

# Early phonetic learning without phonetic categories – Insights from large-scale simulations on realistic input

Thomas Schatz, Naomi H Feldman, Sharon Goldwater, Xuan Nga Cao, Emmanuel Dupoux

### ► To cite this version:

Thomas Schatz, Naomi H Feldman, Sharon Goldwater, Xuan Nga Cao, Emmanuel Dupoux. Early phonetic learning without phonetic categories – Insights from large-scale simulations on realistic input. Proceedings of the National Academy of Sciences of the United States of America, 118 (7), pp.e2001844118, 2021, 10.1073/pnas.2001844118. hal-03070566

## HAL Id: hal-03070566 https://hal.science/hal-03070566

Submitted on 15 Dec 2020  $\,$ 

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers. L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

## Early phonetic learning without phonetic categories Insights from large-scale simulations on realistic input

Thomas Schatz<sup>a,1</sup>, Naomi H. Feldman<sup>a</sup>, Sharon Goldwater<sup>b</sup>, Xuan-Nga Cao<sup>c</sup>, and Emmanuel Dupoux<sup>c,d</sup>

<sup>a</sup> Department of Linguistics & UMIACS, University of Maryland, College Park, MD 20742, USA; <sup>b</sup>School of Informatics, University of Edinburgh, 10 Crichton St, Edinburgh EH8 9AB, UK; <sup>c</sup>Cognitive Machine Learning (ENS - EHESS - PSL Research University - CNRS - INRIA), 2 Rue Simone IFF, 75012 Paris, France; <sup>d</sup>Facebook A.I. Research, 6 Rue Ménars, 75002 Paris, France

This manuscript was compiled on August 7, 2020

Before they even speak, infants become attuned to the sounds of the language(s) they hear, processing native phonetic contrasts more 2 easily than non-native ones (1-3). For example, between 6-8 months and 10-12 months, infants learning American English get better at 4 distinguishing English [J] and [I], as in 'rock' vs 'lock', relative to 5 infants learning Japanese (4). Influential accounts of this early 6 phonetic learning phenomenon initially proposed that infants group 7 sounds into native vowel- and consonant-like phonetic categories-8 like [J] and [I] in English-through a statistical clustering mechanism 9 dubbed 'distributional learning' (5-8). The feasibility of this mech-10 anism for learning phonetic categories has been challenged, how-11 ever (9-16). Here we demonstrate that a distributional learning al-12 gorithm operating on naturalistic speech can predict early phonetic 13 learning as observed in Japanese and American English infants, sug-14 gesting that infants might learn through distributional learning after 15 all. We further show, however, that contrary to the original distri-16 butional learning proposal, our model learns units too brief and too 17 fine-grained acoustically to correspond to phonetic categories. This 18 challenges the influential idea that what infants learn are phonetic 19 categories. More broadly, our work introduces a novel mechanism-20 driven approach to the study of early phonetic learning, together with 21 a quantitative modeling framework that can handle realistic input. 22 This allows, for the first time, accounts of early phonetic learning 23 to be linked to concrete, systematic predictions regarding infants' 24 attunement. 25

Phonetic learning | Language acquisition | Computational modeling

dults have difficulties perceiving consonants and vowels A of foreign languages accurately (17). For example, native 2 Japanese listeners often confuse American English [1] and [1] 3 (as in 'rock' vs 'lock') (18, 19) and native American English 4 listeners often confuse French [u] and [y] (as in 'roue', wheel, 5 versus 'rue', street) (20). This phenomenon is pervasive (21)6 and persistent: even extensive, dedicated training can fail to 7 eradicate these difficulties (22-24). The main proposed expla-8 nations for this effect revolve around the idea that adult speech perception involves a 'native filter': an automatic, involuntary 10 and not very plastic mapping of each incoming sound, foreign 11 or not, onto native phonetic categories, i.e. the vowels and con-12 sonants of the native language (25–29). American English [1] 13 and [1], for example, would be confused by Japanese listeners 14 because their productions can be seen as possible realizations 15 of the same Japanese consonant, giving rise to similar percepts 16 after passing through the 'native Japanese filter'. 17

Surprisingly, these patterns of perceptual confusion arise
very early during language acquisition. Infants learning American English distinguish [I] and [l] more easily than infants

learning Japanese before they even utter their first word (4). 21 Dozens of other instances of such early phonetic learning have 22 been documented, whereby cross-linguistic confusion patterns 23 matching those of adults emerge during the first year of life 24 (2, 3, 30). These observations naturally led to the assump-25 tion that the same mechanism thought to be responsible for 26 adults' perception might be at work in infants, i.e. foreign 27 sounds are being mapped onto native phonetic categories. This 28 assumption—which we will refer to as the *phonetic category* 29 hypothesis—is at the core of the most influential theoretical 30 accounts of early phonetic learning (5-7, 25, 31). 31

The notion of *phonetic category* plays an important role 32 throughout the paper, so requires further definition. It has 33 been used in the literature exclusively to refer to vowel- or 34 consonant-like units. What that means varies to some extent 35 between authors, but there are at least two constant, defin-36 ing characteristics (32). First, phonetic categories have the 37 characteristic size/duration of a vowel or consonant, i.e. the 38 size of a *phoneme*, the 'smallest distinctive unit within the 39 structure of a given language' (17, 33). This can be contrasted 40 with larger units like syllables or words and smaller units like 41 speech segments corresponding to a single period of vocal fold 42 vibration in a vowel. Second, phonetic categories-although 43

#### Significance Statement

Infants become attuned to the sounds of their native language(s) before they even speak. Hypotheses about *what* is being learned by infants have traditionally driven researchers' attempts to understand this surprising phenomenon. Here, we propose to start instead from hypotheses about *how* infants might learn. To implement this *mechanism-driven* approach, we introduce a quantitative modeling framework based on largescale simulation of the learning process on realistic input. It allows, for the first time, learning mechanisms to be systematically linked to testable predictions regarding infants' attunement to their native language(s). Through this framework, we obtain evidence for an account of infants' attunement that challenges established theories about what infants are learning.

<sup>1</sup>To whom correspondence should be addressed. E-mail: thomas.schatz.1986@gmail.com

T.S. and E.D. designed the study; T.S. and X.C. prepared the speech recordings; T.S. trained the models and carried out the discrimination tests. T.S. designed and carried out the statistical analyses. T.S., N.F., S.G. and E.D. designed the tests of the nature of learned representations and T.S. and E.D. implemented them. All authors contributed to writing the manuscript.

The authors declare that they have no conflict of interest.

101 102 103

they may be less abstract than phonemes<sup>\*</sup>—retain a degree of 44 abstractness and never refer to a single acoustic exemplar. For 45 example, we would expect a given vowel or consonant in the 46 47 middle of a word repeated multiple times by the same speaker 48 to be consistently realized as the same phonetic category, de-49 spite some acoustic variation across repetitions. Finally, an added characteristic in the context of early phonetic learning 50 is that phonetic categories are defined relative to a language. 51 What might count as exemplars from separate phonetic cate-52 gories for one language, might belong to the same category in 53

another. 54 The phonetic category hypothesis—that infants learn to 55 process speech in terms of the phonetic categories of their 56 native language—raises a question. How can infants learn 57 about these phonetic categories so early? The most influential 58 proposal in the literature has been that infants form phonetic 59 categories by grouping the sounds they hear on the basis 60 of how they are distributed in a universal (i.e. language-61 62 independent) perceptual space, a statistical clustering process dubbed 'distributional learning' (8, 10, 34, 35). 63

Serious concerns have been raised regarding the feasibility 64 of this proposal, however (12, 36). Existing phonetic category 65 accounts of early phonetic learning assume that speech is being 66 represented phonetic segment by phonetic segment—i.e. for 67 each vowel and consonant separately—along a set of language-68 independent phonetic dimensions (6, 7, 25).<sup>†</sup> Whether it is 69 possible for infants to form such a representation in a way that 70 would enable distributional learning of phonetic categories 71 is questionable, for at least two reasons. First, there is a 72 lack of acoustic-phonetic invariance (37-39): there is not a 73 simple mapping from speech in an arbitrary language to an 74 75 underlying set of universal phonetic dimensions that could act as reliable cues to phonetic categories. Second, phonetic 76 category segmentation—finding reliable language-independent 77 cues to boundaries between phonetic segments (i.e. individual 78 vowels and consonants)—is a hard problem (37). It is clear 79 that finding a solution to these problems for a given language 80 is ultimately feasible, as literate adults readily solve them for 81 their native language. Assuming that infants are able to solve 82 them from birth in a language-universal fashion is a much 83 stronger hypothesis, however, with little empirical support. 84

Evidence from modeling studies reinforces these concerns. 85 Initial modeling work investigating the feasibility of learning 86 phonetic categories through distributional learning sidestepped 87 88 the lack of invariance and phonetic category segmentation problems by focusing on drastically simplified learning conditions 89 (40-45), but subsequent studies considering more realistic 90 variability have failed to learn phonetic categories accurately 91 (9, 12, 14, 15, 46, 47) (see Supplementary Discussion 1). 92

These results have largely been interpreted as a challenge 93 to the idea that distributional learning is how infants learn 94 phonetic categories. Additional learning mechanisms tapping 95 into other sources of information plausibly available to infants 96 have been proposed (9-12, 14, 15, 36, 46, 47), but existing 97 feasibility results for such complementary mechanisms still 98 assume that the phonetic category segmentation problem has 99 somehow been solved and do not consider the full variability of 100

natural speech (9, 12, 14, 15, 43, 46–48). Attempts to extend them to more realistic learning conditions have failed (13, 16) (see Supplementary Discussion 1).

Here, we propose a different interpretation for the observed 104 difficulty in forming phonetic categories through distributional 105 learning: it might indicate that *what* infants learn are not 106 phonetic categories. We are not aware of empirical results 107 establishing that infants learn phonetic categories, and indeed, 108 the *phonetic category hypothesis* is not universally accepted. 109 Some of the earliest accounts of early phonetic learning were 110 based on syllable-level categories and/or on continuous rep-111 resentations without any explicit category representations<sup>‡</sup> 112 (49-52). Although they appear to have largely fallen out of 113 favor, we know of no empirical findings refuting them. 114

We present evidence in favor of this alternative interpreta-115 tion, first by showing that a distributional learning mechanism 116 applied to raw, unsegmented, unlabeled continuous speech 117 signal predicts early phonetic learning as observed in Ameri-118 can English- and Japanese-learning infants-thereby providing 119 the first realistic proof of feasibility for any account of early 120 phonetic learning. We then show that the speech units learned 121 through this mechanism are too brief and too acoustically 122 variable to correspond to phonetic categories. 123

We rely on two key innovations. First, whereas previous 124 studies followed an *outcome-driven* approach to the study 125 of early phonetic learning-starting from assumptions about 126 what was learned, before seeking plausible mechanisms to 127 learn it—we adopt a *mechanism-driven* approach—focusing 128 first on the question of *how* infants might plausibly learn 129 from realistic input, and seeking to characterize what was 130 learned only a posteriori. Second, we introduce a quantitative 131 modeling framework suitable to implement this approach at 132 scale using realistic input. This involves explicitly simulating 133 both the ecological learning process taking place at home and 134 the assessment of infants' discrimination abilities in the lab. 135

Beyond the immediate results, the framework we introduce is the first to provide a feasible way of linking accounts of early phonetic learning to systematic predictions regarding the empirical phenomenon they seek to explain, i.e. the observed cross-linguistic differences in infants' phonetic discrimination.

#### Approach

We start from a possible learning mechanism. We simulate 142 the learning process in infants by implementing this mecha-143 nism computationally and training it on naturalistic speech 144 recordings in a target language—either Japanese or American 145 English. This yields a candidate model for the early phonetic 146 knowledge of, say, a Japanese infant. Next, we assess the 147 model's ability to discriminate phonetic contrasts of Amer-148 ican English and Japanese—for example American English 149 [J] vs [l]—by simulating a discrimination task using speech 150 stimuli corresponding to this contrast. We test whether the 151 predicted discrimination patterns agree with the available em-152 pirical record on cross-linguistic differences between American 153

<sup>\*</sup>For example, the same phoneme might be realized as different phonetic categories depending on the preceding and following sounds or on characteristics of the speaker.

<sup>&</sup>lt;sup>†</sup> In some accounts, the phonetic dimensions are assumed to be 'acoustic' (25)—e.g. formant frequencies—in other they are 'articulatory' (6)—e.g. the degree of vocal tract opening at a constriction—and some accounts remain noncommittal (7).

<sup>&</sup>lt;sup>‡</sup>Note that the claims in all the relevant theoretical accounts are for the formation of *explicit* representations, in the sense that they are assumed to be available for manipulation by downstream cognitive processes at later developmental stages (see e.g. (7)). Thus, even if one might be tempted to say that phonetic categories are *implicitly* present in some sense in a representation—for example in a continuous representation exhibiting sharp increases in discriminability across phonetic category boundaries (49)—unless a plausible mechanism by which downstream cognitive processes could explicitly read out phonetic categories from that representation is provided, together with evidence that infants actually use this mechanism, this would not be sufficient to support the early phonetic category acquisition hypothesis.



Fig. 1. Gaussian mixture model training and representation extraction, illustrated for a model with three Gaussian components. In practice the number of Gaussian components is learned from the data and much higher. (a) Model training: the learning algorithm extracts moderate-dimensional (d=39) descriptors of the local shape of the signal spectrum at time points regularly sampled every 10ms (speech frames). These descriptors are then considered as having been generated by a mixture of Gaussian probability distributions, and parameters for this mixture that assign high probability to the observed descriptors are learned. (b) Model test: the sequence of spectral-shape descriptor a test stimulus (possibly in a language different from the training language) are extracted and the model representation for that stimulus is obtained as the sequence of posterior probability vectors resulting from mapping each descriptor to its probability of having been generated by each of the Gaussian components in the learned mixture.

English and Japanese-learning infante. Finally, we investigate whether *what* has been learned by the model corresponds to the phonetic categories of the model's 'native' language (i.e. its training language).

To identify a promising learning mechanism, we build on 158 recent advances in the field of machine learning, and more 159 specifically in unsupervised representation learning for speech 160 technology, which have established that, given only raw, un-161 transcribed, unsegmented speech recordings, it is possible to 162 learn representations that accurately discriminate the phonetic 163 categories of a language (53-70). The learning algorithms con-164 sidered have been argued to be particularly relevant for model-165 ing how infants learn in general, and learn language in partic-166 ular (71). Among available *learning algorithms*, we select the 167 one at the core of the winning entries in the Zerospeech 2015 168 and 2017 international competitions in unsupervised speech 169 representation learning (58, 59, 69). Remarkably, it is based 170 on a Gaussian mixture clustering mechanism—illustrated in 171 Figure 1 (a)—that can straightforwardly be interpreted as a 172 form of distributional learning (8, 10). A different *input repre*-173 sentation to the Gaussian mixture is used than in previously 174 proposed implementations of distributional learning, however 175 (9, 12, 14, 40, 42, 44, 45). Simple descriptors of the shape 176 of the speech signal's short-term auditory spectrum sampled 177 at regular points in time (every 10ms) (72) are used instead 178 of traditional phonetic measurements obtained separately for 179 each vowel and consonant, such as formant frequencies or 180 harmonic amplitudes.<sup>§</sup> This type of input representation only 181 assumes basic auditory abilities from infants, which are known 182 to be fully operational shortly after birth (75), and has been 183 proposed previously as a potential way to get around both 184 185 the lack of invariance and the phonetic category segmentation problems in the context of adult word recognition (37). A 186 second difference from previous implementations of distribu-187 tional learning is in the *output representation*. Test stimuli 188 are represented as sequences of posterior probability vectors 189 (posteriorgrams) over K Gaussian components in the mixture 190 (Figure 1 (b)), rather than simply being assigned to the most 191

Table 1. Language, speech register, duration and number of speakers of training and test sets for our four corpora of speech recordings

| Corpus                     | Language    | Reg.   | Duration |      | No. speakers |      |
|----------------------------|-------------|--------|----------|------|--------------|------|
|                            |             |        | Train    | Test | Train        | Test |
| R-Eng (84)                 | Am. English | Read   | 19h30    | 9h39 | 96           | 47   |
| R-Jap ( <mark>85</mark> )  | Japanese    | Read   | 19h33    | 9h40 | 96           | 47   |
| Sp-Eng ( <mark>86</mark> ) | Am. English | Spont. | 9h13     | 9h01 | 20           | 20   |
| Sp-Jap ( <mark>87</mark> ) | Japanese    | Spont. | 9h11     | 8h57 | 20           | 20   |

likely Gaussian component. These continuous representations have been shown to support accurate discrimination of native phonetic categories in the Zerospeech challenges.

To simulate the infants' learning process, we expose the 195 selected learning algorithm to a realistic model of the linguistic 196 input to the child, in the form of raw, unsegmented, untran-197 scribed, multi-speaker continuous speech signal in a target 198 language (either Japanese or American English). We select 199 recordings of adult speech made with near field, high quality 200 microphones in two speech registers which cover the range of 201 articulatory clarity that infants may encounter. On one end of 202 the range, we use spontaneous adult directed speech, and on 203 the other, we use read speech; these two speaking registers are 204 crossed with the language factor (English, Japanese), resulting 205 in four corpora, each split into a training set and a test set 206 (Table 1). We would have liked to use recordings made in 207 infant's naturalistic environments, but no such dataset of suf-208 ficient audio quality was available for this study. It is unclear 209 whether or how using infant-directed speech would impact re-210 sults: the issue of whether infant directed speech is beneficial 211 for phonetic learning has been debated, with arguments in 212 both directions (76-83). We train a separate model for each 213 of the four training sets, allowing us to check that our results 214 hold across different speech registers and recording conditions. 215 We also train separate models on 10 subsets of each training 216 set for several choices of subset sizes, allowing us to assess the 217 effects of varying the amount of input data and the variability 218 due to the choice of training data for a given input size. 219

We next evaluate whether the trained 'Japanese native' and 220 'American-English native' models correctly predict early phonetic learning as observed in Japanese-learning and American 222

<sup>&</sup>lt;sup>S</sup>There was a previous attempt to model infant phonetic learning from such spectrogram-like auditory representations of continuous speech (73, 74), but we are the first to combine this modeling approach with a suitable evaluation methodology.

