, iʊ˧, iʊ↗, iʊ˩, iʊ˦, ioʊ˥, ioʊ˧, ioʊ↗, ioʊ˩, ioʊ˦, iu˥, iu˧, iu↗, iu˩, iu˦, ʨ, kʰ, l, m, n, ŋ, ɤ˥, ɤ˧, ɤ↗, ɤ˩, ɤ˦, oʊ˥, oʊ˧, oʊ↗, oʊ˩, oʊ˦, pʰ, ʨʰ, ɻ̩, s, ʂ, tʰ, u˥, u˧, u↗, u˩, u˦, uä˥, uä˧, uä↗, uä˩, uä˦, uaɪ˥, uaɪ˧, uaɪ↗, uaɪ˩, uaɪ˦, uə˥, uə˧, uə↗, uə˩, uə˦, ueɪ˥, ueɪ˧, ueɪ↗, ueɪ˩, ueɪ˦, uɔ˥, uɔ˧, uɔ↗, uɔ˩, uɔ˦, y˥, y˧, y↗, y˩, y˦, yɛ˥, yɛ˧, yɛ↗, yɛ˩, yɛ˦, yœ˥, yœ˧, yœ↗, yœ˩, yœ˦, ɕ, ʦ, ʈʂ
CA	Glissando	ə, β, ɛ, ɲ, ʎ, ŋ, ɔ, ʃ, θ, ʒ, a, b, d, ʤ, ʣ, e, eɪ, f, g, h, ɦ, i, j, k, l, m, n, o, p, ɾ, r, s, t, ʧ, u, v, w, x, z
ES	Glissando	β, ɛ, ɲ, ʎ, ŋ, ʃ, θ, a, b, d, ʤ, e, eɪ, f, g, h, i, j, ɟ, k, l, m, n, o, p, ɾ, r, s, t, ʧ, u, w, x

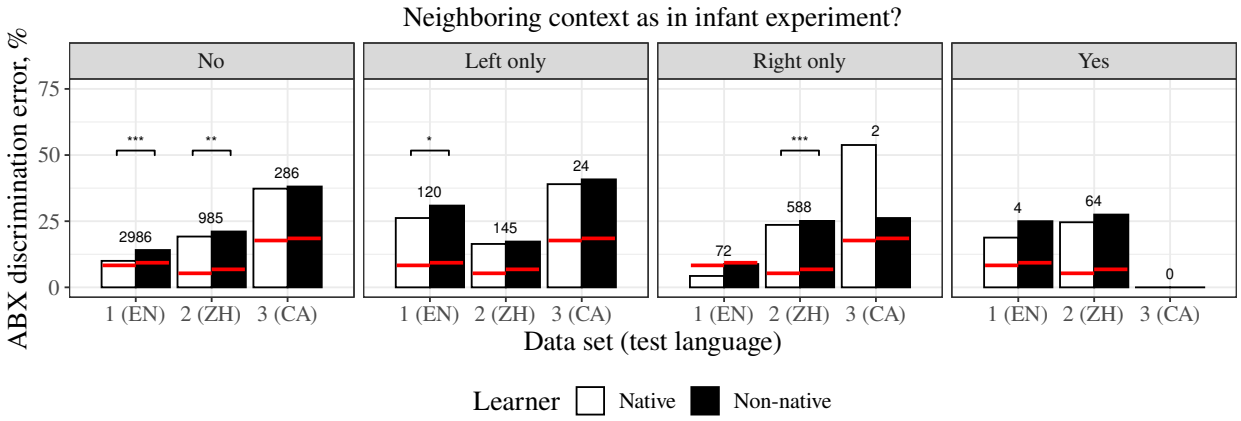


Figure S1: Extended results for Study 1. ABX error rates of the native and non-native DPGMM models in the three discrimination tasks (EN [ɹ]–[l], ZH [ɕ]–[tɕʰ] and CA [e]–[ɛ]), with different degree of control over the neighboring phonetic context of the target phones in the test data (any context, left/right/both contexts as in the original experiments with infants). The number of data pairs (i.e., different speaker–phonetic context combinations) in each test set is shown on top of each bar. Red lines indicate model's error rates averaged over all consonant (for EN and ZH) or all vowel (for CA) contrasts. To match the infant pattern of discrimination, the native model in each pair must show significantly lower error rates than the non-native model. The number of asterisks denotes significance level: *** corresponds to $p < .001$, ** to $p < .01$, and * to $p < .05$.