English-learning infants, respectively, and whether they make 223 novel predictions regarding the differences in speech discrimi-224 nation abilities between these two populations. Because we do 225 not assume that the outcome of infants' learning is adult-like 226 227 knowledge, we can only rely on infant data for evaluation. The 228 absence of specific assumptions a priori about what is going to be learned, and the sparsity of empirical data on infant 229 discrimination, makes this challenging. The algorithm we 230 consider outputs complex, high-dimensional representations 231 (Figure 1 (b)) that are not easy to link to concrete predic-232 tions regarding infant discrimination abilities. Traditional 233 signal detection theory models of discrimination tasks (88)234 cannot handle high-dimensional perceptual representations, 235 while more elaborate (Bayesian) probabilistic models (89) have 236 too many free parameters given the scarcity of available data 237 from infant experiments. We rely instead on the machine ABX238 approach that we previously developed (90, 91). It consists 239 of a simple model of a discrimination task, which can handle 240 any representation format provided the user can provide a 241 reasonable measure of (dis)similarity between representations 242 (90, 91). This is not a detailed model of infant's performance 243 in a specific experiment, but rather a simple and effectively 244 parameterless way to systematically link the complex speech 245 representations produced by our models to predicted discrim-246 ination patterns. For each trained model and each phonetic 247 contrast of interest, we obtain an 'ABX error rate' such that 0% 248 and 50% error indicate perfect and chance-level discrimination, 249 respectively. This allows us to evaluate the qualitative match 250 between the model's discrimination abilities and the available 251 empirical record in infants (see Supplementary Discussion 3 252 for an extended discussion of our approach to interpreting the 253 simulated discrimination errors and relating them to empirical 254 observations, including why it would not be meaningful to 255 seek a quantitative match at this point). 256

Finally, we investigate whether the learned Gaussian com-257 ponents correspond to phonetic categories. We first compare 258 the number of Gaussians in a learned mixture to the num-259 ber of phonemes in the training language (category number 260 test): although a phonetic category can be more concrete than 261 a phoneme, the number of phonetic categories documented 262 in typical linguistic analyses remains on the same order of 263 magnitude as the number of phonemes. We then administer 264 two diagnostic tests based on the two defining characteris-265 tics identified above that any representation corresponding to 266 phonetic categories should pass.<sup>¶</sup> The first characteristic is 267 size/duration: a phonetic category is a phoneme-sized unit 268 (i.e. the size of a vowel or a consonant). Our duration test 269 probes this by measuring the average duration of activation of 270 the learned Gaussian components (a component is taken to be 271 'active' when its posterior probability is higher than all other 272 components), and comparing this to the average duration of 273 activation of units in a baseline system trained to recognize 274 phonemes with explicit supervision. The second characteris-275 tic is abstractness: although phonetic categories can depend 276 on phonetic context<sup>I</sup> and on non-linguistic properties of the 277 speech signal—e.g. the speaker's gender—at a minimum, the 278

central phone in the same word repeated several times by the 279 same speaker is expected to be consistently realized as the 280 same phonetic category. Our *acoustic (in)variance* test probes 281 this by counting the number of distinct representations needed 282 by our model to represent ten occurrences of the central frame 283 of the central phone of the same word either repeated by the 284 same speaker (within speaker condition) or by different speak-285 ers (across speaker condition). We use a generous correction 286 to handle possible misalignment (see Materials and Methods). 287 The last two tests can be related to the phonetic category 288 segmentation and lack of invariance problems: solving the 289 phonetic category segmentation problem involves finding units 290 that would pass the *duration* test, while solving the lack of 291 invariance problem involves finding units that would pass the 292 acoustic (in)variance test. Given the laxity in the use of the 293 concept of phonetic category in the literature, some might be 294 tempted to challenge that even these diagnostic tests can be 295 relied on. If they cannot, however, it is not clear to us how 296 phonetic category accounts of early phonetic learning should 297 be understood as scientifically refutable claims. 298

#### Results

**Overall discrimination.** After having trained a separate model 300 for each of the four possible combinations of language and 301 register, we test whether the models' overall discrimination 302 abilities, like those of infants (2, 3, 30), are specific to their 303 'native' (i.e. training) language. Specifically, for each corpus, 304 we look at overall discrimination errors averaged over all conso-305 nant and vowel contrasts available in a held-out test set from 306 that corpus (See Table 1). We tested each of the two American 307 English-trained and each of the two Japanese-trained models 308 on each of four test sets, yielding a total of  $4 \times 4$  discrimination 309 errors. We tabulated the average errors in terms of 4 conditions 310 depending on the relation between the test set and the training 311 background of the model: native versus non-native contrasts 312 and same versus different register. The results are reported in 313 Figure 2 (see also Figures S1, S4 for non-tabulated results). 314 Panel (a) shows that discrimination performance is higher 315 on average in matched-language conditions (in blue) than in 316 mismatched-language conditions (in red). In contrast, register 317 mismatch has no discernible impact on discrimination perfor-318 mance. A comparison with a supervised phoneme recognizer 319 baseline (Figure S3) shows a similar pattern of results, but 320 with a larger absolute cross-linguistic difference. If we interpret 321 this supervised baseline as a proxy to the adult state, then our 322 model suggests that infant's phonetic representations, while al-323 ready language-specific, remain 'immature'.\*\* Panel (b) shows 324 the robustness of these results, with 81.7% of the 1295 distinct 325 phonetic contrasts tested proving easier to discriminate on the 326 basis of representations from a model trained on the matching 327 language. Taken together, these results suggest that, similar to 328 infants, our models acquire language-specific representations, 329 and that these representations generalize across register. 330

American English [J]-[I] discrimination. Next, we focus on the<br/>specific case of American English [J]-[I] discrimination, for<br/>which Japanese adults show a well-documented deficit (18, 19)<br/>and which has been studied empirically in American English<br/>and Japanese infants (4). While 6- to 8-month-old infants331<br/>332

<sup>&</sup>lt;sup>¶</sup> This provides *necessary* but not *sufficient* conditions for 'phonetic categoriness', but since we will see that the representations learned in our simulations already fail these tests, more fine-grained assessments will not be required.

For example, in the American English word 'top' the phoneme /t/ is realized as an aspirated consonant [t<sup>h</sup>] (i.e. there is a slight delay before the vocal folds start to vibrate after the consonant), whereas in the word 'stop' it is realized as a regular voiceless consonant [t], which might be considered to correspond to a different phonetic category than [t<sup>h</sup>].

<sup>\*\*</sup> This is compatible with empirical evidence that phonetic learning continues into childhood well beyond the first year (see 92-94, for example).



Fig. 2. (a) Average ABX error rates over all consonant and vowel contrasts obtained with our models as a function of the match between the training set and test set language and register. Error bars correspond to plus and minus one standard deviation The 'Native' (blue) of the errors across resampling of the test stimuli speakers. conditions, with training and test in the same language, show fewer discrimination errors than the 'Non-native' (red) conditions, whereas there is little difference in error rate within the 'Native' and within the 'Non-native' conditions. This shows that the models learned native-language specific representations that generalize across register. (b) Letter-value representation (95) of the distribution of 'native advantages across all tested phonetic contrasts (pooled over both languages). The native language advantage is the increase in discrimination error for a contrast of language L1 between a 'L1-native' model and a model trained on the other language. for the same training register. The 'native register' advantage is the increase in error for a contrast of register R1 between a 'R1-native' model and a model trained on the other register, for the same training language. A native language advantage is observed across contrasts (positive advantage for 81.7% of all contrasts) and there is a weaker native register advantage (positive advantage for 60.1% of all contrasts)

from American English and Japanese language backgrounds 336 performed similarly in discriminating this contrast, 10- to 337 12-month-old American English infants outperformed their 338 Japanese peers. We compare the discrimination errors ob-339 tained with each of our four models for American English 340 [J]-[l] and for two controls: the American English [w]-[j] con-341 trast (as in 'wet' versus 'yet'), for which we do not expect a 342 gap in performance between American English and Japanese 343 natives (96), and the average error over all the other conso-344 nant contrasts of American English. For each contrast and 345 for each of the four models, we average discrimination errors 346 obtained on each of the two American English held-out test 347 sets, yielding  $3 \times 4$  discrimination errors. We further average 348 over models with the same 'native' language to obtain  $3 \times 2$ 349 discrimination errors. The results are shown in Figure 3 (see 350 351 also Figures S2 and S6 for untabulated results and a test 352 confirming our results with the synthetic stimuli used in the original infant experiment, respectively). In panel (a), we see 353 that, similar to 10- to 12-month old infants, American English 354 'native' models (in blue) greatly outperform Japanese 'na-355 tive' models (in red) in discriminating American English [1]-[1]. 356 Here again, a supervised phoneme recognizer baseline yields 357 a similar pattern of results, but with larger cross-linguistic 358 359 differences (see Figure S5), again suggesting that the representations learned by the unsupervised models-like those of 360 infants—remain somewhat 'immature'. In panel (b), we see re-361 sults obtained by training ten different models on ten different 362 subsets of the training set of each corpus, varying the sizes of 363 the subsets (see Materials and Methods for more details). It 364 reveals that one hour of input is sufficient for the divergence 365 between the Japanese and English models to emerge robustly, 366 and that this divergence increases with exposure to the native 367

language. While it is difficult to interpret this trajectory relative to absolute quantities of data or discrimination scores, the fact that the cross-linguistic difference increases with more data mirrors the empirical findings from infants (see also an extended discussion of our approach to interpreting the simulated discrimination errors and relating them to empirical data in Supplementary Discussion 3).

Nature of the learned representations. Finally, we consider the 375 nature of the learned representations and test whether what 376 has been learned can be understood in terms of phonetic cat-377 egories. Results are reported in Figure 4 (see also Figure S7) 378 for comparisons with a different supervised baseline). First, 379 looking at the *category number* criterion in Figure 4 (a), we 380 see that our models learned more than ten times as many 381 categories as the number of phonemes in the corresponding 382 languages. Even allowing for notions of phonetic categories 383 more granular than phonemes, we are not aware of any pho-384 netic analysis ever reporting that many allophones in these 385 languages. Second, looking at the *duration* criterion in Fig-386 ure 4 (b), the learned Gaussian units appear to be activated 387 on average for about a quarter the duration of a phoneme. 388 This is shorter than any linguistically identified unit. It shows 389 that the phonetic category segmentation problem has not been 390 solved. Next, looking at the acoustic (in)variance criterion in 391 Figure 4 (c) and (d)—for the within and across speakers condi-392 tions, respectively—we see that our models require on average 393 around two distinct representations to represent ten tokens of 394 the same phonetic category without speaker variability, and 395 three distinct representations across different speakers. The 396 supervised phoneme recognizer baseline establishes that our 397 results cannot be explained by defective test stimuli. Instead, 398 this result shows that the learned units are finer-grained than 399 phonetic categories along the spectral axis, and that the lack of 400 invariance problem has not been solved. Based on these tests, 401 we can conclude that the learned units do not correspond to 402 phonetic categories in any meaningful sense of the term. 403

#### Discussion

Through explicit simulation of the learning process under re-405 alistic learning conditions, we showed that several aspects of 406 early phonetic learning as observed in American English and 407 Japanese infants can be correctly predicted through a distribu-408 tional learning (i.e. clustering) mechanism applied to simple 409 spectrogram-like auditory features sampled at regular time 410 intervals. This is the first time that a potential mechanism for 411 early phonetic learning is shown to be feasible under realistic 412 learning conditions. We further showed that the learned speech 413 units are too brief and too acoustically variable to correspond 414 to the vowel- and consonant-like 'phonetic categories' posited 415 in earlier accounts of early phonetic learning. 416

404

Distributional learning has been an influential hypothesis 417 in language acquisition for over a decade (8, 10, 35). Previous 418 modeling results questioning the feasibility of learning phonetic 419 categories through distributional learning have traditionally 420 been interpreted as challenging the learning mechanism (9– 421 12, 14, 15, 36, 46, 47), but we have instead suggested that 422 such results may be better interpreted as challenging the 423 idea that phonetic categories are the outcome of early pho-424 netic learning. Supporting this view, we showed that when 425 the requirement to learn phonetic categories is abandoned, 426



Fig. 3. (a) ABX error rates for the American English [*i*]-[*i*] contrast and two controls: American English [*w*]-[*j*] and average over all American English consonant contrasts (C-C). Error rates are reported for two conditions: average over models trained on American English and average over models trained on Japanese. Error bars correspond to plus and minus one standard deviation of the errors across resampling of the test stimuli speakers. Similar to infants, the Japanese 'native' models exhibit a specific deficit for American English [*j*]-[*l*] discrimination compared to the 'American English' models. (b) The robustness of the effect observed in panel (a) to changes in the training stimuli and their dependence on the amount of input are assessed by training separate models on independent subsets of the training data of each corpus of varying duration (see Materials and Methods). For each selected duration (except when using the full training set), ten independent subsets are selected and ten independent models are trained. We report mean discrimination errors for American English [*i*]-[*l*] and [*w*]-[*j*] as a function the amount of input data, with error bands indicating plus or minus one standard deviation. The results show that a deficit in American English [*i*]-[*l*] discrimination for 'Japanese-native' models robustly emerges with as little as 1 h of training data.

distributional learning on its own can be very effective, lead-427 ing to the first realistic demonstration of feasibility-using 428 unsegmented, untranscribed speech signal as input-for any 429 mechanism for early phonetic learning. Our results are still 430 compatible with the idea that mechanisms tapping into other 431 relevant sources of information might complement distribu-432 tional learning—an idea supported by evidence that infants 433 learn from some of these sources in the lab (97-103)—but 434 they suggest that those other sources of information may not 435 436 play a role as crucial as previously thought (10). Our findings also join recent accounts of 'word segmentation' (104) and 437 the 'language familiarity effect' (105) in questioning whether 438 we might have been over-attributing linguistic knowledge to 439 pre-verbal infants across the board. 440

A new account of early phonetic learning. Our results suggest 441 an account of phonetic learning that substantially differs from 442 existing ones. Whereas previous proposals have been primarily 443 motivated through an *outcome-driven* perspective—starting 444 from assumptions about what it is about language that is 445 learned—the motivation for the proposed account comes from 446 447 a *mechanism-driven* perspective—starting from assumptions about how learning might proceed from the infant's input. 448 This contrast is readily apparent in the choice of the initial 449 speech representation upon which the early phonetic learning 450 process operates (the input representation). Previous accounts 451 assumed speech to be represented innately through a set of 452 universal (i.e. language-independent) phonetic feature detec-453 tors (5-7, 25, 31, 49-52). The influential phonetic category 454 accounts furthermore assumed these features to be available 455

phonetic segment by phonetic segment (i.e. for each vowel and 456 consonant separately) (5-7, 25, 31). While these assumptions 457 are attractive from an *outcome-driven* perspective—they con-458 nect transparently to phonological theories in linguistics and 459 theories of adult speech perception that assume a decomposi-460 tion of speech into phoneme-sized segments defined in terms 461 of abstract phonological features—from a mechanism-driven 462 perspective, both assumptions are difficult to reconcile with 463 the continuous speech signal that infants hear. The lack of 464 acoustic-phonetic invariance problem challenges the idea of 465 phonetic feature detectors, and the phonetic category segmen-466 tation problem challenges the idea that the relevant features 467 are segment-based (37-39). The proposed account does not 468 assume either problem to be solved by infants at birth. In-469 stead, it relies on basic auditory abilities that are available to 470 neonates (75), using simple auditory descriptors of the speech 471 spectrum obtained regularly along the time axis. This type 472 of spectrogram-like representation is effective in speech tech-473 nology applications (72) and can be seen as the output of 474 a simple model of the peripheral auditory system (91, chap. 475 3), which is fully operational shortly after birth (75). Such 476 representations have also been proposed before as an effective 477 way to get around both the lack of invariance and the phonetic 478 category segmentation problems in the context of adult word 479 recognition (37) and can outperform representations based on 480 traditional phonetic measurements (like formant frequencies) 481 as predictors of adult speech perception (106-110). 482

While the input representation is different, the learning 483 mechanism in the proposed account—distributional learning is similar to what had originally been proposed in phonetic 483



row: American English mative models. Bottom row: Japanese roative models. Models are tested on read speech in their 'native' language. (a) Number of units learned by the models. Gaussian mixtures discover ten to wenty times more categories than there are phonemes in the training language, exceeding any reasonable count for phonetic categories. (b) Average duration of activation of the learned units. The average duration of activation of the duration of a same word by the resulting distribution over units are shown. Learned Gaussian units get activated on average for about the quarter of the duration of a phoneme. They are thus much too 'short' to correstand to phonetic categories. (b) Average number of distinct representations for the central frame of the central phone for ten repetitions of a same word by the same speaker, gorrected for possible missing distribution over word types are shown. The phoneme recogniser baseline reliably identifies the ten tokens as exemplars form a common phonetic category, whereas our Gaussian mixture models typically maintain on the order of two distinct representations to baseline reliably identifies the ten tokens as exemplars form a common phonetic categories. (d) As in (c) parkwith repetitions of a same word by the same speakers, showing that the learned Gaussian units are not speaker-independent.

category accounts. Infants' abilities, both in the lab (8, 35)486 and in ecological conditions (34), are consistent with such a 487 learning mechanism. Moreover, when applied to the input 488 representation considered in this paper, distributional learning 489 is adaptive in that it yields speech representations that can 490 support remarkably accurate discrimination of the phonetic 491 categories of the training language, outperforming a number of 492 alternatives that have been proposed for unsupervised speech 493 representation learning (58, 59, 69). 494

As a consequence of our mechanism-driven approach, what 495 has been learned needs to be determined a posteriori based 496 on the outcomes of learning simulations. The speech units 497 learned under the proposed account accurately model infants' 498 discrimination, but are too brief and acoustically variable 499 to correspond to phonetic categories, failing in particular to 500 provide a solution to the lack of invariance and phonetic 501 category segmentation problems (37). Such brief units do 502 not correspond to any previously identified linguistic unit (32)503 (see Supplementary Discussion 4 for a discussion of possible 504 reasons why the language acquisition process might involve 505 the learning by infants of a representation with no established 506 linguistic interpretation, and a discussion of the biological 507 and psychological plausibility of the learned representation), 508 and it will be interesting to try to further understand their 509 nature. However, since there is no guarantee that a simple 510 characterization exists, we leave this issue for future work. 511

Phonetic categories are often assumed as precursors in ac-512 counts of phenomena occurring later in the course of language 513 acquisition. Our account does not necessarily conflict with 514 this view, as phonetic categories may be learned later in de-515 velopment, before phonological acquisition. Alternatively, the 516 influential PRIMIR account of early language acquisition (7) 517 proposes that infants learn in parallel about the phonetics, 518 word-forms, and phonology of their native language, but do 519

not develop abstract phonemic representations until well into<br/>their second year of life. Although *PRIMIR* explicitly assumes<br/>phonetic learning to be *phonetic category* learning, other as-<br/>pects of their proposed framework do not depend on that<br/>assumption, and our framework may be able to stand in for<br/>the phonetic learning process they assume.522<br/>524

To sum up, we introduced and motivated a new account of 526 early phonetic learning and showed that it is feasible under 527 realistic learning conditions, which cannot be said of any other 528 account at this time. Importantly, this does not constitute 529 decisive evidence for our account over alternatives. Our pri-530 mary focus has been on modeling cross-linguistic differences 531 in the perception of one contrast, [1]-[1]; further work is neces-532 sary to determine to what extent our results extend to other 533 contrasts and languages (111). Furthermore, an absence of 534 feasibility proof does not amount to a proof of infeasibility. 535 While we have preliminary evidence that simply forcing the 536 model to learn fewer categories is unlikely to be sufficient (Fig-537 ures S9 and S10), recently proposed partial solutions to the 538 phonetic category segmentation problem (e.g. (112-114)) and 539 to the lack of invariance problem (115) (see also Supplemen-540 tary Discussion 2 regarding the choice of model initialization) 541 might yet lead to a feasible phonetic category-based account, 542 for example. In addition, a number of other representation 543 learning algorithms proposed in the context of unsupervised 544 speech technologies and building on recent developments in the 545 field of machine learning have yet to be investigated (53-70). 546 They might provide concrete implementations of previously 547 proposed accounts of early phonetic learning or suggest new 548 ones altogether. This leaves us with a large space of appeal-549 ing theoretical possibilities, making it premature to commit 550 to a particular account. Candidate accounts should instead 551 be evaluated on their ability to predict empirical data on 552 early phonetic learning, which brings us to the second main 553

60

554 contribution of this article.