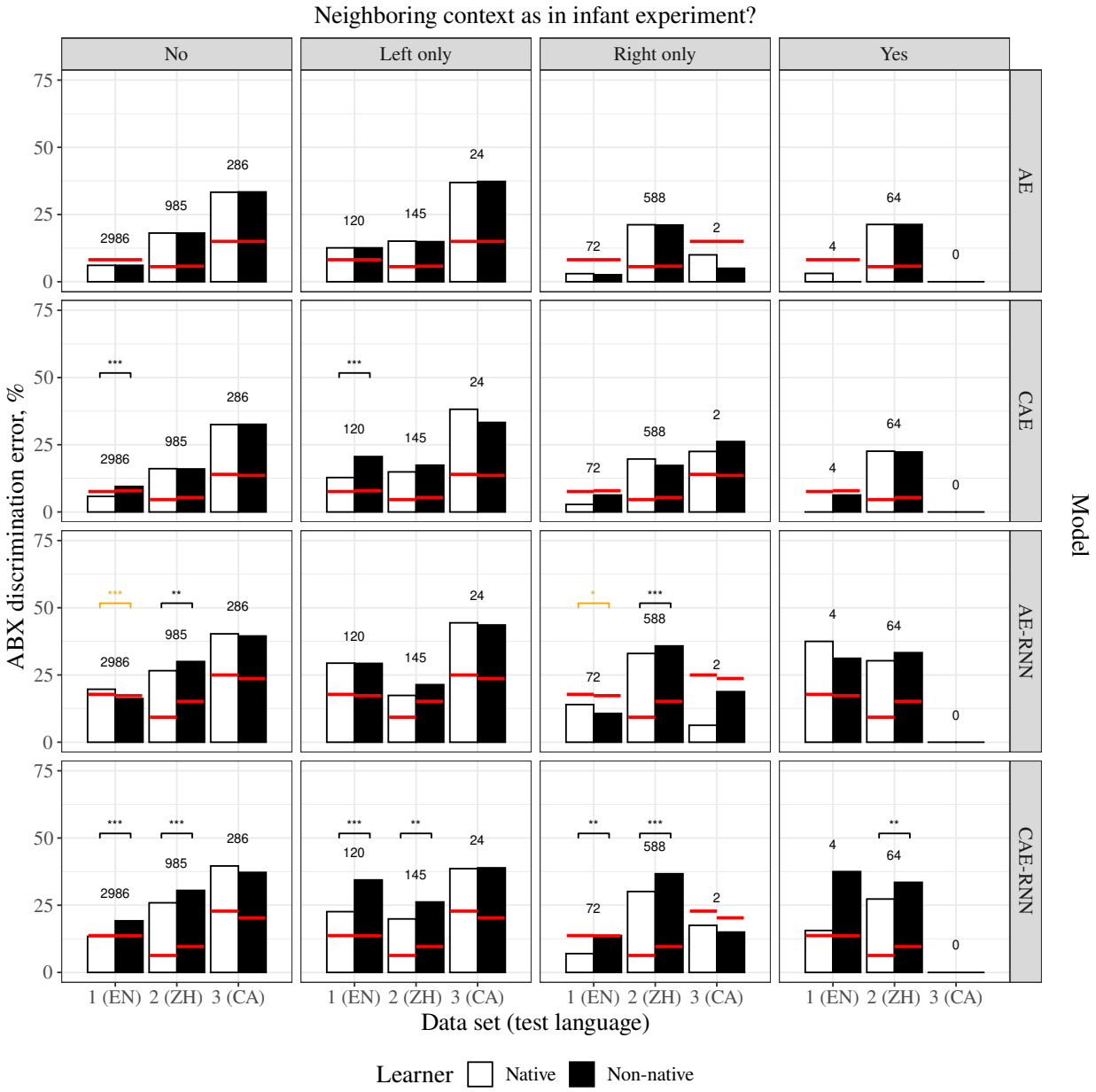


Figure S2: Extended results for Study 2. ABX error rates of the native and non-native neural network models in the three discrimination tasks (EN [ɹ]–[l], ZH [ɕ]–[tɕʰ] and CA [e]–[ɛ]), with different degree of control over the neighboring phonetic context of the target phones in the test data (any context, left/right/both contexts as in the original experiments with infants). The number of data pairs (i.e., different speaker–phonetic context combinations) in each test set is shown on top of each bar. Red lines indicate models' error rates averaged over all consonant (for EN and ZH) or all vowel (for CA) contrasts. To match the infant pattern of discrimination, the native model in each pair must show significantly lower error rates than the non-native model (black brackets), the inverse pattern is wrong even if the difference is significant (orange brackets). The number of asterisks denotes significance level: *** corresponds to $p < .001$, ** to $p < .01$, and * to $p < .05$.

S3 Representation analyses for Catalan and Spanish models

In this section, we report the results of the models' representation analyses (i.e., clustering and classification) for Catalan and Spanish. While the analyses in Section 7 are carried out across corpora (i.e., each model is trained on one corpus and then tested on another corpus of the same language), for both Catalan and Spanish we only had one corpus, and we run the analyses on two samples from that corpus.

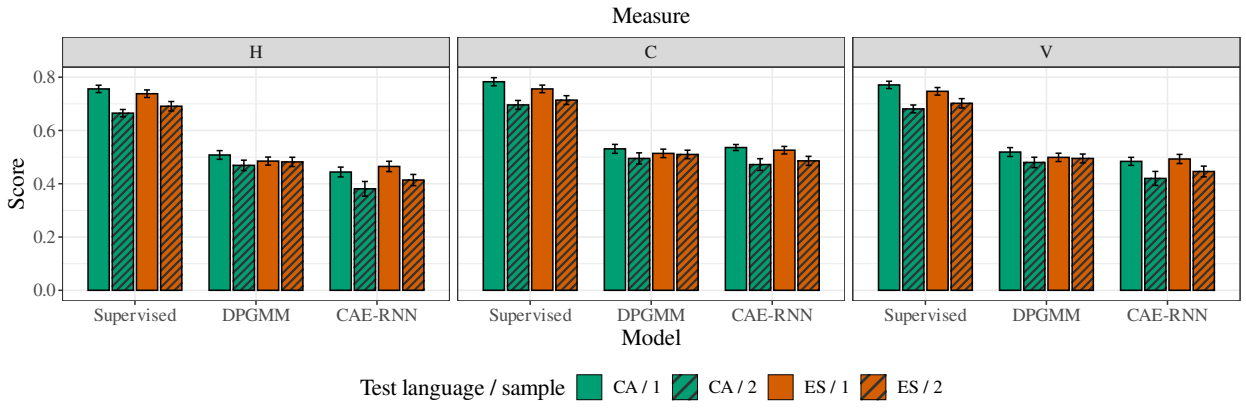


Figure S3: Quality of the unsupervised clusters, found by clustering the representations in the two most successful models (DPGMM and CAE-RNN) and the supervised baseline in Catalan and Spanish. Plots display homogeneity (left), completeness (middle), and V-measure (right) averaged over the 10 data splits, error bars show standard error of the mean.

The quality of the unsupervised clusters for Catalan and Spanish is shown in Figure S3. The general pattern of results is similar to that in other languages (cf. Figure 4): the supervised model shows the highest quality on all three measures ($0.55 \leq H \leq 0.66$, $0.56 \leq C \leq 0.67$, $0.56 \leq V \leq 0.67$), and the perceptual space learning models show lower quality ($0.14 \leq H \leq 0.37$, $0.28 \leq C \leq 0.44$, $0.19 \leq V \leq 0.38$), with the DPGMM on average being somewhat better at this clustering task than the CAE-RNN. We observe a similar pattern in Figure S4, which shows how well the models' clusters map onto true phone labels in Catalan and Spanish. Again, these results are similar to what we report for the other languages (cf. Figure 6).