Toward predictive theories of early phonetic learning. Almost 555 since the original empirical observation of early phonetic 556 learning (1), a number of theoretical accounts of the phe-557 nomenon have co-existed (6, 25, 49, 50). This theoretical 558 under-determination has typically been thought to result from 559 the scarcity of empirical data from infant experiments. We ar-560 gue instead that the main limiting factor on our understanding 561 of early phonetic learning might have been the lack—on the 562 theory side—of a practical method to link proposed accounts 563 of phonetic learning with concrete, systematic predictions re-564 garding the empirical discrimination data they seek to explain. 565 Establishing such a systematic link has been challenging due 566 to the necessity of dealing with the actual speech signal, with 567 all its associated complexity. The modeling framework we 568 introduce provides, for the first time, a practical and scalable 569 way to overcome these challenges and obtain the desired link 570 for phonetic learning theories—a major methodological ad-571 vance, given the fundamental epistemological importance of 572 linking explanandum and explanans in scientific theories (116). 573

Our mechanism-driven approach to obtaining predictions— 574 which can be applied to any phonetic learning model imple-575 mented in our framework—consists first of explicitly simulating 576 the early phonetic learning process as it happens outside of 577 the lab, which results in a trained model capable of mapping 578 any speech input to a model representation for that input. 579 The measurement of infants' perceptual abilities in labora-580 tory settings—including their discrimination of any phonetic 581 contrast—can then be simulated on the basis of the model's 582 representations of the relevant experimental stimuli. Finally, 583 phonetic contrasts for which a significant cross-linguistic differ-584 ence is robustly predicted can be identified through a careful 585 statistical analysis of the simulated discrimination judgments 586 (see Supplementary Materials and Methods 4). As an illus-587 tration of how such predictions can be generated, we report 588 specific predictions made by our distributional learning model 589 in Table S1 (see also Supplementary Discussion 5). 590

Although explicit simulations of the phonetic learning pro-591 cess have been carried out before (9, 12, 14, 15, 40-49, 73, 74), 592 those have typically been evaluated based on whether they 593 learned phonetic categories, and have not been directly used 594 to make predictions regarding infants' discrimination abilities. 595 An outcome-driven approach to making predictions regarding 596 discrimination has typically been adopted instead, starting 597 598 from the assumption that phonetic categories are the outcome 599 of learning. To the best of our knowledge this has never resulted in the kind of systematic predictions we report here, 600 however (see Supplementary Discussion 6 for a discussion of 601 the limits of previous approaches and of the key innovations 602 underlying the success of our framework). 603

Our framework readily generates novel, empirically testable, 604 predictions regarding infants' discrimination, yet further com-605 606 putational modeling is called for before we return to experiments. Indeed, existing data—collected over more than three 607 decades of research (2, 3, 21, 30)—might already suffice to dis-608 tinguish between different learning mechanisms. To make that 609 determination, and to decide which contrasts would be most 610 useful to test next in case more data are needed, many more 611 learning mechanisms and training/test language pairs will 612 need to be studied. Even for a specified learning mechanism 613 and training/test datasets, multiple implementations should 614

ideally be compared (e.g. testing different parameter settings 615 for the input representations or the clustering algorithm), as 616 implementational choices that weren't initially considered to 617 be important might nevertheless have an effect on the result-618 ing predictions and thus need to be included in our theories. 619 Conversely, features of the model that may seem important a620 priori (e.g. the type of clustering algorithm used) might turn 621 out to have little effect on the learning outcomes in practice. 622

Cognitive science has not traditionally made use of such 623 large-scale modeling, but recent advances in computing power, 624 large datasets, and machine learning algorithms make this 625 approach more feasible than ever before (71). Together with 626 ongoing efforts in the field to collect empirical data on a 627 large scale—such as large-scale recordings of infants' learning 628 environment at home (117) and large-scale assessment of in-629 fants' learning outcomes (118, 119)—our modeling approach 630 opens the path towards a much deeper understanding of early 631 language acquisition. 632

#### Materials and Methods

633 634

Datasets. We used speech recordings from four corpora: two corpora 635 of read news articles—a subset of the Wall Street Journal corpus 636 of American English (84) (WSJ) and the Globalphone corpus of 637 Japanese (85) (GPJ)—and two corpora of spontaneous speech—the 638 Buckeye corpus of American English (86) (BUC) and a subset of 639 the corpus of spontaneous Japanese (87) (CSJ). As we are primarily 640 interested in the effect of training language on discrimination abili-641 ties, we sought to remove possibly confounding differences between 642 the two read corpora and between the two spontaneous corpora. 643 Specifically, we randomly sampled sub-corpora while matching total 644 duration, number and gender of speakers and amount of speech per 645 speaker. We made no effort to match corpora within a language, 646 as the differences (for example in the total duration and number 647 of speakers) only serve to reinforce the generality of any result 648 holding true for both registers. Each of the sampled subsets was 649 further randomly divided into a training and a test set (see Table 650 1), satisfying three conditions: the test set lasts approximately ten 651 hours; no speaker is present in both the training and test set; the 652 training and test sets for the two read corpora, and separately for 653 the two spontaneous corpora, remain matched on overall duration, 654 number of speakers of each gender and distribution of duration per 655 speaker of each gender. To carry out analyses taking into account 656 the effect of input size and of the choice of input data, we further 657 divided each training set in ten with each  $1/10^{th}$  subset containing 658 an equal proportion of the speech samples from each speaker in the 659 original training set. We then divided each of the  $1/10^{th}$  subset in 660 ten again following the same procedure and select the first subset 661 to obtain ten  $1/100^{th}$  subsets. Finally, we iterated the procedure 662 one more time to obtain ten  $1/1000^{th}$  subsets. See Supplementary 663 Materials and Methods 1 for additional information. 664

Signal processing, models and inference. The raw speech signal is 665 decomposed into a sequence of overlapping 25ms-long frames sam-666 pled every 10ms and moderate-dimensional (d=39) descriptors of 667 the spectral shape of each frame are then extracted, describing how 668 energy in the signal spreads across different frequency channels. 669 The descriptors are comprised of 13 mel-frequency cepstral coeffi-670 cients (MFCC) with their first and second time derivatives. These 671 coefficients correspond approximately to the principal components 672 of spectral slices in a log-spectrogram of the signal, where the spec-673 trogram frequency channels are selected on a mel frequency scale 674 (linear for lower frequency and logarithmic for higher frequencies. 675 matching the frequency selectivity of the human ear). 676

For each corpus, the set of all spectral-shape descriptors for the corpus' training set is modeled as a large i.i.d. sample from a probabilistic generative model. The generative model is a Gaussian mixture model with no restrictions on the form of covariance matrices and with a Dirichlet process prior over its parameters with



Fig. 5. Generative Gaussian mixture model with Dirichlet process prior with normalinverse-Wishart base measure, represented as a graphical model in plate notation based on the stick-breaking construction of Dirichlet processes.

Normal-inverse-Wishart base measure. The generative model is depicted as a graphical model in plate notation in Figure 5, where nis the number of input descriptors,  $(X_1, X_2, ..., X_n)$  are the random variables from which the observed descriptors are assumed to be sampled and the other elements are latent variables and hyperparameters. The depicted variables have the following conditional distributions:

$$\begin{array}{ccccccc} X_i & | & z_i, (\mu_1, \mu_2, \ldots), (\Lambda_1, \Lambda_2, \ldots) & \sim & \mathcal{N}(\mu_{z_i}, \Lambda_{z_i}^{-1}) \\ \mu_k & | & \Lambda_k, \mu_0, \lambda & \sim & \mathcal{N}(\mu_0, (\lambda\Lambda_k)^{-1}) \\ \Lambda_k & | & \Lambda_0, \nu & \sim & \mathcal{W}(\Lambda_0, \nu) \\ z_i & | & \pi & \sim & \text{Multi}(\pi) \\ \pi & | & \alpha & \sim & \text{SB}(\alpha) \end{array}$$

for any  $1 \leq i \leq n$ , for any  $k \in \{1, 2, ...\}$ , with  $\mathcal{N}$  the multivari-690 ate Gaussian distribution,  $\mathcal{W}$  the Wishart distribution, *Multi* the 691 692 generalisation of the usual multinomial probability distribution to an infinite discrete support and SB, the mixing weights generating 693 distribution from the stick-breaking representation of Dirichlet pro-694 cesses (120). Mixture parameters with high posterior probability 695 given the observed input features vectors and the prior are found 696 697 using an efficient parallel Markov chain Monte Carlo sampler (121). Following previous work (61, 66), model initialization is performed 698 by partitioning training points uniformly at random into ten clus-699 ters and the hyperparameters are set as follows:  $\alpha$  to 1,  $\mu_0$  to the 700 average of all input features vectors,  $\lambda$  to 1,  $\lambda_0$  to the inverse of the 701 covariance of all input feature vectors and  $\nu$  to 42 (i.e. the spectral 702 shape descriptors dimension plus three). We additionally train a 703 model on each of the ten  $1/10^{th}$ ,  $1/100^{th}$  and  $1/1000^{th}$  training 704 705 subsets of each of the four corpora, following the same procedure. Given a trained Gaussian mixture with K components, mix-706 ing weights  $(\pi_1, \pi_2, ..., \pi_K)$ , means  $(\mu_1, \mu_2, ..., \mu_K)$  and covariance 707 matrices  $(\Sigma_1, \Sigma_2, ..., \Sigma_K)$ , we extract a test stimulus representa-708 tion from the sequence  $(x_1, x_2, ..., x_m)$  of spectral-shape descrip-709 710 tors for that stimulus, as the sequence of posterior probability vectors  $(p_1, p_2, ..., p_m)$  where for any frame  $i, 1 \leq i \leq m$ , 711  $p_i = (p_{i1}, p_{i2}, ..., p_{iK})$ , with, for any  $1 \le k \le K$ : 712

713 
$$p_{ik} = \frac{\pi_k \mathcal{N}(x_i | \mu_k, \Sigma_k)}{\sum_{j=1}^K \pi_j \mathcal{N}(x_i | \mu_j, \Sigma_j)}$$

As a baseline, we also train a phoneme recognizer on the training set of each corpus, with explicit supervision (i.e. phonemic transcriptions of the training stimuli). We extract frame-level posterior probabilities at two granularity levels: actual phonemes—the *phoneme recognizer* baseline—and individual states of the contextual hidden Markov models—the *ASR phone state* baseline. See Supplementary Materials and Methods 2 for additional information.

Discrimination tests. Discriminability between model representa-721 tions for phonetic contrasts of interest is assessed using machine 722 723 ABX discrimination errors (90, 91). Discrimination is assessed in context, defined as the preceding and following sound and the iden-724 tity of the speaker. For example, discrimination of American English 725 [u] versus [i] is assessed in each available context independently, 726 yielding-for instance-a separate discrimination error rate for test 727 stimuli in [b]\_[t] phonetic context, as in 'boot' versus 'beet', as 728 spoken by a specified speaker. Other possible factors of variability, 729 such as word boundaries or syllable position are not controlled. For 730 731 each model, each test corpus and each phonemic contrast in that test corpus (as specified by the corpus' phonemic transcriptions), 732 we obtain a discrimination error for each context in which the con-733 trasted phonemes occur at least twice in the test corpus' test set. 734

To avoid combinatorial explosion in the number of ABX triplets to be considered, a randomly selected subset of five occurrences is used to compute discrimination errors when a phoneme occurs more than five times in a given context. An aggregated ABX error rate is obtained for each combination of model, test corpus and phonemic contrast, by averaging the context-specific error rates over speakers and phonetic contexts, in that order. 740

Model representations are extracted for the whole test sets, and the part corresponding to a specific occurrence of a phonetic category is then obtained by selecting representation frames centered on time points located between the start and end times for that occurrence, as specified by the test set's forced aligned phonemic transcriptions. Given model representations  $\Delta = (\delta_1, \delta_2, ..., \delta_{n_{\delta}})$ and  $\Xi = (\xi_1, \xi_2, ..., \xi_{n_{\xi}})$  for  $n_{\delta}$  tokens of phonetic category  $\delta$  and  $n_{\xi}$  tokens of phonetic category  $\xi$ , the non-symmetrized Machine ABX discrimination error between  $\delta$  and  $\xi$  is then estimated as the proportion of representation triplets a, b, x, with a and x taken from  $\Delta$  and b taken from  $\Xi$ , such that x is closer to b than to a, i.e.:

$$\hat{e}(\Delta, \Xi) := \frac{1}{n_{\delta}(n_{\delta} - 1)n_{\xi}} \sum_{a=1}^{n_{\delta}} \sum_{b=1}^{n_{\xi}} \sum_{\substack{x=1\\x \neq a}}^{n_{\delta}} \left[ \mathbbm{1}_{d(\xi_{b}, \delta_{x}) < d(\delta_{a}, \delta_{x})} + \frac{1}{2} \mathbbm{1}_{d(\xi_{b}, \delta_{x}) = d(\delta_{a}, \delta_{x})} \right]$$

where 1 is the indicator function returning 1 when its predicate is true and 0 otherwise and d is a dissimilarity function taking a pair of model representations as input and returning a real number (with higher values indicating more dissimilar representations). The (symmetric) Machine ABX discrimination error between  $\delta$  and  $\xi$ is then obtained as: 747

$$\hat{\epsilon}(\Delta, \Xi) = \hat{\epsilon}(\Xi, \Delta) := \frac{1}{2} [\hat{e}(\Delta, \Xi) + \hat{e}(\Xi, \Delta)].$$
 748

As realizations of phonetic categories vary in duration, we need 749 a dissimilarity function d that can handle model representations 750 with variable length. This is done, following established practice 751 (28, 29, 56, 58, 69), by measuring the average dissimilarity along a 752 time-alignment of the two representations obtained through dynamic 753 time warping (122), where the dissimilarity between model repre-754 sentations for individual frames is measured with the symmetrized 755 Kullback-Leibler divergence for posterior probability vectors and 756 with the angular distance for spectral shape descriptors. 757

Analysis of learned representations. Learned units are taken to be<br/>the Gaussian components for the Gaussian mixture models, the<br/>phoneme models for the phoneme recognizer baseline, and the phone<br/>state models for the ASR phone state baseline. Since experimental<br/>studies of phonetic categories are typically performed with citation<br/>form stimuli, we study how each model represents stimuli from the<br/>matched-language read speech corpus' test set.759<br/>760<br/>760To study average durations of activation we exclude any761<br/>762

To study average durations of activation we exclude any utterance-initial or utterance-final silence from the analysis, as well as any utterance for which utterance-medial silence was detected during the forced alignment. The average duration of activation for a given unit is computed by averaging over all episodes in the test utterances during which that unit becomes dominant, i.e. has the highest posterior probability among all units. Each of these episodes is defined as a continuous sequence of speech frames during which the unit remains dominant without interruptions, with duration equal to that number of speech frames times 10ms.

766

767

768

769

770

771

772

773

774

The acoustic (in)variance of the learned units is probed by 775 looking at multiple repetitions of a single word and testing whether 776 the dominant unit at the central frame of the central phone of the 777 word remains the same for all repetitions. Specifically, we count 778 the number of distinct dominant units occurring at the central 779 frame of the central phone for ten repetitions of the same word. To 780 compensate for possible misalignment of the central phones' central 781 frames (e.g. due to slightly different time courses in the acoustic 782 realization of the phonetic segment and/or small errors in the forced 783 alignment), we allow the dominant unit at the central frame to 784 be replaced by any unit that was dominant at some point within 785 the previous or following 46ms (thus covering a 92ms slice of time 786

corresponding to the average duration of a phoneme in our read 787

788 speech test sets), provided it can bring down the overall count of

distinct dominant units for the ten occurrences (see Supplementary 789 790 Materials and Methods 3 for more information). We consider

791 two conditions: in the within-speaker condition, the test stimuli

are uttered by the same speaker ten times; in the across-speaker 792

793 condition, they are uttered by ten different speakers one time. See

Supplementary Materials and Methods 3 for more information on 794

the stimulus selection procedure. 795

Data and code availability. The datasets analysed in this study are 796 797 publicly available from the commercial vendors and research institutions holding their copyrights (84-87). Datasets generated during 798 the course of the study are available from the corresponding author 799 upon reasonable request. Code to reproduce the results will be 800

made available in a public GitHub repository upon publication. 801

ACKNOWLEDGMENTS. We thank the editor and anonymous re-802 viewers for their helpful comments on the manuscript. We also thank 803 Yevgen Matusevych for helpful comments on the manuscript. This 804 research was supported by ERC-2011-AdG-295810 BOOTPHON, 805 ANR-10-0001-02 PSL\*, ANR-10-LABX-0087 IEC, ANR-17-EURE-806 0017, NSF BCS-1734245, ESRC ES/R006660/1, JSMF 220020374 807 and by a grant from Facebook AI Research. 808

- 809 1. JF Werker, RC Tees, Cross-language speech perception: evidence for perceptual reorganization during the first year of life. Infant behavior development 7, 49-63 (1984). 810
- 2. JF Werker, RC Tees, Influences on infant speech processing: Toward a new synthesis. 811 Annu. review psychology 50, 509–535 (1999). 812
- S Tsuji, A Cristia, Perceptual attunement in vowels: A meta-analysis. Dev. psychobiology 813 3. 56, 179-191 (2014). 814
- PK Kuhl, et al., Infants show a facilitation effect for native language phonetic perception 815 4. between 6 and 12 months. Dev. science 9, F13-F21 (2006). 816
- 5. PK Kuhl, et al., Phonetic learning as a pathway to language: new data and native language 817 magnet theory expanded (nlm-e). Philos. Transactions Royal Soc. B: Biol. Sci. 363, 979-818 1000 (2007). 819
- 6. CT Best., et al., The emergence of native-language phonological influences in infants; A 820 perceptual assimilation model. The development speech perception: The transition from 821 822 speech sounds to spoken words 167, 233-277 (1994).
- JF Werker, S Curtin, Primir: A developmental framework of infant speech processing. Lang. 823 7 824 learning development 1, 197-234 (2005).
- 825 8 J Mave, JF Werker, L Gerken, Infant sensitivity to distributional information can affect phonetic discrimination. Cognition 82, B101-B111 (2002). 826
- F Adriaans, D Swingley, Distributional learning of vowel categories is supported by prosody 827 9 in infant-directed speech in Proc. COGSCI. (2012). 828
- 10. 829 JF Werker, HH Yeung, KA Yoshida, How do infants become experts at native-speech perception? Curr. Dir. Psychol. Sci. 21, 221-226 (2012). 830
- 831 11. O Räsänen, Computational modeling of phonetic and lexical learning in early language 832 acquisition: existing models and future directions. Speech Commun. 54, 975-997 (2012).
- 833 12. NH Feldman, TL Griffiths, S Goldwater, JL Morgan, A role for the developing lexicon in 834 phonetic category acquisition. Psychol. review 120, 751 (2013).
- 835 13. RAH Bion, K Miyazawa, H Kikuchi, R Mazuka, Learning phonemic vowel length from natu-836 ralistic recordings of Japanese infant-directed speech. PLoS ONE 8, e51594 (2013).
- 837 14. S Frank, N Feldman, S Goldwater, Weak semantic context helps phonetic learning in a 838 model of infant language acquisition in Proc. ACL. (2014).
- 839 15 F Adriaans, D Swingley, Prosodic exaggeration within infant-directed speech: Conse-840 quences for vowel learnability. The J. Acoust. Soc. Am. 141, 3070-3078 (2017).
  - 16 S Antetomaso, et al., Modeling phonetic category learning from natural acoustic data. (Cascadilla Press), (2017).
  - 17 E Sapir, An introduction to the study of speech. (1921).

841

842

843

844

845

848

849

853

854

855

856

857

858

- H Goto, Auditory perception by normal Japanese adults of the sounds "L" and "R". Neu-18. ropsychologia 9, 317-323 (1971).
- K Miyawaki, et al., An effect of linguistic experience: The discrimination of [r] and [I] by native 846 19 speakers of Japanese and English. Percept. & Psychophys. 18, 331-340 (1975) 847
  - 20. W Strange, ES Levy, FF Law, Cross-language categorization of french and german vowels by naïve american listeners. The J. Acoust. Soc. Am. 126, 1461-1476 (2009).
- 850 21. W Strange, Speech perception and linguistic experience: Issues in cross-language research. 851 (York Press), (1995)
- JS Logan, SE Lively, DB Pisoni, Training japanese listeners to identify english/r/and/l: A first 852 22. report. The J. Acoust. Soc. Am. 89, 874-886 (1991).
  - 23. P Iverson, V Hazan, K Bannister, Phonetic training with acoustic cue manipulations: A comparison of methods for teaching english/r/-/l/to japanese adults. The J. Acoust. Soc. Am. 118, 3267-3278 (2005)
  - 24. ES Levy, W Strange, Perception of french vowels by american english adults with and with out french language experience. J. Phonetics 36, 141-157 (2008).
- PK Kuhl, KA Williams, F Lacerda, KN Stevens, B Lindblom, Linguistic experience alters 859 860 phonetic perception in infants by 6 months of age. Science 255, 606-608 (1992)
- 26. JE Flege, Second language speech learning: Theory, findings, and problems. Vol. 92, pp. 861 862 233-277 (1995).

CT Best, A direct realist view of cross-language speech perception. (York Press), pp. 171-27. 863 206 (1995) 864

865

866

867

868

869

870

871

875

876

877

882

883

884

885

886

887

888

889

890

891

895

899

900

901

902

903

904

913

914

915

916

917

918

919

920

921

922

923

924

925

926

930

932

933

934

935

936

937

- 28. T Schatz, F Bach, E Dupoux, Evaluating automatic speech recognition systems as quantitative models of cross-lingual phonetic category perception. The J. Acoust. Soc. Am. 143, EL372-EL378 (2018).
- 29 T Schatz, NH Feldman, Neural network vs. hmm speech recognition systems as models of human cross-linguistic phonetic perception in Proc. CCN. (2018).
- 30. J Gervain, J Mehler, Speech perception and language acquisition in the first year of life. Annu. review psychology 61, 191–218 (2010).
- 31 PK Kuhl, Innate predispositions and the effects of experience in speech perception; The na-872 tive language magnet theory in Developmental neurocognition: Speech and face processing 873 in the first year of life. (Springer), pp. 259-274 (1993). 874
- 32 N Kazanina, JS Bowers, W Idsardi, Phonemes: Lexical access and beyond. Psychon. bulletin & review 25, 560-585 (2018).
- 33 NS Trubetzkoy, Principles of phonology. (1969).
- A Cristia, Fine-grained variation in caregivers'/s/predicts their infants'/s/category. The J. 878 34. Acoust. Soc. Am. 129, 3271-3280 (2011). 879
- A Cristia, Can infants learn phonology in the lab? a meta-analytic answer. Cognition 170, 35. 880 312-327 (2018). 881
- 36 D Swingley, Contributions of infant word learning to language development. Philos. Transactions Royal Soc. B: Biol. Sci. 364, 3617-3632 (2009).
- DH Klatt, Speech perception: A model of acoustic-phonetic analysis and lexical access. pp. 37. 243-288 (1980)
- D Shankweiler, W Strange, R Verbrugge, Speech and the problem of perceptual constancy. 38 pp. 315-345 (1977).
- 39 I Appelbaum, The lack of invariance problem and the goal of speech perception in Proc. ICSLP. (1996).
- B De Boer, PK Kuhl, Investigating the role of infant-directed speech with a computer model. Acoust. Res. Lett. Online 4, 129-134 (2003).
- 41. MH Coen, Self-supervised acquisition of vowels in american english in Proc. AAAI. (2006). 892 GK Vallabha, JL McClelland, F Pons, JF Werker, S Amano, Unsupervised learning of vowel 893 894
- categories from infant-directed speech. Proc. Natl. Acad. Sci. 104, 13273-13278 (2007). 43. B Gauthier, R Shi, Y Xu, Learning phonetic categories by tracking movements. Cognition 103, 80-106 (2007).
- 896 B McMurray, RN Aslin, JC Toscano, Statistical learning of phonetic categories: insights from 897 a computational approach. Dev. science 12, 369-378 (2009). 898
- C Jones, F Meakins, S Muawiyath, Learning vowel categories from maternal speech in gurindji kriol. Lang. Learn. 62, 1052-1078 (2012).
- B Dillon, E Dunbar, W Idsardi, A single-stage approach to learning phonological categories: Insights from inuktitut. Cogn. Sci. 37, 344-377 (2013).
- F Adriaans, Effects of consonantal context on the learnability of vowel categories from infant-47. directed speech. The J. Acoust. Soc. Am. 144, EL20-EL25 (2018).
- H Rasilo, O Räsänen, UK Laine, Feedback and imitation by a caregiver guides a virtual 905 infant to learn native phonemes and the skill of speech inversion. Speech Commun. 55, 906 909-931 (2013). 907
- FH Guenther, MN Gjaja, The perceptual magnet effect as an emergent property of neural 49. 908 map formation. The J. Acoust. Soc. Am. 100, 1111-1121 (1996). 909
- PW Jusczyk, Developing Phonological Categories from the Speech Signal, (York, Timonium, 50 910 MD), pp. 17-64 (1992). 911 912
- 51 PW Jusczyk. From general to language-specific capacities: the WRAPSA model of how speech perception develops. J. Phonetics 21, 3-28 (1993).
- 52. P Jusczyk. The discovery of spoken language (1997)
- B Varadarajan, S Khudanpur, E Dupoux, Unsupervised learning of acoustic sub-word units 53. in Proc. ACL. (2008).
- AS Park, JR Glass, Unsupervised pattern discovery in speech. IEEE Transactions on Audio. 54 Speech, Lang. Process. 16, 186-197 (2008).
- 55. Cy Lee, J Glass, A nonparametric bayesian approach to acoustic model discovery in Proc. ACL. (2012).
- A Jansen. et al., A summary of the 2012 jhu clsp workshop on zero resource speech tech-56. nologies and models of early language acquisition in Proc. ICASSP. (IEEE), (2013)
- 57. G Synnaeve, T Schatz, E Dupoux, Phonetics embedding learning with side information in Proc. SLT), (2014). 58. M Versteegh, et al., The zero resource speech challenge 2015 in Proc. INTERSPEECH.
- (2015)
- 59. M Versteegh, X Anguera, A Jansen, E Dupoux, The zero resource speech challenge 2015: 927 Proposed approaches and results. Procedia Comput. Sci. 81, 67-72 (2016). 928
- 60 L Ondel, L Burget, J Černocký, Variational inference for acoustic unit discovery. Procedia 929 Comput. Sci. 81, 80-86 (2016) 931
- 61. H Chen, CC Leung, L Xie, B Ma, H Li, Parallel inference of dirichlet process gaussian mixture models for unsupervised acoustic modeling: A feasibility study in Proc. INTERSPEECH. (2015).
- 62. R Thiolliere, E Dunbar, G Synnaeve, M Versteegh, E Dupoux, A hybrid dynamic time warping-deep neural network architecture for unsupervised acoustic modeling in Proc. IN-TERSPEECH. (2015).
- H Kamper, M Elsner, A Jansen, S Goldwater, Unsupervised neural network based feature 63. extraction using weak top-down constraints in Proc. ICASSP. (2015).
- 64 D Renshaw, H Kamper, A Jansen, S Goldwater, A comparison of neural network methods 939 for unsupervised representation learning on the zero resource speech challenge in Proc. 940 INTERSPEECH. (2015). 941
- N Zeghidour, G Synnaeve, M Versteegh, E Dupoux, A deep scattering spectrum-deep 942 siamese network pipeline for unsupervised acoustic modeling in Proc. ICASSP. (2016). 943 944
- 66 M Heck, S Sakti, S Nakamura, Unsupervised linear discriminant analysis for supporting dpgmm clustering in the zero resource scenario. Procedia Comput. Sci. 81, 73-79 (2016). 945
- M Heck, S Sakti, S Nakamura, Feature optimized dpgmm clustering for unsupervised sub-946

word modeling: A contribution to zerospeech 2017 in Proc. ASRU. (2017). 947

951

952

954

963

964

965

966

967

968

969

970

971

972

973

974

975

976

977

978

979 980

981 982

983

984 985

986

987

988

989

990

991

992

993

994

996

997

998

999 1000

1001 1002

1003

1004

- WN Hsu, Y Zhang, J Glass, Unsupervised learning of disentangled and interpretable repre-948 68 sentations from sequential data in Proc. NEURIPS. (2017). 949
- 69. E Dunbar, et al., The zero resource speech challenge 2017 in Proc. ASRU. (2017). 950
  - 70. J Chorowski, RJ Weiss, S Bengio, Avd Oord, Unsupervised speech representation learning using wavenet autoencoders. CoRR abs/1901.08810 (2019).
- 953 71. E Dupoux, Cognitive science in the era of artificial intelligence: A roadmap for reverseengineering the infant language-learner. Cognition 173, 43-59 (2018).
- 955 72. P Mermelstein, Distance measures for speech recognition, psychological and instrumental. 956 Pattern recognition artificial intelligence 116, 91-103 (1976).
- 957 73. K Miyazawa, H Kikuchi, R Mazuka, Unsupervised learning of vowels from continuous 958 speech based on self-organized phoneme acquisition model in Proc. INTERSPEECH. 959 (2010)
- 960 74. K Miyazawa, H Miura, H Kikuchi, R Mazuka, The multi timescale phoneme acquisition model 961 of the self-organizing based on the dynamic features in Proc. INTERSPEECH. (2011).
- 962 JR Saffran, JF Werker, LA Werner, The infant's auditory world: Hearing, speech, and the 75. beginnings of language. (Wiley Online Library), (2006).
  - PK Kuhl, et al., Cross-language analysis of phonetic units in language addressed to infants. 76. Science 277, 684-686 (1997).
  - 77. A Fernald, Speech to infants as hyperspeech: Knowledge-driven processes in early word recognition. Phonetica 57, 242-254 (2000).
  - 78. B McMurray, KA Kovack-Lesh, D Goodwin, W McEchron, Infant directed speech and the development of speech perception: Enhancing development or an unintended consequence? Cognition 129, 362-378 (2013)
  - 79. A Cristia, A Seidl, The hyperarticulation hypothesis of infant-directed speech. J. Child Lang. 41, 913-934 (2014).
  - A Martin, et al., Mothers speak less clearly to infants than to adults: A comprehensive test of the hyperarticulation hypothesis. Psychol. science 26, 341-347 (2015).
  - 81. B Ludusan, A Seidl, E Dupoux, A Cristia, Motif discovery in infant-and adult-directed speech in Proc. CogACLL. (2015)
  - 82. BS Eaves Jr, NH Feldman, TL Griffiths, P Shafto, Infant-directed speech is consistent with teaching. Psychol. review 123, 758 (2016).
  - 83. A Guevara-Rukoz, et al., Are words easier to learn from infant-than adult-directed speech? a quantitative corpus-based investigation. Cogn. science 42, 1586-1617 (2018).
  - 84. DB Paul, JM Baker, The design for the wall street journal-based csr corpus in Proc. Workshop on Speech and Natural Language. (1992).
  - 85. T Schultz, Globalphone: a multilingual speech and text database developed at karlsruhe university. in Proc. INTERSPEECH. (2002).
  - 86. MA Pitt, K Johnson, E Hume, S Kiesling, W Raymond, The buckeye corpus of conversational speech: Labeling conventions and a test of transcriber reliability. Speech Commun. 45, 89-95 (2005).
  - 87. K Maekawa, Corpus of spontaneous japanese: Its design and evaluation in Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition. (2003).
  - 88. NA Macmillan, CD Creelman, Detection theory: A user's guide. (Psychology press), (2004).
  - NH Feldman, TL Griffiths, JL Morgan, The influence of categories on perception: Explaining 89. the perceptual magnet effect as optimal statistical inference. Psychol. review 116, 752 (2009).
- 90. T Schatz, et al., Evaluating speech features with the minimal-pair abx task; Analysis of the classical mfc/plp pipeline in Proc. INTERSPEECH. (2013). 995
  - 91 T Schatz, Ph.D. thesis (Université Paris 6) (2016).
  - DK Burnham, Developmental loss of speech perception: Exposure to and experience with 92. a first language. Appl. Psycholinguist. 7, 207-240 (1986).
  - 93. V Hazan, S Barrett, The development of phonemic categorization in children aged 6-12. J. phonetics 28, 377-396 (2000).
  - 94. K Idemaru, LL Holt, The developmental trajectory of children's perception and production of english/r/-/l. The J. Acoust. Soc. Am. 133, 4232-4246 (2013).
  - 95. H Hofmann, K Kafadar, H Wickham, Letter-value plots: Boxplots for large data (The American Statistician) (2011).
- T Tsushima, et al., Discrimination of english/rl/and/wy/by japanese infants at 6-12 months: 1005 96. 1006 language-specific developmental changes in speech perception abilities in Proc. ICSLP. 1007 (1994)
- 1008 97. PK Kuhl, FM Tsao, HM Liu, Foreign-language experience in infancy: Effects of short-term 1009 exposure and social interaction on phonetic learning. Proc. Natl. Acad. Sci. 100, 9096-9101 1010 (2003)
- 1011 98. T Teinonen, RN Aslin, P Alku, G Csibra, Visual speech contributes to phonetic learning in 1012 6-month-old infants. Cognition 108, 850-855 (2008).
- 1013 99. HH Yeung, JF Werker, Learning words' sounds before learning how words sound: 9-month-1014 olds use distinct objects as cues to categorize speech information. Cognition 113, 234-243 1015 (2009)
- NH Feldman, EB Myers, KS White, TL Griffiths, JL Morgan, Word-level information influ-1016 100. 1017 ences phonetic learning in adults and infants. Cognition 127, 427-438 (2013).
- 1018 101. N Mani, S Schneider, Speaker identity supports phonetic category learning. J. Exp. Psychol. Hum. Percept. Perform. 39, 623 (2013). 1019
- 102. HH Yeung, T Nazzi, Object labeling influences infant phonetic learning and generalization. 1020 Cognition 132, 151-163 (2014). 1021
- 1022 103. HH Yeung, LM Chen, JF Werker, Referential labeling can facilitate phonetic learning in 1023 infancy. Child development 85, 1036-1049 (2014).
- 104. C Bergmann, L Ten Bosch, P Fikkert, L Boves, A computational model to investigate as-1024 umptions in the headturn preference procedure. Front. psychology 4, 676 (2013). 1025
- 105. CA Thorburn, NH Feldman, T Schatz, A quantitative model of the language familiarity effect 1026 1027 in infancy in Proc. CCN. (2019).
- JL Schwartz, LJ Boë, N Vallée, C Abry, The dispersion-focalization theory of vowel systems. 1028 106 1029 J. phonetics 25, 255-286 (1997).
- 1030 107. SA Zahorian, AJ Jagharghi, Spectral-shape features versus formants as acoustic correlates

for yowels. The J. Acoust. Soc. Am. 94, 1966-1982 (1993). 1031 108 M Ito, J Tsuchida, M Yano, On the effectiveness of whole spectral shape for yowel perception. 1032 The J. Acoust. Soc. Am. 110, 1141-1149 (2001). 1033