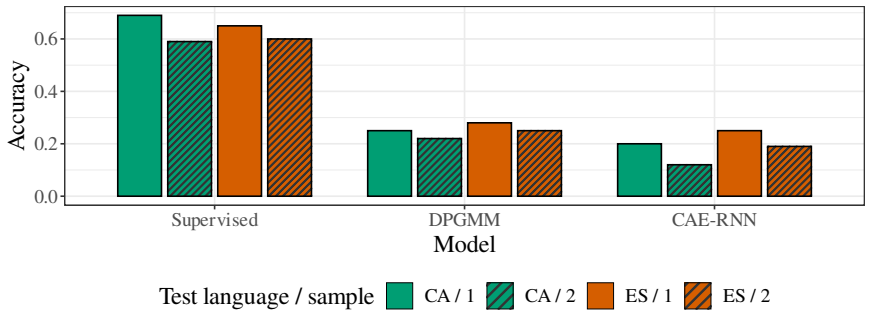


Figure S4: Accuracy of the mapping between true phone labels and the best matching clusters in Catalan and Spanish, computed using the Hungarian Algorithm.

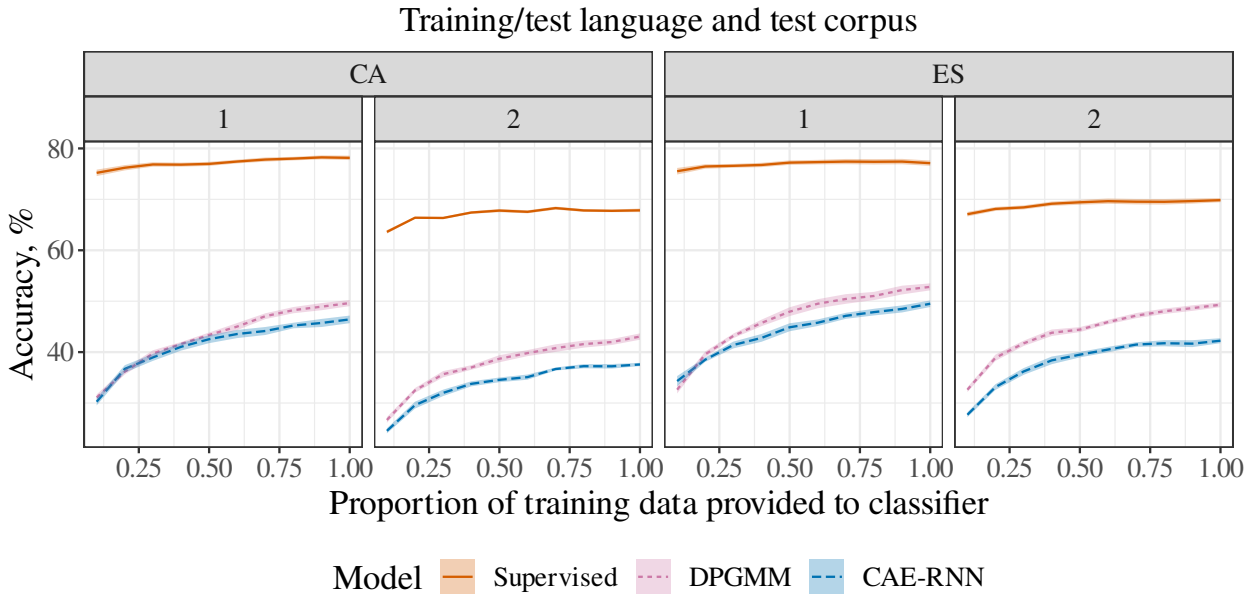


Figure S5: Accuracy of the k-NN phone classifier trained and tested on the representations of phone instances in the DPGMM and the CAE-RNN model for Catalan and Spanish. Bands show the standard error of the mean for 10 random training–test data splits.

Phone classification results for Catalan and Spanish (Figure S5), again, show patterns that we also observe in the other languages (cf. Figure 7). Specifically, the supervised model achieves the highest accuracy among all models, and it does so already after seeing a small share of the data (20–30%). For the DPGMM and the CAE-RNN, the accuracy increases with the amount of the training data, and the DPGMM achieves higher accuracy than the CAE-RNN.

Overall, the main patterns of results for Catalan and Spanish reported in this section, as well as the performance in absolute terms, are consistent with what we observe in English, Japanese,

and Mandarin. This suggests that the structure of the perceptual spaces in our models does not fundamentally change, whether the training and test data come from the same corpus or from different corpora. At the same time, we can speculate that, because phones may be more consistent within than across corpora, the clustering and classification performance of our models for Catalan and Spanish would have been lower, had the training and test data come from different corpora. This hypothetical result would indicate a lower quality of Catalan and Spanish representations compared to the other languages, which would support our suggestion that the training/test data for Catalan is noisy and explain our results on crosslinguistic discrimination, i.e., the models' inability to correctly predict the infants' pattern for the target Catalan contrast.

S4 Full confusion matrices

In Section 7.1 (Figure 5), we presented fragments of confusion matrices between the true phone labels and the clusters extracted from the DPGMM and the CAE-RNN representations. Here, we present full confusion matrices for reference: Figures S6 and S7 show the results for the DPGMM and the CAE-RNN, respectively. We observe that in the DPGMM representations most of the confusions occur for acoustically similar phones (e.g., vowels or fricative consonants), whereas in the CAE-RNN representations some clusters (with indices 36, 0, and 2) cross many phones with very different acoustic characteristics.

Figure S6: A confusion matrix between the true phone labels vs. DPGMM-based best matching clusters on English Buckeye data. The clusters are obtained using unsupervised agglomerative clustering on the DPGMM representations of 100 instances for each English phone, and the matching is done by solving the assignment problems between true categories and predicted clusters using the Hungarian Algorithm. KL divergence is used as a distance measure.