1034

1038

1039

1043

1050

1051

1054

1055

1056

1057

- 109. MR Molis, Evaluating models of vowel perception, The J. Acoust. Soc. Am, 111, 2433-2434 (2005).
- 1035 JM Hillenbrand, RA Houde, RT Gayvert, Speech perception based on spectral peaks versus 110. 1036 spectral shape. The J. Acoust. Soc. Am. 119, 4041-4054 (2006). 1037
- Y Matusevych, T Schatz, H Kamper, NH Feldman, S Goldwater, Evaluating computational 111. models of infant phonetic learning across languages in Proc. COGSCI. (2020).
- 112 G Aversano, A Esposito, M Marinaro, A new text-independent method for phoneme segmen-1040 tation in Proc. MWSCAS. (2001). 1041 1042
- 113. O Rasanen, Basic cuts revisited: Temporal segmentation of speech into phone-like units with statistical learning at a pre-linguistic level in Proc. COGSCI. (2014).
- 114. P Michel, O Rasanen, R Thiolliere, E Dupoux, Blind phoneme segmentation with temporal 1044 prediction errors in Proc. ACL. (2017). 1045
- E Hermann, S Goldwater, Multilingual bottleneck features for subword modeling in zero-115. 1046 resource languages in Proc. Interspeech. (2018). 1047
- 116. CG Hempel, P Oppenheim, Studies in the logic of explanation. Philos. science 15, 135-175 1048 (1948)1049
- 117. M VanDam, et al., Homebank: An online repository of daylong child-centered audio recordings in Seminars in speech and language. (2016).
- MC Frank, et al., A collaborative approach to infant research: Promoting reproducibility, best 118. 1052 practices, and theory-building. Infancy 22, 421-435 (2017). 1053
- C Bergmann, et al., Promoting replicability in developmental research through metaanalyses: Insights from language acquisition research. Child development 89, 1996-2009 (2018).
- 120. J Sethuraman, A constructive definition of dirichlet priors. Stat. sinica 4, 639-650 (1994).
- J Chang, JW Fisher III, Parallel sampling of dp mixture models using sub-cluster splits in Proc. NEURIPS. (2013).
- 1059 122. TK Vintsyuk, Speech discrimination by dynamic programming. Cybern. Syst. Analysis 4, 1060 52-57 (1968). 1061



## <sup>2</sup> Supplementary Information for

- Early phonetic learning without phonetic categories
- <sup>4</sup> Insights from large-scale simulations on realistic input
- 5 Thomas Schatz, Naomi H. Feldman, Sharon Goldwater, Xuan-Nga Cao and Emmanuel Dupoux
- 6 Thomas Schatz.

1

7 E-mail: thomas.schatz.1986@gmail.com

#### 8 This PDF file includes:

- <sup>9</sup> Supplementary text
- <sup>10</sup> Figs. S1 to S11
- 11 Table S1
- 12 References for SI reference citations

#### **13** Supporting Information Text

#### 14 Supplementary Materials and Methods.

15 1. Datasets. The BUC and GPJ corpora annotations present a number of inconsistencies and were curated in-house. In particular,
 16 readers for the GPJ corpus often need several takes before they read an utterance correctly and the failed takes are included in
 17 the original corpus. We only keep the final take for each sentence. For the two spontaneous speech corpora, we keep disfluencies

the original corpus. We only keep the final take for each sentence. For the two spontaneous speech corpora, we keep distituencies typical of spontaneous speech (such as hesitations, word fragments, pronunciation errors, fillers, etc.), but remove parts that

were not phonetically transcribed or that include other kinds of noise or silence (96.11% and 80.38% of all utterances are kept for the BUC and CSJ corpora, respectively).

Phonetic transcriptions for the two read speech corpora are obtained by combining the read text with a phonetic dictionary. For the two spontaneous speech corpora, a manual phonetic transcription of the recordings is used. Word units, which are not directly apparent in the Japanese writing system, are obtained from the phonetic transcriptions by a Japanese morphological parser for the read Japanese corpus. For the spontaneous Japanese corpus, we use the provided 'Long Word Units' as words. We exclude phonemes occurring with frequency less than 1 in 10,000 by removing any utterance in which they occur and we harmonize the transcriptions in order to have the same phonemic inventory for the read and spontaneous corpora for each

27 language. No phonemes are excluded for the American English corpora. For the Japanese corpora, a few geminate consonants

are excluded (/b:/, /z:/, /h:/, /d:/, /z:/, / $\phi$ :/ for both corpora and /ts:/ for the GPJ corpus only). The retained phonemic

inventory for American English consists of 24 consonants (/p/, /t/, /k/, /b/, /d/, /g/, /f/, /v/, / $\theta$ /, / $\delta$ /, /s/, /z/, /J/, / $\sigma$ /, / $\sigma$ /

33 / g', (5), (11), (1

 $^{22}$  /d/, /g/, /s/, /c/, /s:/, /c/, /z/, /z/, /ts/, /ts/, /tc/, /tc/, /m/, /n/, /n/, /h/, / $\phi$ /, /r/, /w/, /j/) and 10 vowels (/ä/, /e/,

<sup>33</sup> /i/, /o/, /ui/, /äː/, /eː/, /iː/, /oː/, /uː/). For each corpus, timestamps are obtained for the phonetic transcriptions through

forced alignment with an automatic speech recognition (ASR) system (same architecture for the acoustic model as for the

<sup>35</sup> phoneme recognizer baseline described in Section 2 below, trained on the full corpus).

2. Phoneme recognizer baselines. As a baseline, we also train a phoneme recognizer on the training set of each corpus, with 36 explicit supervision (i.e. providing the phonemic transcriptions of the training stimuli along with the waveforms). Specifically, 37 38 we use the Kaldi toolkit (1) for automatic speech recognition (ASR) to train a hidden Markov model Gaussian mixture model (HMM-GMM) acoustic model and a phoneme-level bigram language model for each training set. The same training recipe 39 (adapted from the Wall Street Journal corpus recipe), with the same parameters is used to train a separate model on each of 40 the four corpora. The acoustic model takes the form of a probabilistic generative model with each phoneme modeled as a set of 41 contextual variants that are allowed to depend on word-position and preceding and following phonemes. Each variant is itself 42 modeled as a tri-state hidden Markov model with diagonal covariance Gaussian mixture emission probabilities. The models are 43 adapted to speakers both during training and test through feature-space maximum likelihood linear regression (fMLLR). See 44 the Kaldi toolkit documentation for more detail (http://kaldi-asr.org/doc/). 45

The trained acoustic and language models are combined (with kaldi acoustic scale parameter set to 0.1) to obtain representations of test stimuli (possibly in a 'foreign' language) under the form of a sequence of frame-level Viterbi-smoothed posterior probability vectors. We extract frame-level posterior probabilities at two granularity levels: actual phonemes—to which we refer as the *phoneme recognizer* baseline—and individual states of the contextual hidden Markov models—to which we refer as the *ASR phone state* baseline.

#### 51 3. Analysis of learned representations.

Correction for possible misalignment in the acoustic (in)variance test. We compensate for possible misalignment 52 of the central phones' central frames by allowing the dominant unit at the central frame to be replaced by any unit that was 53 dominant at some point within the previous or following 46ms, provided this brings down the overall count of distinct dominant 54 units for the ten occurrences. Finding the optimal way to assign dominant units under this constraint corresponds to solving an 55 56 instance of the NP-complete minimal hitting set size problem (2). We are able to solve the problem exactly in most cases, due to the small size of the considered instances. In the few cases where we are not able to solve the problem exactly, our solver 57 provides a lower bound on the number of representations and we use a greedy search to obtain an upper bound. Although the 58 effect on the results is very small, we report lower bounds for the Gaussian mixture models and upper bounds for the phoneme 59 recognizer and ASR phone state baselines, in order to be maximally conservative. 60

Stimulus selection for the acoustic (in)variance test. To avoid potentially mispronounced short function words and possible co-articulation effect across word boundaries, for the acoustic (in)variance test, we select only words of at least five phonemes and study their central phoneme(s).\* We sample uniformly at random a subset of ten occurrences (by a single speaker or by at least ten distinct speakers, depending on the condition) for each such word with enough repetitions in the test set. We report results averaged over ten independent runs of this stimulus sampling procedure. The results are also averaged over the two possible 'central phone' positions for words of even length and—in the within-speaker condition—over all available speakers for a given word type. This yields one average number of distinct dominant units per tested word type. The number

<sup>\*</sup>This stimulus selection procedure was only applied for the acoustic (in)variance test and has the effect of making the test more conservative—i.e. the learned representations would look even more variable without this restriction. Other analyses were not restricted to such words, and all model training was carried out with unfiltered continuous speech that contained words of all different lengths in unsegmented whole sentences.

of available word types matching the specified conditions is 13 (within speaker) and 476 (across speaker) for the American English test stimuli and 83 (within speaker) and 408 (across speaker) for the Japanese test stimuli. As an example, here are the word types selected for the within-speaker American English condition: unquote, billion, dollars, hundred, company, market, million, mister, nineteen, percent, seven, seventy, thousand. For the within-speaker condition, we additionally listened to each test stimulus to identify potential mispronounced, noisy or misaligned stimuli and we checked that excluding these stimuli from the analysis (0/83 word types, 4/1048 word tokens excluded for American English; 14/168 word types, 204/2217 word tokens excluded for Japanese) did not affect the overall pattern of results (Figure S8).

4. Deriving systematic model predictions. We systematically seek phonetic contrasts of American English and of Japanese for 75 which the learning mechanism under study robustly predicts a significant cross-linguistic difference in discrimination between 76 Japanese- and American English-learning infants. By robust we mean that (a) a significant difference in discrimination errors 77 between models trained on American English and Japanese is consistently found across possible choices for the training and 78 test registers, and (b) that the magnitude of this difference does not decrease when the amount of training input is increased. 79 The former criterion allows us to rule out effects that would reflect peculiarities of the training and/or test stimuli rather than 80 an intrinsic property of the language pair under study. The latter criterion allows us to rule out transient effects that might 81 reflect peculiarities of the model initialization and/or be unlikely to be observed empirically. 82

We define the predicted cross-linguistic effect for a phonetic contrast as the expected difference in average ABX discrimination 83 error between an 'American English-native' and a 'Japanese-native' model on that contrast, where the expectation is taken 84 over the choice of American English model, Japanese model, test speaker, phonetic context, and choice of the a, b, and x85 acoustic tokens given the contrast, speaker and phonetic context. For each contrast, we perform statistical significance tests 86 separately for each of the 8 possible combinations of training register for the American English model, training register for the 87 Japanese model, and test register. We use the models trained on the  $1/10^{th}$  training sets of each corpus for these significance 88 tests, which allows us to take into account variance due to the model training procedure (including the choice of input data) in 89 addition to that due to the choice of test stimuli. We estimate the predicted cross-linguistic effect and its variance and use 90 those estimates to conduct asymptotic bilateral z-tests of the hypothesis that the cross-linguistic effect is different from 0. We 91 also estimate the effects (but not the variances) using the full training sets, which allows us to test whether the observed effects 92 increase (in absolute value) with the amount of input data. We report a robust predicted cross-linguistic effect for a contrast if 93 each of the estimated effects for that contrast (for each of the 8 possible combination of training and test registers) is in the 94 95 same direction and significantly different from 0 in our asymptotic bilateral z-test, with Benjamini-Yekutieli (3) correction for 96 multiple correlated comparisons at level  $\alpha = 0.05$ ; and if the estimated effect for models trained on the full training sets are in the same direction and larger in absolute value than the corresponding effects estimated for models trained on the  $1/10^{th}$ 97 subsets. 98

In what follows, we first formally define the predicted cross-linguistic effect for a phonetic contrast  $P_1, P_2$ . We then discuss how to estimate the effect in practice from finite samples of models trained on Japanese and trained on American English, and finite samples of test acoustic tokens from phonetic categories  $P_1$  and  $P_2$ . Finally, we explain in detail how the statistical significance of the estimated effects can be assessed.

Effect of interest. We are interested in the predicted cross-linguistic effect for a phonetic contrast  $P_1, P_2$ , i.e. the expected difference in average ABX discrimination error between a model trained on language  $L_1$  and a model trained on language  $L_2$ , which we denote as  $\delta(P_1, P_2, L_1, L_2)$  and define formally below.<sup>†</sup> Let us consider a model M trained on input language L, input register  $R_I$  and input amount  $A_I$ , and tested on phonetic category P from test language  $L_T$  in phonetic context C(preceding and following phonetic category) from test speaker S with test register  $R_T$ . Let us note

$$p_{P,L,R_I,A_I,L_T,R_T}(\mathfrak{R} \mid M, S, C),$$

the probability distribution over model representations  $\Re$ , where we treat the trained model M, test speaker S and test context C as conditioning random variables and assume fixed values for the other parameters. Then, the predicted cross-linguistic effect for phonetic contrast  $P_1, P_2$  and training languages  $L_1, L_2$  is defined as

$$\delta(P_1, P_2, L_1, L_2) := \mathbf{E}_{M_1, M_2, S, C}[\epsilon(P_1, P_2, M_1, S, C) - \epsilon(P_1, P_2, M_2, S, C)],$$

109 where

•  $M_x$  for x in  $\{1, 2\}$  is a randomly sampled trained model for input language  $L_x$ , training register  $R_{I,x}$  and input amount  $A_{I,x}$ ;

- S is a randomly chosen test speaker and C is a context chosen uniformly at random among available test phonetic contexts, for test language  $L_T$ , test register  $R_T$  and test phonetic contrast $(P_1, P_2)$ ;
  - $\epsilon(P_1, P_2, M_x, S, C)$  is the symmetric ABX discrimination error, defined as

$$\epsilon(P_1, P_2, M_x, S, C) := \frac{1}{2} [e(P_1, P_2, M_x, S, C) + e(P_2, P_1, M_x, S, C)],$$

<sup>&</sup>lt;sup>†</sup>This is for a given choice of input registers  $R_{I,1}$  and  $R_{I,2}$  and input amounts  $A_{I,1}$  and  $A_{I,2}$  for each model, and of test language  $L_T$  and test register  $R_T$  (which we constrain to be the same for the two tested phonetic categories in our experiments). To avoid clutter, we do not indicate these dependencies explicitly in the notation.

with

114

$$e(P_1, P_2, M_x, S, C) := p[d(A, X) < d(B, X)] + \frac{1}{2}p[d(A, X) = d(B, X)],$$

for A, X drawn independently from  $p_{P_1,L}(\mathfrak{R} \mid M_x, S, C)$  and B drawn from  $p_{P_2,L}(\mathfrak{R} \mid M_x, S, C)$ .

<sup>115</sup> This is the quantity we seek to estimate, given our trained models in English and Japanese, and the particular acoustic tokens <sup>116</sup> in our corpora from the phonetic categories we would like to test.

**Estimation of the effect.** In order to obtain a sample of model representations  $S_{P,M,L_T,R_T,S,C}$  for each relevant combination of the index variables, we extract a representation of each test acoustic token for each model M.<sup>‡</sup> For each combination of test language  $L_T$ , test register  $R_T$ , test speaker S and test phonetic context C, we obtain a sample of up to 5 acoustic realizations of each phonetic category from the test corpus. For each combination of training language L, training register  $R_I$ , we obtain ne model trained on the full training set and 10 models that are each trained on  $1/10^{th}$  of it.

Given these samples from the distributions of model representations of test stimuli, we define the following estimator of  $\delta(P_1, P_2, L_1, L_2)$ ,

$$\hat{\delta}(\mathcal{S}, \mathcal{M}_{1}, \mathcal{M}_{2}) := \frac{1}{|\mathcal{S}|} \sum_{S \in \mathcal{S}} \frac{1}{|\mathcal{C}(S)|} \sum_{C \in \mathcal{C}(S)} \left( \frac{1}{|\mathcal{M}_{1}|} \sum_{M_{1} \in \mathcal{M}_{1}} \hat{\epsilon}(S_{P_{1}, M_{1}, S, C}, S_{P_{2}, M_{1}, S, C}) - \frac{1}{|\mathcal{M}_{2}|} \sum_{M_{2} \in \mathcal{M}_{2}} \hat{\epsilon}(S_{P_{1}, M_{2}, S, C}, S_{P_{2}, M_{2}, S, C}) \right),$$

where S is the set of sampled test speakers, C(S) is the set of contexts available for the target contrast from test speaker S,  $\mathcal{M}_1$  and  $\mathcal{M}_2$  are the sampled models for training language  $L_1$  and  $L_2$  respectively and  $\hat{\epsilon}$  is the estimator for the ABX discrimination error defined in the Material and Methods section of the main text.

Provided there is no systematic bias in how phonetic contexts are missing from the sample of any particular test speaker,  $\hat{\delta}(S, \mathcal{M}_1, \mathcal{M}_2)$  can be shown to be an unbiased estimator of  $\delta(P_1, P_2, L_1, L_2)$ .