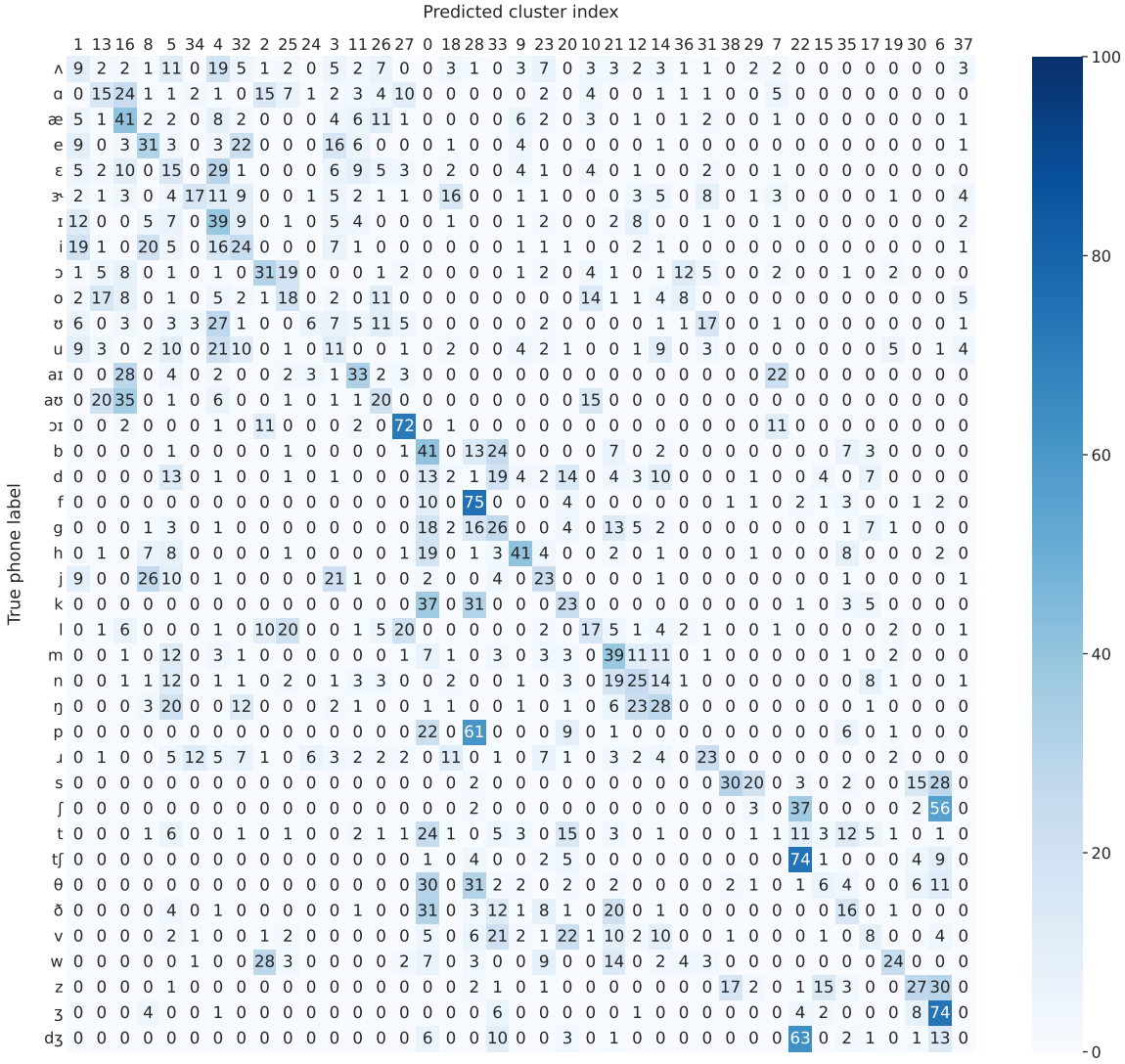
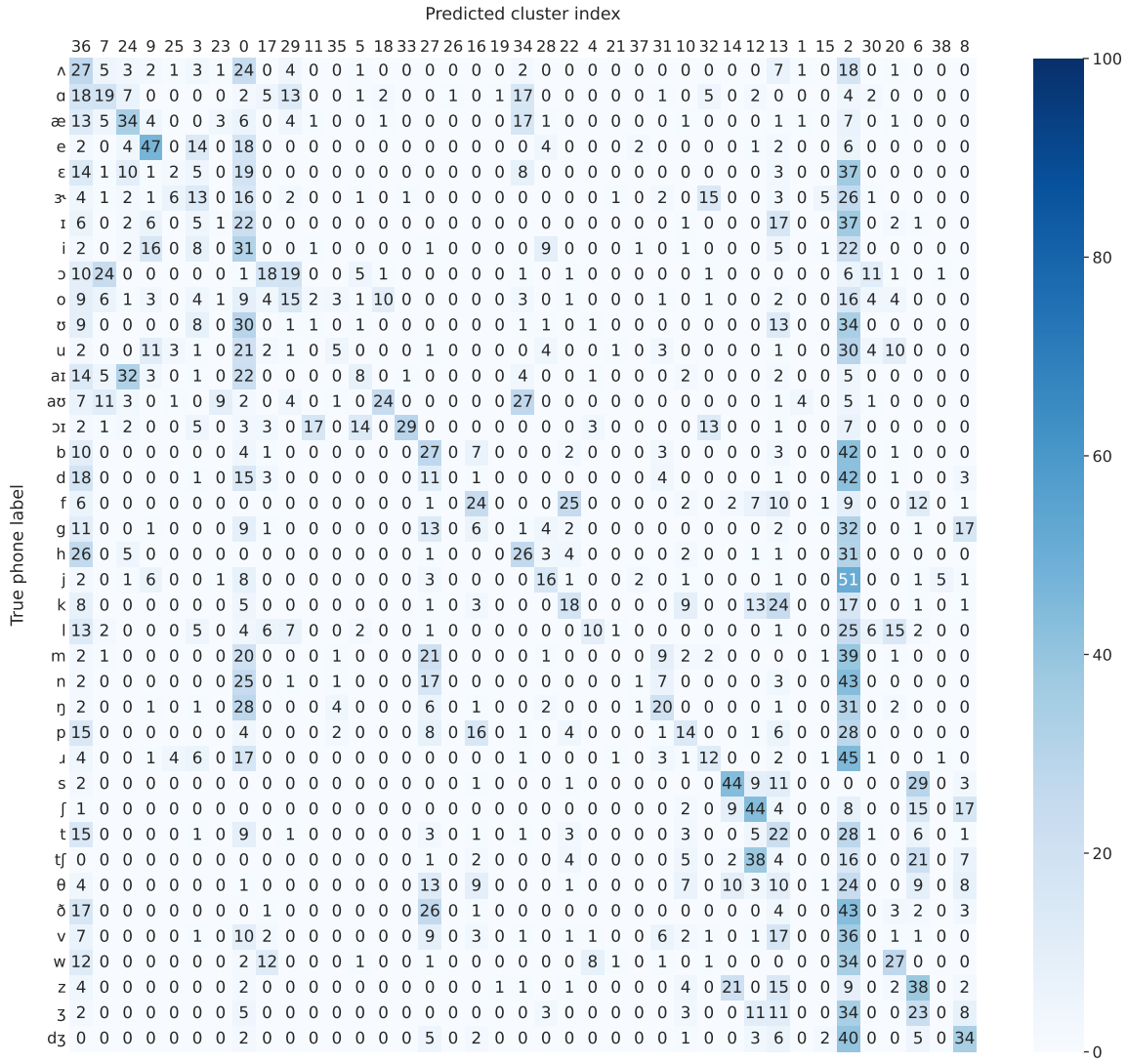


Figure S7: A confusion matrix between the true phone labels vs. CAE-RNN-based best matching clusters on English Buckeye data. The clusters are obtained using unsupervised agglomerative clustering on the CAE-RNN representations of 100 instances for each English phone, and the matching is done by solving the assignment problems between true categories and predicted clusters using the Hungarian Algorithm. Angular distance is used as a distance measure.



S5 CAE-RNN simulations with Euclidean distances

In the previous section, the confusion matrices are constructed using the default distance measures we used in all our simulations (KL divergence for the DPGMM and angular distance for the CAE-RNN). Angular distance may be less suitable for clustering than KL divergence, and ideally we would use KL divergence for the CAE-RNN representations as well. However, this was not possible because KL divergence between two vectors requires them to be proper distributions, which is not the case for the CAE-RNN representations. Therefore, in Figure S8 we provide a confusion matrix for the CAE-RNN representations constructed using Euclidean distance, which is more suitable for clustering than angular distance. Comparing this confusion matrix to the one in Figure S7, we can see that, unlike angular distance, Euclidean distance does not yield clusters that span across multiple phones with very different acoustic characteristics.

To test how the choice of distance measure affected our representation analyses, we computed the cluster quality for the CAE-RNN representations using Euclidean distance, and it was not substantially different. Furthermore, to ensure that our choice of distance measure did not affect our main results on the target phone contrast discrimination in Study 2, we used our CAE-RNN models trained on English and Japanese to run ABX phone discrimination experiments on the target English [ɹ]–[l] contrast using Euclidean distance. We compared the results with the original discrimination results using angular distance, and the absolute discrimination rates were very similar: when Euclidean distance was used, the rates were 13.9% (English model) vs. 19.3% (Japanese model), while the original results using angular distance were 13.4% vs. 19.2% for English and Japanese model, respectively, suggesting that the choice of a distance measure likely did not play a significant role in our experiments.

Figure S8: A confusion matrix between the true phone labels vs. CAE-RNN-based best matching clusters on English Buckeye data. The clusters are obtained using unsupervised agglomerative clustering on the CAE-RNN representations of 100 instances for each English phone, and the matching is done by solving the assignment problems between true categories and predicted clusters using the Hungarian Algorithm. Euclidean distance is used as a distance measure.