Significance testing. We want to assess the contrasts for which a significant cross-linguistic difference in discriminability is observed. In order to do assess significance, we need a test statistic with a known distribution. For given  $P_1$ ,  $P_2$ ,  $L_1$ ,  $L_2$ , we define

135

$$\hat{D}(S, M_1, M_2) := \frac{1}{|\mathcal{C}(S)|} \sum_{C \in \mathcal{C}(S)} \left[ \hat{\epsilon}(S_{P_1, M_1, S, C}, S_{P_2, M_1, S, C}) - \hat{\epsilon}(S_{P_1, M_2, S, C}, S_{P_2, M_2, S, C}) \right]$$

134 It is straightforward to check that

j

$$\hat{\delta}(\mathcal{S}, \mathcal{M}_1, \mathcal{M}_2) = \frac{1}{|\mathcal{S}||\mathcal{M}_1||\mathcal{M}_2|} \sum_{\substack{S \in \mathcal{S} \\ M_1 \in \mathcal{M}_1 \\ M_2 \in \mathcal{M}_2}} \hat{D}(S, M_1, M_2).$$

 $\hat{\delta}(S, \mathcal{M}_1, \mathcal{M}_2)$  can thus be interpreted as a (generalized) U-statistic with kernel  $\hat{D}$  of order 3 and degree (1, 1, 1) (4), applied to mutually independent i.i.d. samples  $S, \mathcal{M}_1$  and  $\mathcal{M}_2$  (where an element S of S is effectively a sample of up to five acoustic tokens for each phonetic context available from speaker S for the target phonetic contrast).

Assuming this U-statistic is not degenerate, we can apply the central limit theorem for U-statistics (4) to obtain that

140
$$\frac{\delta(\mathcal{S},\mathcal{M}_1,\mathcal{M}_2)}{\mathbf{Var}[\hat{\delta}(\mathcal{S},\mathcal{M}_1,\mathcal{M}_2)]}$$

has an asymptotic normal distribution with mean  $\delta(P_1, P_2, L_1, L_2)$  and variance 1. Provided we can estimate the variance of the estimator  $\operatorname{Var}[\hat{\delta}(\mathcal{S}, \mathcal{M}_1, \mathcal{M}_2)]$ , this result allows us to perform asymptotic z-tests of  $\mathcal{H}_0$ :  $\delta(P_1, P_2, L_1, L_2) = 0$  versus  $\mathcal{H}_1$ :  $\delta(P_1, P_2, L_1, L_2) \neq 0$ . We provide the required estimator  $\hat{V}(\mathcal{S}, \mathcal{M}_1, \mathcal{M}_2)$  of  $\operatorname{Var}[\hat{\delta}(\mathcal{S}, \mathcal{M}_1, \mathcal{M}_2)]$  in the next section.

<sup>144</sup> Estimation of the variance of  $\hat{\delta}$ . The previous section showed that given an estimate  $\hat{V}(S, \mathcal{M}_1, \mathcal{M}_2)$  of the variance <sup>145</sup>  $\operatorname{Var}[\hat{\delta}(S, \mathcal{M}_1, \mathcal{M}_2)]$ , we can compute statistical significance of the estimated differences in discrimination error between <sup>146</sup> languages. In this section we derive such an estimator.

We first find an expression for  $\operatorname{Var}[\hat{\delta}(\mathcal{S}, \mathcal{M}_1, \mathcal{M}_2)]$ , then derive an estimator from it. We use  $n_1$  to denote the number of test speakers,  $|\mathcal{S}|$ ,  $n_2$  to denote the number of models trained on language  $L_1$ ,  $|\mathcal{M}_1|$ , and  $n_3$  to denote the number of models trained on language  $L_2$ ,  $|\mathcal{M}_2|$ . We can express the variance using the standard decomposition for the variance of a U statistic (4),

$$\mathbf{Var}[\hat{\delta}(\mathcal{S},\mathcal{M}_{1},\mathcal{M}_{2})] = \frac{1}{n_{1}n_{2}n_{3}}[(n_{1}-1)(n_{2}-1)\sigma_{001}^{2} + (n_{1}-1)(n_{3}-1)\sigma_{010}^{2} + (n_{2}-1)(n_{3}-1)\sigma_{100}^{2} + (n_{1}-1)\sigma_{011}^{2} + (n_{2}-1)\sigma_{101}^{2} + (n_{3}-1)\sigma_{110}^{2} + \sigma_{111}^{2}]$$

where  $\sigma_{xyz}^2$  denotes the covariance between  $\hat{D}(s_1, a_1, j_1)$  and  $\hat{D}(s_2, a_2, j_2)$  for two triplets  $(s_1, a_1, j_1), (s_2, a_2, j_2)$  formed of a randomly sampled combination of a test speaker, an American English model, and a Japanese model, with the subscripts x, y,

<sup>&</sup>lt;sup>‡</sup> Possibly with some missing data, as not all possible phonetic contexts occur for each speaker and each phonetic category in any given test set.

and z indicating whether the two test speakers, American English models and Japanese models, respectively, are constrained to be identical (subscript 0) or not (subscript 1). For example,

$$\sigma_{000}^{2} = \mathbb{E}_{s_{1},s_{2},a_{1},a_{2},j_{1},j_{2}} [\hat{D}(s_{1},a_{1},j_{1})\hat{D}(s_{2},a_{2},j_{2})] - (\mathbb{E}_{s,a,j}[\hat{D}(s,a,j])^{2} = 0;$$
  

$$\sigma_{111}^{2} = \mathbb{E}_{s,a,j} [\hat{D}(s,a,j)^{2}] - (\mathbb{E}_{s,a,j}[\hat{D}(s,a,j])^{2};$$
  

$$\sigma_{001}^{2} = \mathbb{E}_{s_{1},s_{2},a_{1},a_{2},j} [\hat{D}(s_{1},a_{1},j)\hat{D}(s_{2},a_{2},j)] - (\mathbb{E}_{s,a,j}[\hat{D}(s,a,j])^{2}.$$

We now use the above variance decomposition to derive an estimator. Let us define the order 3, degree (2, 2, 2) kernel  $\psi_{k_1k_2k_3}$  for some strictly positive integers  $k_1, k_2, k_3$ , as follows

$$\begin{split} \psi_{k_1k_2k_3}(s_1,s_2,a_1,a_2,j_1,j_2) &\coloneqq \frac{1}{k_1k_2k_3} [(k_1-1)(k_2-1)(\hat{D}(s_1,a_1,j_1)\hat{D}(s_2,a_2,j_1) - \hat{D}(s_1,a_1,j_1)\hat{D}(s_2,a_2,j_2)) \\ &\quad + (k_1-1)(k_3-1)(\hat{D}(s_1,a_1,j_1)\hat{D}(s_2,a_1,j_2) - \hat{D}(s_1,a_1,j_1)\hat{D}(s_2,a_2,j_2)) \\ &\quad + (k_2-1)(k_3-1)(\hat{D}(s_1,a_1,j_1)\hat{D}(s_1,a_2,j_2) - \hat{D}(s_1,a_1,j_1)\hat{D}(s_2,a_2,j_2)) \\ &\quad + (k_1-1)(\hat{D}(s_1,a_1,j_1)\hat{D}(s_2,a_1,j_1) - \hat{D}(s_1,a_1,j_1)\hat{D}(s_2,a_2,j_2)) \\ &\quad + (k_2-1)(\hat{D}(s_1,a_1,j_1)\hat{D}(s_1,a_2,j_1) - \hat{D}(s_1,a_1,j_1)\hat{D}(s_2,a_2,j_2)) \\ &\quad + (k_3-1)(\hat{D}(s_1,a_1,j_1)\hat{D}(s_1,a_1,j_2) - \hat{D}(s_1,a_1,j_1)\hat{D}(s_2,a_2,j_2)) \\ &\quad + (\hat{D}(s_1,a_1,j_1)\hat{D}(s_1,a_1,j_1) - \hat{D}(s_1,a_1,j_1)\hat{D}(s_2,a_2,j_2)) \\ \end{split}$$

Let us consider some arbitrary orderings  $(s_1, s_2, ..., s_{n_1})$ ,  $(a_1, a_2, ..., a_{n_2})$  and  $(j_1, j_2, ..., j_{n_3})$  of  $\mathcal{S}$ ,  $\mathcal{M}_1$ , and  $\mathcal{M}_2$ , respectively. Let us also note  $(n \ k)$ , for any integers n and k, the set of all integer k-tuples  $(i_1, i_2, ..., i_k)$  such that  $1 \le i_1 < i_2 < ... < i_k \le n$ .

It is straightforward to show that  $\psi_{n_1n_2n_3}$  is an unbiased estimator for  $\operatorname{Var}[\hat{\delta}(\mathcal{S}, \mathcal{M}_1, \mathcal{M}_2)]$ , leading to the following symmetric unbiased estimator based on all of the available data

$$\hat{V}(\mathcal{S},\mathcal{M}_1,\mathcal{M}_2) := \frac{1}{\binom{n_1}{2}\binom{n_2}{2}\binom{n_3}{2}} \sum_{\substack{i_1,i_2 \in (n_1 \ 2) \\ j_1,j_2 \in (n_2 \ 2) \\ k_1,k_2 \in (n_3 \ 2)}} \psi^S_{n_1n_2n_3}(s_{i_1},s_{i_2},a_{j_1},a_{j_2},j_{k_1},j_{k_2}),$$

where  $\psi_{n_1n_2n_3}^S$  is the symmetrized version of  $\psi_{n_1n_2n_3}$ 

$$\psi_{n_1n_2n_3}^S(s_1, s_2, a_1, a_2, j_1, j_2) := \frac{1}{(2!)^3} \sum_{\substack{i_1, i_2 \in S_2\\j_1, j_2 \in S_2\\k_1, k_2 \in S_2}} \psi_{n_1n_2n_3}(s_{i_1}, s_{i_2}, a_{j_1}, a_{j_2}, j_{k_1}, j_{k_2}),$$

with  $S_2 = \{(1,2), (2,1)\}$  the set of all permutations of  $\{1,2\}$ .

With this estimator for the variance of  $\delta(S, \mathcal{M}_1, \mathcal{M}_2)$ , we can now conduct a z-test over the test statistic defined in the previous section to compute statistical significance of cross-linguistic discrimination differences.

#### 152 Supplementary Discussion.

1. Input idealization in computational modeling of early phonetic learning. Modeling studies investigating the feasibility of potential 153 learning mechanisms for early phonetic learning have typically relied on input idealizations that sidestep the lack of invariance 154 problem and the phonetic segmentation problem, and cannot therefore alleviate the feasibility concerns related to these 155 156 problems. In initial modeling work investigating the feasibility of learning phonetic categories through distributional learning 157 (5-9), the phonetic category segmentation problem was either simply assumed to have been solved (7-9), or the input speech was assumed to consist of exemplars from a restricted number of pre-segmented or isolated syllable types, that were furthermore 158 chosen such that automatic segmentation of the vowel nucleus based on voicing cues would be easy (5, 6). The impact of the 159 lack of invariance problem was minimized by artificially limiting the variability of the input. Specifically, the input speech 160 signal was: chosen from a restricted set of phonemes (5-9); occurring in a restricted set of phonetic contexts (5-7); uttered 161 by a (very) restricted set of speakers (5, 9); available to the learner in a manually encoded (7-9) and/or restricted (5-9)162 phonetic feature space; drawn from synthetic parametric sound distributions fitted to corpus data rather than using corpus data 163 164 directly (7, 8). Subsequent studies considered slightly more realistic variability and found that distributional learning was not sufficient anymore to learn phonetic categories accurately (10-16) and proposed additional learning mechanisms tapping into 165 other sources of information plausibly available to infants to complement distributional learning. However, demonstrations of 166 feasibility for the proposed mechanisms still assumed the phonetic category segmentation problem to be solved (10-12, 14-16)167 and/or did not fully address the lack of invariance problem by not considering the full variability of natural speech (10-16). 168 Specifically, input speech signal was: chosen from a restricted set of phonemes (10-12, 14-16); occurring in a restricted set of 169 phonetic contexts (12, 14, 16); uttered by a very restricted set of speakers (10, 11, 13, 15, 16); available to the learner in a 170 manually encoded (9, 10, 12, 14–16) and/or restricted (10–12, 14–16) phonetic feature space; drawn from synthetic parametric 171 sound distributions fitted to corpus data rather than using corpus data directly (11-14). Existing attempts to extend some 172

<sup>173</sup> of these results to more realistic learning conditions have failed (17, 18). The few studies that attempted to model infant

phonetic learning from naturalistic, unsegmented speech input remained inconclusive for lack of a suitable evaluation method (19, 20). Finally, we know of only one demonstration of feasibility for an account of early phonetic learning in which the outcome of learning is not phonetic categories (21). It also assumes the phonetic category segmentation problem to be solved and minimizes the impact of the lack of invariance problem by artificially limiting the variability of the input speech.

178 Modeling assumptions are necessary in any model—for example, our approach ignores the visual component of speech and uses adult-directed rather than child-directed speech—but they should be critically examined to assess their suitability relative 179 to the research objectives. For example, whereas the assumptions typically made in previous studies were all geared toward 180 making the learning problem easier—by sidestepping the lack of invariance and phonetic segmentation problems—we focus, as 181 much as possible, on modeling assumptions that make it harder. This means that in our framework, positive feasibility results 182 constitute much stronger evidence. Our framework is not devoid of modeling assumptions that make the learning problem 183 easier; for example, we consider speech input consisting of speech from a single speaker at a time, captured by a close-range 184 microphone, and with no overlap with environmental sounds. However, we make many fewer such simplifying assumptions than 185 previous models and we are careful not to sidestep the phonetic category segmentation and the lack of invariance problems in 186 particular. This ensures that our simulations are suitable to address feasibility concerns related to these problems. 187

2. Model initialization, learning procedure and convergence. Following Chen et al. (22), the parameters of our Gaussian mixture models are learned through the exact Markov chain Monte-Carlo (MCMC) sampling algorithm proposed in Chang & Fisher (23). This algorithm combines, in a principled way, Gibbs sampling of the parameters of instantiated mixture components (i.e. the clusters with non-empty membership at any given point in the algorithm execution) with sampling of split and merge moves that increase or reduce the number of instantiated mixture components. It is designed to combine good statistical convergence properties with computational efficiency, and in particular to allow the parallelization of the computations to accommodate large training datasets.

We also follow Chen et al. (22) for model initialization. They used the default initialization procedure in the implementation proposed by Chang & Fisher (23), which consists of assigning each data point in the training set uniformly at random to one of ten initial clusters. The mean vector and covariance matrix for each of these ten initial clusters is then taken as the mean and covariance of the points assigned to that cluster. The weights of each of the cluster in the initial mixture is obtained by drawing from a Dirichlet distribution with ten categories and concentration parameter whose *i*-th component, for  $1 \le i \le 10$ , is the number of points that were initially assigned to the *i*-th cluster.

In theory, the initial state should not influence the learning outcomes when using this algorithm. The sampling algorithm 201 we use comes with the usual guarantees (for sampling algorithms) of global convergence to the true posterior in the limit (23), 202 so that in principle, the initialization procedure should not matter if we run the sampling procedure for long enough. The main 203 issue in practice is that there is usually no definitive way to determine when it has been 'long enough'. In our case, we look at 204 the number of learned categories as a function of the number of sampling iterations (Figure S11). We see that this number 205 is largely stabilized after about 600 iterations for all the models we train. This suggests that training the models for 1500 206 sampling iterations (per parameter), as we do—again following the example of (22)—is sufficient for model convergence. We 207 also see that cross-linguistic differences emerge quite robustly on independent runs for models trained on one to two hours of 208 speech input (Figure 3(b)). Thus, we are reasonably confident that the models have converged. 209

Still, we cannot completely rule out the possibility that running the algorithm for longer might ultimately lead to a different 210 outcome (e.g. to units corresponding to phonetic categories), and that a different setting of the initial state might lead to that 211 outcome faster. This leads us to consider the biological and psychological plausibility of the initialization procedure we used. 212 A prominent proposal in the literature (see 24, for example)—motivated by observations of a certain 'language-readiness' of 213 the human brain at birth and even before (25)—is that infants start with an innately specified, 'universal' mapping from an 214 auditory space to a phonetic space, which is then progressively altered as they gain experience with their native language. 215 However, there have not yet been proposals for a concrete implementation of such a mapping (although see 26, for a possible 216 technical solution). 217

This view is not universally shared. An alternative hypothesis has been argued to be fully compatible with the empirical record (e.g. 27, 28), according to which the observation of 'universal' phonetic discrimination abilities in newborns would correspond to an initial mode of perception of a purely auditory nature, in the absence of any mapping to phonetic space. Under this view, phonetic representations would be initiated through some form of random mapping, and subsequently refined through experience-dependent plasticity. One benefit of this latter view is that it assumes less in terms of what needs to be genetically specified than an innate universal mapping between acoustic and phonetic space.

As discussed in the main text, MFCC input features can be interpreted as the output of a (very) simple model of the peripheral auditory system, and our approach to initialization can thus be understood as an implementation of this latter view. We are not aware of many empirical constraints on what would constitute a plausible random initialization of the phonetic clusters within this auditory space, and our initialization procedure represents one possible, albeit admittedly arbitrary, solution.

229 3. Interpretation of simulated discrimination errors and relation to empirical observations. To evaluate our trained models, we expose them to appropriate test stimuli (e.g. exemplars of [1] and [1]) and simulate discrimination tasks using the models' representation of these stimuli. Here, we discuss our criteria to decide if the models successfully account for early phonetic learning on the basis of the resulting discrimination errors. For the purpose of this article, we deem our models successful if they can account for the

cross-linguistic differences in discrimination abilities observed in infants in the first year of life for the Japanese/American
 English language pair we study.

The results to be accounted for come from a 2006 study by Kuhl and colleagues (29), since we are not aware of other studies 235 directly comparing the phonetic discrimination abilities of Japanese and American English infants in the first year. Using 236 237 a conditioned head turning paradigm, they found no significant difference between American English and Japanese infants 238 ability to discriminate a synthetic [Ia] stimulus from a synthetic [Ia] stimulus at 6-8 months. Both groups answered correctly on about 65% of test trials. In contrast, at 10-12 months, American English infants were found to to be significantly more accurate 239 than Japanese infants in the same task. American English infants answered correctly on about 75% of trials while Japanese 240 infants answered correctly on about 60% of trials. All four groups discriminated the stimuli significantly above chance. When 241 comparing across ages, American English 10-12 month olds were found to be significantly better at discriminating the stimuli 242 than their 6-8 month old counterparts, whereas Japanese 10-12 month olds were not found to be significantly worse than their 243 6-8 month old counterparts (but see 30). We adopt the standard interpretation that these results reflect infants' discrimination 244 of the [1-1] contrast, and not just of the two specific stimuli tested in the experiment. We therefore test our models both on 245 those specific stimuli (Figure S6), and on other instances of [1] and [1] (Figure 3). However, we do not assume these observations 246 of early phonetic learning in infants to mean that 10-12 month old infants have formed adult-like representations; while this is 247 a common view in the literature, it is premised on the phonetic category hypothesis we are contesting. In particular, we do not 248 take the results from Kuhl et al. (29) to necessarily indicate that Japanese 10-12 month olds have become nearly deaf to the 249 [1]-[1] distinction, or that American English 10-12 month olds learned to discriminate it perfectly.<sup>§</sup> 250

Given our current state of knowledge about infant cognition, there are some quantitative aspects of these results that 251 we cannot hope to model, even in principle. First, we cannot hope to model the quantitative values of the error rates or d' 252 measurements characterizing infant discrimination in these experiments, as these values depend strongly on the specifics of 253 the experiments in ways that are not well understood (33). This uncertainty might potentially be accounted for through free 254 parameters in the model, but fitting those parameters would not be feasible due to the limited number of datapoints available 255 to constrain them.<sup>¶</sup> Second, we do not know the precise correspondence between an infant of a particular age and a model 256 presented with a particular amount and quality of data. The quality and quantity of data in infants' environments does not 257 directly translate into their intake (34), the data they use for learning. In addition, some of the differences in infants' behavior 258 at different ages might also stem from developmental factors not directly related to perception, and these are not included in 259 our model. Moreover, we do not know whether infants rely solely on learned representations for discrimination, even when those 260 representations are just starting to be formed and might be unreliable, or whether they initially rely on language-universal 261 input features for discrimination, and then smoothly transition to relying on the learned language-specific representations as 262 the amount of training data increases. This prevents us from interpreting the change in discrimination errors as a function of 263 the amount of training input given to the model on Figure 3(b) directly as a developmental trajectory for example. 264

Because we cannot hope to get a quantitative match in either the absolute discrimination scores or the absolute quantity of 265 training data, we focus on modeling qualitative aspects of the empirical results. This means showing that American English 266 models discriminate [1] and [1] better than Japanese models do. We find this qualitative effect both with the original stimuli 267 from Kuhl et al. (29), and with a broader set of speech stimuli drawn from American English speech corpora. Figure S6 shows 268 that with small amounts of training data, the dissimilarity between the two original stimuli is roughly similar for all models. 269 As the amount of training data increases, the two stimuli become more dissimilar for the American English models, while 270 their dissimilarity stays roughly the same for the Japanese models. When tested on a broader set of [1] and [1] stimuli, all 271 models get better at discriminating this contrast as the amount of training data increases, but a clear cross-linguistic difference 272 nevertheless emerges (Figure 3(b)). As noted above, there are a number of reasons why the direction of change in absolute 273 error rates might not be reliable; but in both simulations, the increasing separation between English and Japanese models with 274 increasing training data qualitatively matches the empirical pattern. 275

A limitation of this study is that it focuses on one language pair, limiting the relevant empirical record to mostly one study 276 (29). Mugitani and colleagues (35) suggested that vowel length perception at 10 months could be similar in American English 277 and Japanese listeners; our models appear broadly consistent with this hypothesis, as we find no systematic difference in 278 Japanese vowel length discrimination between the Japanese and American English models (see Supplementary Discussion 5). 279 However, we do not focus on this result, as Mugitani and colleagues (35) did not directly test American English 10 month olds, 280 and recent evidence suggests that the development of vowel length perception, for Japanese listeners at least, might be more 281 282 complicated than once thought (36). As argued in the main discussion, in the longer term our modeling framework will allow evaluating the proposed learning mechanism against the empirical record on further language pairs, comparing it with other 283 possible learning mechanisms, and designing empirical tests of their predictions. 284

We are not aiming to model adult data, nor are we able to interpret absolute error rates relative to infant data. Thus, the absolute levels of the discrimination errors we obtain have little bearing on our main conclusions. However, it is still interesting to get a sense of how those absolute error rates might be interpreted. To this end, we added a supervised phoneme recognizer baseline as a possible approximation of an adult-like state, In general, the supervised baselines show larger cross-linguistic differences than our (unsupervised) models do. For the [x]-[l] contrast, for example, the absolute difference in discrimination errors between 'native' and 'non-native' models is about four times as large for the supervised phoneme recognizers as for the

<sup>§</sup> This view is supported by empirical evidence that American English infants' perception of [J]-[I] develops well beyond the first year of life (31). See also Feldman et al. (32).

<sup>&</sup>lt;sup>¶</sup>One potential solution might be to pool infant data across many experiments to try and calibrate task models. However, it is unclear whether this strategy could be successful, because of the heterogeneity in the way infant experiments are carried out in practice.

This is different from its role in Figures 4, S7, S9 and S10, where it is used as a possible embodiment of the linguistic notion of phonetic category.

<sup>291</sup> unsupervised models. These larger crosslinguistic differences are driven by decreased performance of the supervised baselines

on the 'non-native' language and increased performance on the 'native' language (Figures S3, S5), though improvement on

the 'native' language does not appear robust to a register change (Figure S3). These results show that the proposed learning

mechanisms for early phonetic learning is compatible with the view that one-year-olds have not yet formed mature, adult-like speech representations (32).\*\*

We additionally included an unlearned 'auditory' input features baseline (with distances computed directly between sequence 296 of MFCC input vectors) in Figures S3, S5, as a possible approximation of discrimination on the basis of a language-universal 297 auditory representation. This baseline performs surprisingly well relative to both the supervised baseline and the unsupervised 298 models in discriminating some phonetic contrasts. On average, the 'native' models do better than the baseline, and the 299 'non-native' models do worse, as expected (Figure S3). However, this is not true for every contrast, as can be seen for [1]-[1] 300 and [w]-[j] on Figure S5. There are a number of possible ways to interpret this result.<sup>††</sup> This might reflect a shortcoming 301 common to both the unsupervised models and supervised baselines for these contrasts. It might also be that, in order to 302 catch up with the input features baseline, our models require larger amount of training input (Figure 3(b)) or input that is 303 more similar to what infants hear (39). Finally, another possibility is that high level language-specific representation might 304 need to be combined with information-rich auditory representation (40) to enable accurate phonetic discrimination of certain 305 contrasts—as appears to be the case in humans (41). 306

4. Interpretation and plausibility of the learned representations. It might seem surprising for infants to be learning—as part of the 307 language acquisition process—units such as those we find, with no established linguistic interpretation. Given the relative 308 evolutionary recency of the language faculty in humans (42), however, early phonetic learning might be grounded in domain-309 general perceptual learning mechanisms (43, 44), the outcome of which might not conform to a purely linguistic interpretation. 310 Supporting this view are observations of early perceptual attunement in other modalities than speech perception—for example 311 in face (45), voice (46), pitch (47, 48), music (49) and linguistic sign (50) perception—and in other animals than humans—for 312 example for conspecific vocalizations in rats (51), for music in mice (52) and for faces in macaques (53). Furthermore, there 313 is evidence that the physiological mechanisms governing the onset and offset of perceptual attunement might be similar in 314 these different modalities and conserved from mouse to man (54-56). Furthermore, from a more adaptive/functional point 315 of view, phonetic categories embody sophisticated linguistic knowledge and inferring them from scratch might simply be 316 too difficult. The learned representations under the proposed account support remarkably accurate discrimination of native 317 language word-forms (22, 57–59)—a criterion for which early phonetic representations have been proposed to be optimized 318 (60-62). They could thus serve as a more robust intermediate point in a bootstrapping process (63) ultimately leading to 319 language proficiency. 320

Another question that arises is whether the learned representations are biologically and psychologically plausible given 321 their relatively high dimensionality—between 444 and 899 learned categories, with posterior probability vectors of matching 322 dimension. It is questionable whether infants—or even adults—would be able to explicitly access and manipulate such detailed 323 representations of the phonetics of very short stretches of speech. We believe, however, that the learned units are plausible 324 at least as lower-level perceptual representations. Such high-capacity intermediate representations are commonly postulated 325 in other domains of adult and infant cognition—for example, as part of the 'core' object recognition and the 'core' spatial 326 navigation systems (64), with corresponding computational models typically featuring representations in even higher dimensions 327 than the ones we consider here (65-68). Computation over such high-capacity representations is likely to be costly and might 328 be limited to a restricted set of operations—including the formation of integrated similarity or familiarity judgments, for 329 example. Such representations are typically seen as supporting the operation of largely subconscious cognitive processes and 330 allowing the formation of higher-level, lower-capacity, representations over which computations can be carried out more flexibly 331 (see 69, for example). 332

5. Systematic model predictions. We provide a concrete demonstration of our framework's ability to link accounts of early phonetic learning to systematic predictions regarding the empirical phenomenon they seek to explain by reporting in Table S1 phonetic contrasts of Japanese and American English for which the distributional learning mechanism we study robustly predicts a significant difference in discrimination abilities between learners of those languages. Note that nothing in our method—which we present in detail in Supplementary Materials and Methods 4—is specific to the particular distributional learning mechanism studied in this article. It applies directly to any learning mechanism taking actual speech signal as input, as long as a reasonable way to measure the (dis)similarity between the learned representations of relevant test stimuli can be provided.

Reassuringly, we find that American English [1]-[1] is among the contrasts robustly predicted to be significantly harder to discriminate for Japanese-learning infants. Only two other contrasts of American English are predicted to be robustly harder to discriminate for Japanese-learning infants, both involving the rhotacized vowel [3]. We are not aware of empirical comparisons of Japanese- and American English-learning infants (and even adults) having been carried out so far for these contrasts. No contrast of Japanese is predicted to be robustly harder for American-English-learning infants.

6. Advantages of our approach over traditional approaches to making predictions. Our approach to linking a learning mechanism to systematic predictions regarding infant phonetic discrimination relies on explicit simulations of the learning process. Such simulations have been carried out before (5–16, 19–21, 70), however this never resulted in concrete predictions regarding

<sup>\*\*</sup> This view is supported among other things by evidence of continued phonetic learning well after the first year (see e.g. 31, 37, 38)

<sup>&</sup>lt;sup>+†</sup>We do not attempt to decide between these possible interpretations here, as this is not directly relevant to our main conclusions.

infants' discrimination abilities. One reason is that previous simulation studies were conducted in the context of *outcome*-348 driven approaches and therefore focused on testing whether phonetic categories could be learned, rather than on predicting 349 discrimination patterns observed in infants. There are also methodological limitations that would have have severely limited 350 the possibility of obtaining systematic predictions in these studies. One of them is the drastically simplified input used in 351 352 most studies. Influences of the phonetic context on cross-linguistic differences in discrimination abilities (71) might fail to be 353 captured when the training data is restricted to just a few contexts, for example. Or meaningful predictions might be impossible for non-native contrasts falling into part of the phonetic space that is not represented in the input when it contains only a 354 subset of the phonetic categories of the training language (e.g. if the input consists exclusively of vowels represented in terms 355 of their formant frequencies). Even for the studies that did attempt to model infant phonetic learning from realistic speech 356 input (19, 20), the lack of a suitable evaluation method to handle the complex speech representations typically produced by 357 algorithms learning from raw speech without supervision would have prevented the derivation of systematic predictions. Indeed, 358 as we already noted, traditional signal detection theory models of discrimination tasks (72) cannot handle high-dimensional 359 input representations, while more elaborate Bayesian probabilistic models (73) typically have too many free parameters to be 360 practical. Moreover, traditional evaluation methods for representation learning algorithms from the machine learning literature 361 typically assess performance on downstream tasks such as supervised classification, or against known cluster labels, rather than 362 on the discrimination abilities measured in infants. Finally, the procurement of appropriate test stimuli for all the phonetic 363 contrasts for which predictions are to be obtained, and the need for a sound statistical methodology to separate signal from 364 noise in the large number of resulting predictions, would have presented two additional challenges. 365

In principle, an alternative to our mechanism-driven approach would be to obtain predictions by relying on pre-specified 366 notions of the outcome of learning. In phonetic category accounts, for example, predictions could be made based on how the 367 phonetic categories from the test language map onto the phonetic categories of the native language. This has been the standard 368 approach in the field until now, but to the best of our knowledge, has never resulted in the kind of systematic predictions 369 we report here. Its scalability is limited by two central difficulties related to the intrinsic complexity of the speech signal. 370 First, given that detailed aspects of the speech signal can strongly affect discrimination abilities (71, 74), making systematic 371 predictions would require extraordinarily detailed phonetic descriptions of the whole phonetic space in all of the relevant 372 languages. Such descriptions are not available at the required scale at present, and conducting detailed phonetic analyses to 373 obtain them would represent a colossal undertaking. Second, even on a small scale, how to carry out the required phonetic 374 375 analyses is not clear. Arbitrary decisions would have to be made, for example, regarding which phonetic dimensions to include, how to characterize these dimensions acoustically, how to characterize discrete categories in the presence of gradient effects, 376 and how to concretely relate the observed cross-linguistic phonetic differences to predicted discrimination abilities. Some of this 377 methodological uncertainty has been sidestepped in practice by relying on empirical assimilation patterns—adults' judgments 378 regarding what sound from their native language is most similar to a non-native stimulus—to guide the derivation of predictions 379 in an ad hoc fashion. This is not a scalable solution, however, given the costs associated with human experimentation. It also 380 fails to explain how the observed assimilation patterns arise in the first place. 381

Our modeling framework provides the first practical, scalable way to link accounts of early phonetic learning to systematic 382 predictions regardings infant phonetic discrimination. Key innovations underlying the success of our framework relative to 383 previous approaches include a focus on mechanisms rather than outcomes, and on mechanisms capable of learning from 384 naturalistic speech in particular, resulting in models capable of making systematic predictions. The testing of these models 385 at scale relies on further important innovations. One of them is the use of large forced-aligned databases of transcribed 386 continuous speech recordings to procure relevant test stimuli. Another is the use of the machine ABX test to link model 387 representation of test stimuli to concrete, systematic predictions regarding infants' discrimination abilities. The machine 388 ABX test is an automatized, parameterless measure of discriminability that is computationally tractable, statistically efficient, 389 and can handle representations in essentially any format, as long as a reasonable way to measure the similarity between the 390 speech representations to be evaluated can be provided, making it easy to compare the predictions from different models 391 (75). The rationale for such an evaluation method, with a focus on simplicity of use and scalability—rather than seeking to 392 provide a detailed model of infants' behavior in a particular experimental paradigm—is the idea that different discrimination 393 tasks all index a common perceptual process and should result in qualitatively similar discrimination patterns—an idea that 394 has received empirical support from the signal detection literature (72). Finally, another important innovation is the careful 395 statistical analysis—taking into account noise sources in both model training and evaluation (see Supplementary Materials and 396 Methods 4)—which allows us to tease out reliable effects in the large number of generated predictions. 397

Table S1. Phonetic contrasts for which a significant difference in discriminability between American English- and Japanese-learning infants is *robustly* predicted by the proposed distributional learning mechanism. That is, for each possible choice of training and test register, these contrasts show a significant difference in discrimination errors between models trained on American English and Japanese, and the magnitude of this difference does not decrease as the training data size is increased. See Supplementary Materials and Methods 4 for justification of these criteria and details of the method.

| Language    | Contrast                | Easier for<br>learners of | Average difference in<br>discrimination error |
|-------------|-------------------------|---------------------------|---|
| Am. English | [3 <sup>L</sup> ] - [I] | Am. English               | 5.4%  |
| Am. English | [3r] - [V]              | Am. English               | 4.8%  |
| Am. English | [l] - [l]               | Am. English               | 3.7%  |





Fig. S1. Average ABX error rates over all consonant and vowel contrasts obtained with each of our four Gaussian mixture models on each of the four test sets. Error bars correspond to plus and minus one standard deviation of the errors across resampling of the test stimuli speakers. On all four test sets, 'native' models make fewer discrimination errors than 'non-native' models, illustrating the robustness of the observed native advantage.





Fig. S2. ABX error rates for the American English [µ]-[I] contrast and two controls: American English [w]-[j] and average over all American English consonant contrasts. Error-rates are reported for each of the four trained Gaussian mixture models and each of the two American English test sets. Error bars correspond to plus and minus one standard deviation of the errors across resampling of the test stimuli speakers. Results show that the specific deficit for American English [µ]-[I] discrimination for 'Japanese' models compared to 'American English' models is robustly observed across all training and test conditions.



Fig. S3. Average ABX error rates over all consonant and vowel contrasts obtained with unsupervised Gaussian mixture models (GMM), with a supervised phoneme recogniser baseline (HMM) and with an input features (MFCC) baseline, as a function of the match between the training set and test set language and register. Error bars correspond to plus and minus one standard deviation of the errors across resampling of the test stimuli speakers. For both Gaussian mixture models and the phoneme recogniser baseline, the 'Native' (blue) conditions, with training and test in the same language, show fewer discrimination errors than the 'Non-native' (red) conditions. Also, in both cases the 'Native' conditions show fewer errors than the input features baseline, while 'non-native' conditions show more errors. However, the native language effect (difference between 'native' and 'non-native' models) is bigger for the supervised than the unsupervised models. Also, whereas the unsupervised models generalise very well across registers, the supervised models appear to overfit the training register.



Fig. S4. Letter-value plots(76) of the distribution of 'native' advantages across all tested phonetic contrasts (pooled over both languages) for the unsupervised Gaussian mixture models (GMM) and the supervised phoneme recogniser baseline (HMM). The native language advantage is the increase in discrimination error for a contrast of language L1 between a 'L1-native' model and a model trained on the other language, keeping the training register constant. The 'native register' advantage is the increase in error for a contrast of register R1 between a 'R1-native' model and a model trained on the other register, keeping the training language constant. For both types of models and a model trained on the other register, the other language' conditions is not driven by just a few contrasts. The 'native register' only seems to play a role for the supervised models. In particular supervised models trained on read speech appear to have trouble discriminating spontaneous speech stimuli, while supervised models trained on spontaneous speech do not have problem discriminating read speech stimuli.



**Fig. S5.** ABX error rates for the American English [*i*]-[I] contrast and two controls: American English [*w*]-[j] and average over all American English consonant contrasts (C-C). Error rates averaged over the two American English test sets and across model's training registers are reported for the unsupervised Gaussian mixture models (GMM), the supervised phoneme recogniser baseline (HMM) and the input features baseline. Error bars correspond to plus and minus one standard deviation of the errors across with both the unsupervised Gaussian mixtures and the supervised phoneme recognisers. The size of the deficit is larger for the supervised baseline, though, which we can interpret as the unsupervised GAMM models producing somewhat immature representations of speech, like those of human infants (37), while the supervised HMM models produce more adult-like representations. Another interesting result is that the supervised American English models ('native' condition, in blue) do not outperform the input features baseline in the supervised case and underperform it in the unsupervised case. This suggests that some of the detailed information relevant to discrimination that was present in the input features was not preserved through the learning of a different representation of the speech signal in terms of discrete Gaussian components (see Supplementary Discussion 3 for further discussion).



**Fig. S6.** Dissimilarity between the trained models' representation of a synthetized /ra/ stimulus and a synthetized /la/ stimulus as a function of the amount of input. These stimuli are those used in the empirical study which showed the emergence of a cross-linguistic difference in discriminability of these stimuli between Japanese- and American English-learning infants (29). For each selected duration (except when using the full training set), ten independent subsets are selected and ten independent models are trained. Solid lines indicate the average dissimilarity, with error bands indicating plus or minus one standard deviation. The dissimilarity corresponds to the average of the Kullback-Leibler divergence between posteriorgram representations of the stimuli along the dynamic time warping alignment path, expressed in bits (see Material and Methods). As the amount of input data increases, there does not appear to be much of a change in the dissimilarity of the two stimuli for the Japanese models, whereas there is sharp increase in dissimilarity for the American English models, especially between the 1-2h and 10-20h of training input. This is remarkably consistent with the empirically observed behavior of infants tested with these stimuli is observed in the ability of Japanese-learning infants to discriminate these stimuli between 6-8 and 10-12 months of age, whereas American English infants became better at it (29). The predicted cross-linguistic difference between American English and Japanese learners appears to require more input to be observed reliably when testing the models with synthetic stimuli than with natural stimuli (cf. Figure 3).



Fig. S7. As in Figure 4, with an additional ASR phone-state baseline (cf. Supplementary Materials and Methods 2). The Gaussian units in the learned (unsupervised) Gaussian mixtures are more similar to the phone-state units than to the phoneme units in the supervised baseline, although some differences remain. Even though the phone states are more numerous than the Gaussian components (a), they remain activated slightly longer on average (b) and they are better aligned with phonetic categories in terms of linguistic content, both within-speakers (c) and across speakers (d).



Read speech model





**Fig. S9.** As in Figure 4, with results for models trained on  $1/10^{th}$  subsets of the full training sets added in baby blue (these models already show a reliable cross-linguistic difference in [1]-[1] discriminability between 'American English' and 'Japanese' models, see Figure 3(b)). For the duration and acoustic (in)variance analyses (panels b c. d), results are averaged over the ten such models trained for each training corpus before standard deviations are estimated. For the number of learned units analysis (panel a), error bars show the standard deviations across the ten trained models. Models trained on  $1/10^{th}$  subsets learn much fewer categories (about one fourth as many). This is closer to the typical number of phonemes or of phonetic categories one would expect in a language. Yet, these learned units remain qualitatively different from phonetic categories as shown by the duration and acoustic (in)variance analyses (panels b, c, d). Although their average duration of activation are a few millisecond longer than for models trained on  $1/10^{th}$  subsets also appear slightly more acoustically invariant, with number of distinct units in the acoustic (in)variance tests about 80% that of the models trained on the full training sets (panels c, d). This remains much more variable than the phoneme recognizer baseline, however. Furthermore, for the acoustic (in)variance analyses models, so indicated by the results on Figure S10. Overall these analyses suggest that the failure of our models to learn of our models to learn phonetic categories cannot be attributed solely to their learning of too many categories.



**Fig. S10.** As in Figure S9 (c, d), but without applying a correction for possible misalignments of the forced-aligned phone centers (Supplementary Materials and Methods 3). For the phoneme recognizer baseline, we see that the average number of distinct units for ten repetitions of a same word shows a small increase compared to the condition with correction for misalignment, with up to about 33% more distinct units (which remains less than what was found for the unsupervised models, *with correction*). In contrast the average number of distinct units without any correction are misalignment of the phone centers is not a very common issue—as the phoneme recognizer baseline manages to find largely invariant units without any correction—suggesting that our main acoustic (in)variance analyses overestimate the acoustic invariance of the units learned by our unsupervised models by a sizable margin.



**Fig. S11.** As a convergence check, we plot the number of learned units (i.e. Gaussian components in the sampled mixture) as a function of the number of sampling iterations. Confidence bands indicate mean +/- one standard deviation in number of learned units for models trained on independent subsets. For models trained on the full corpus no confidence bands is available. The number of learned units remains stable after about 600 iterations for all models we trained, suggesting 1500 iterations was enough for our models to converge. For models trained on subsets of the full training set, we also see through the confidence bands that the number of learned categories does not depend a lot on the particular subset selected. Finally, we see evidence that for models trained on small amounts of data, the size of the training set appears to predict the number of learned units amounts of input (full training sets are about 20 hours long for models trained on read speech and about 10 hours long for model trained on spontaneous speech) learn similar amounts of ateagories initially (for  $1/1000^{th}$  and  $1/100^{th}$  training subsets), but as the size of the training sets larger (starting with  $1/10^{th}$  training subsets), models trained on similar amount of American English. This suggests that the number of learned units for the models trained on larger amounts of showing cross-linguistic differences in discrimination—does not simply reflect the amount of training input, but also the gaultative characteristics of the training sets.

#### 398 References

- 1. D Povey, et al., The kaldi speech recognition toolkit in *Proc. ASRU.* (2011).
- 2. J Flum, M Grohe, Parameterized Complexity Theory. (Springer), pp. 10–17 (2006).
- 3. Y Benjamini, D Yekutieli, The control of the false discovery rate in multiple testing under dependency. *The annals statistics* 29, 1165–1188 (2001).
- 403 4. J Lee, U-statistics: Theory and Practice. (CRC Press), (1990).
- 5. B De Boer, PK Kuhl, Investigating the role of infant-directed speech with a computer model. Acoust. Res. Lett. Online 4, 129–134 (2003).
- 6. MH Coen, Self-supervised acquisition of vowels in american english in *Proc. AAAI*. (2006).
- 407 7. GK Vallabha, JL McClelland, F Pons, JF Werker, S Amano, Unsupervised learning of vowel categories from infant-directed
   408 speech. Proc. Natl. Acad. Sci. 104, 13273–13278 (2007).
- 8. B McMurray, RN Aslin, JC Toscano, Statistical learning of phonetic categories: insights from a computational approach.
   *Dev. science* 12, 369–378 (2009).
- 9. C Jones, F Meakins, S Muawiyath, Learning vowel categories from maternal speech in gurindji kriol. Lang. Learn. 62, 1052–1078 (2012).
- <sup>413</sup> 10. F Adriaans, D Swingley, Distributional learning of vowel categories is supported by prosody in infant-directed speech in
   <sup>414</sup> Proc. COGSCI. (2012).
- <sup>415</sup> 11. B Dillon, E Dunbar, W Idsardi, A single-stage approach to learning phonological categories: Insights from inuktitut. Cogn.
   <sup>416</sup> Sci. 37, 344–377 (2013).
- 12. NH Feldman, TL Griffiths, S Goldwater, JL Morgan, A role for the developing lexicon in phonetic category acquisition.
   Psychol. review 120, 751 (2013).
- H Rasilo, O Räsänen, UK Laine, Feedback and imitation by a caregiver guides a virtual infant to learn native phonemes
   and the skill of speech inversion. Speech Commun. 55, 909–931 (2013).
- 14. S Frank, N Feldman, S Goldwater, Weak semantic context helps phonetic learning in a model of infant language acquisition
   in *Proc. ACL.* (2014).
- F Adriaans, D Swingley, Prosodic exaggeration within infant-directed speech: Consequences for vowel learnability. The J.
   Acoust. Soc. Am. 141, 3070–3078 (2017).
- 16. F Adriaans, Effects of consonantal context on the learnability of vowel categories from infant-directed speech. The J.
   Acoust. Soc. Am. 144, EL20–EL25 (2018).
- 17. RAH Bion, K Miyazawa, H Kikuchi, R Mazuka, Learning phonemic vowel length from naturalistic recordings of Japanese
   infant-directed speech. *PLoS ONE* 8, e51594 (2013).
- 429 18. S Antetomaso, et al., Modeling phonetic category learning from natural acoustic data. (Cascadilla Press), (2017).
- 430 19. K Miyazawa, H Kikuchi, R Mazuka, Unsupervised learning of vowels from continuous speech based on self-organized
   431 phoneme acquisition model in *Proc. INTERSPEECH.* (2010).
- 432 20. K Miyazawa, H Miura, H Kikuchi, R Mazuka, The multi timescale phoneme acquisition model of the self-organizing based
   433 on the dynamic features in *Proc. INTERSPEECH.* (2011).
- FH Guenther, MN Gjaja, The perceptual magnet effect as an emergent property of neural map formation. The J. Acoust.
   Soc. Am. 100, 1111–1121 (1996).
- 436 22. H Chen, CC Leung, L Xie, B Ma, H Li, Parallel inference of dirichlet process gaussian mixture models for unsupervised
   437 acoustic modeling: A feasibility study in *Proc. ISCA*. (2015).
- 438 23. J Chang, JW Fisher III, Parallel sampling of dp mixture models using sub-cluster splits in Proc. NEURIPS. (2013).
- 439 24. PK Kuhl, et al., Phonetic learning as a pathway to language: new data and native language magnet theory expanded
   (nlm-e). *Philos. Transactions Royal Soc. B: Biol. Sci.* 363, 979–1000 (2007).
- 441 25. G Dehaene-Lambertz, The human infant brain: A neural architecture able to learn language. *Psychon. bulletin & review* 442 24, 48-55 (2017).
- <sup>443</sup> 26. E Hermann, S Goldwater, Multilingual bottleneck features for subword modeling in zero-resource languages in *Proc.* <sup>444</sup> *INTERSPEECH.* (2018).
- 445 27. K Chládková, N Paillereau, The what and when of universal perception: A review of early speech sound acquisition. Lang.
   446 Learn. n/a (2020).
- 447 28. K Behnke, The Acquisition of Phonetic Categories in Young Infants: A Self-Organizing Artificial Neural Network Approach,
   448 MPI series in psycholinguistics. (MPI, Nijmegen), (1998).
- <sup>449</sup> 29. PK Kuhl, et al., Infants show a facilitation effect for native language phonetic perception between 6 and 12 months. *Dev.* <sup>450</sup> science 9, F13–F21 (2006).
- 30. T Tsushima, et al., Discrimination of english/rl/and/wy/by japanese infants at 6-12 months: language-specific develop mental changes in speech perception abilities in *Proc. ICSLP.* (1994).
- 31. K Idemaru, LL Holt, The developmental trajectory of children's perception and production of english/r/-/l. The J. Acoust.
   Soc. Am. 133, 4232–4246 (2013).
- 455 32. NH Feldman, S Goldwater, E Dupoux, T Schatz, Do infants really learn phonetic categories? Submitted (2020).
- 456 33. S Tsuji, A Cristia, Perceptual attunement in vowels: A meta-analysis. Dev. psychobiology 56, 179–191 (2014).
- 457 34. A Gagliardi, J Lidz, Statistical insensitivity in the acquisition of Tsez noun classes. Language 90, 58–89 (2014).
- 458 35. R Mugitani, et al., Perception of vowel length by japanese-and english-learning infants. Dev. psychology 45, 236 (2009).

- 459 36. Y Sato, Y Sogabe, R Mazuka, Discrimination of phonemic vowel length by japanese infants. Dev. Psychol. 46, 106 (2010).
- 460 37. DK Burnham, Developmental loss of speech perception: Exposure to and experience with a first language. Appl. 461 Psycholinguist. 7, 207–240 (1986).
- 462 38. V Hazan, S Barrett, The development of phonemic categorization in children aged 6–12. J. phonetics 28, 377–396 (2000).
- 39. R Li, T Schatz, Y Matusevych, S Goldwater, NH Feldman, Input matters in the modeling of early phonetic learning in
   *Proc. COGSCI.* (2020).
- 465 40. E Dunbar, et al., The zero resource speech challenge 2019: TTS without T. CoRR abs/1904.11469 (2019).
- 466 41. DB Pisoni, Auditory and phonetic memory codes in the discrimination of consonants and vowels. *Percept. Psychophys.* 13, 253–260 (1973).
- 468 42. WT Fitch, *The evolution of language*. (Cambridge University Press), (2010).
- 469 43. LS Scott, O Pascalis, CA Nelson, A domain-general theory of the development of perceptual discrimination. Curr.
   470 directions psychological science 16, 197–201 (2007).
- 47. 44. D Maurer, JF Werker, Perceptual narrowing during infancy: A comparison of language and faces. *Dev. Psychobiol.* 56, 154–178 (2014).
- 473 45. DJ Kelly, et al., The other-race effect develops during infancy: Evidence of perceptual narrowing. *Psychol. Sci.* 18, 1084–1089 (2007).
- 475 46. RH Friendly, D Rendall, LJ Trainor, Learning to differentiate individuals by their voices: Infants' individuation of 476 native-and foreign-species voices. *Dev. psychobiology* **56**, 228–237 (2014).
- 477 47. DJ Levitin, RJ Zatorre, On the nature of early music training and absolute pitch: A reply to brown, sachs, cammuso, and 478 folstein. *Music. Perception: An Interdiscip. J.* **21**, 105–110 (2003).
- 479 48. FA Russo, DL Windell, LL Cuddy, Learning the "special note": Evidence for a critical period for absolute pitch
   460 acquisition. *Music. Perception: An Interdiscip. J.* 21, 119–127 (2003).
- 49. EE Hannon, SE Trehub, Tuning in to musical rhythms: Infants learn more readily than adults. *Proc. Natl. Acad. Sci.*482 102, 12639–12643 (2005).
- 50. SB Palmer, L Fais, RM Golinkoff, JF Werker, Perceptual narrowing of linguistic sign occurs in the 1st year of life. *Child development* 83, 543-553 (2012).
- <sup>485</sup> 51. S Bao, Perceptual learning in the developing auditory cortex. Eur. J. Neurosci. **41**, 718–724 (2015).
- 486 52. EJ Yang, EW Lin, TK Hensch, Critical period for acoustic preference in mice. *Proc. Natl. Acad. Sci.* 109, 17213–17220 (2012).
- 488 53. EA Simpson, et al., Face detection and the development of own-species bias in infant macaques. Child development 88,
   489 103–113 (2017).
- 490 54. WM Weikum, TF Oberlander, TK Hensch, JF Werker, Prenatal exposure to antidepressants and depressed maternal
   491 mood alter trajectory of infant speech perception. *Proc. Natl. Acad. Sci.* 109, 17221–17227 (2012).
- 55. J Gervain, et al., Valproate reopens critical-period learning of absolute pitch. Front. systems neuroscience 7, 102 (2013).
   56. JF Werker, TK Hensch, Critical periods in speech perception: new directions. Annu. review psychology 66, 173–196
- 433 50. 51 Werker, TK Hensch, Critical periods in speech perception. new directions. Anna. review psychology **66**, 115–13 494 (2015).
- <sup>495</sup> 57. M Versteegh, X Anguera, A Jansen, E Dupoux, The zero resource speech challenge 2015: Proposed approaches and results.
   <sup>496</sup> Procedia Comput. Sci. 81, 67–72 (2016).
- 497 58. M Heck, S Sakti, S Nakamura, Unsupervised linear discriminant analysis for supporting dpgmm clustering in the zero
   498 resource scenario. *Procedia Comput. Sci.* 81, 73–79 (2016).
- 499 59. M Heck, S Sakti, S Nakamura, Feature optimized dpgmm clustering for unsupervised subword modeling: A contribution
   500 to zerospeech 2017 in *Proc. ASRU*. (2017).
- 501 60. PW Jusczyk, Developing Phonological Categories from the Speech Signal. (York, Timonium, MD), pp. 17–64 (1992).
- For general to language-specific capacities: the WRAPSA model of how speech perception develops. J.
   *Phonetics* 21, 3–28 (1993).
- <sup>504</sup> 62. P Jusczyk, The discovery of spoken language (1997).
- 63. E Dupoux, Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant languagelearner. Cognition 173, 43–59 (2018).
- 64. ES Spelke, SA Lee, Core systems of geometry in animal minds. *Philos. Transactions Royal Soc. B: Biol. Sci.* **367**, 2784–2793 (2012).
- 65. DL Yamins, JJ DiCarlo, Using goal-driven deep learning models to understand sensory cortex. Nat. neuroscience 19, 356–365 (2016).
- 66. C Zhuang, et al., Unsupervised neural network models of the ventral visual stream. bioRxiv (2020).
- 512 67. KL Stachenfeld, MM Botvinick, SJ Gershman, The hippocampus as a predictive map. Nat. neuroscience 20, 1643 (2017).
- 68. JC Whittington, et al., The tolman-eichenbaum machine: Unifying space and relational memory through generalisation in
   the hippocampal formation. bioRxiv (2019).
- 69. C Frith, Making up the mind: How the brain creates our mental world. (John Wiley & Sons), (2013).
- <sup>516</sup> 70. B Gauthier, R Shi, Y Xu, Learning phonetic categories by tracking movements. *Cognition* **103**, 80–106 (2007).
- 517 71. ES Levy, W Strange, Effects of consonantal context on perception of french rounded vowels by american english adults
   518 with and without french language experience. *The J. Acoust. Soc. Am.* 111, 2361–2362 (2002).
- <sup>519</sup> 72. NA Macmillan, CD Creelman, *Detection theory: A user's guide*. (Psychology press), (2004).

- 73. NH Feldman, TL Griffiths, JL Morgan, The influence of categories on perception: Explaining the perceptual magnet effect
   as optimal statistical inference. *Psychol. review* 116, 752 (2009).
- <sup>522</sup> 74. JF Werker, S Curtin, Primir: A developmental framework of infant speech processing. Lang. learning development 1,
   <sup>523</sup> 197–234 (2005).
- <sup>524</sup> 75. T Schatz, Ph.D. thesis (Université Paris 6) (2016).
- <sup>525</sup> 76. H Hofmann, K Kafadar, H Wickham, Letter-value plots: Boxplots for large data, (had.co.nz), Technical report (2011).