Early phonetic learning without phonetic categories

Insights from large-scale simulations on realistic input

Thomas Schatz[a,1], **Naomi H. Feldman**[a], **Sharon Goldwater**[b], **Xuan-Nga Cao**[c], **and Emmanuel Dupoux**[c,d]

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

This manuscript was compiled on January 28, 2021

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

Phonetic learning | Language acquisition | Computational modeling

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

Surprisingly, these patterns of perceptual confusion arise very early during language acquisition. Infants learning American English distinguish [ɹ] and [l] more easily than infants learning Japanese before they even utter their first word (4). Dozens of other instances of such *early phonetic learning* have been documented, whereby cross-linguistic confusion patterns matching those of adults emerge during the first year of life (2, 3, 30). These observations naturally led to the assumption that the same mechanism thought to be responsible for adults' perception might be at work in infants, i.e. foreign sounds are being mapped onto native phonetic categories. This assumption—which we will refer to as the *phonetic category hypothesis*—is at the core of the most influential theoretical accounts of early phonetic learning (5–7, 25, 31).

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

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 large-scale 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.

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.

[1]To whom correspondence should be addressed. E-mail: thomas.schatz.1986@gmail.com

they may be less abstract than phonemes[*]—retain a degree of abstractness and never refer to a single acoustic exemplar. For example, we would expect a given vowel or consonant in the middle of a word repeated multiple times by the same speaker to be consistently realized as the same phonetic category, despite some acoustic variation across repetitions. Finally, an added characteristic in the context of early phonetic learning is that phonetic categories are defined relative to a language. What might count as exemplars from separate phonetic categories for one language, might belong to the same category in another.

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

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

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

These results have largely been interpreted as a challenge to the idea that distributional learning is *how* infants learn phonetic categories. Additional learning mechanisms tapping into other sources of information plausibly available to infants have been proposed (9–12, 14, 15, 36, 46, 47), but existing feasibility results for such complementary mechanisms still assume that the phonetic category segmentation problem has somehow been solved and do not consider the full variability of 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 difficulty in forming phonetic categories through distributional learning: it might indicate that *what* infants learn are not phonetic categories. We are not aware of empirical results establishing that infants learn phonetic categories, and indeed, the *phonetic category hypothesis* is not universally accepted. Some of the earliest accounts of early phonetic learning were based on syllable-level categories and/or on continuous representations without any explicit category representations[‡] (49–52). Although they appear to have largely fallen out of favor, we know of no empirical findings refuting them.

We present evidence in favor of this alternative interpretation, first by showing that a distributional learning mechanism applied to raw, unsegmented, unlabeled continuous speech signal predicts early phonetic learning as observed in American English- and Japanese-learning infants—thereby providing the first realistic proof of feasibility for any account of early phonetic learning. We then show that the speech units learned through this mechanism are too brief and too acoustically variable to correspond to phonetic categories.

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

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 the learning process in infants by implementing this mechanism computationally and training it on naturalistic speech recordings in a target language—either Japanese or American English. This yields a candidate model for the early phonetic knowledge of, say, a Japanese infant. Next, we assess the model's ability to discriminate phonetic contrasts of American English and Japanese—for example American English [ɹ] vs [l]—by simulating a discrimination task using speech stimuli corresponding to this contrast. We test whether the predicted discrimination patterns agree with the available empirical record on cross-linguistic differences between American

[*] 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.

[†] 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).

[‡] 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.

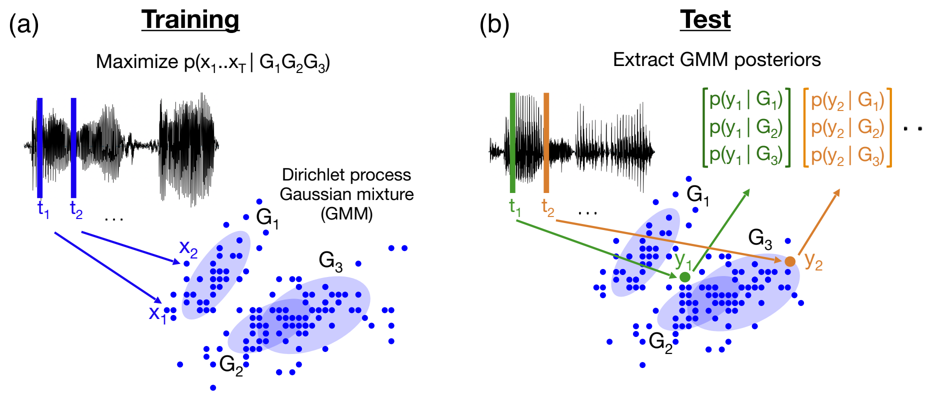

Fig. 1. Gaussian mixture model training and representation extraction, illustrated for a model with three Gaussian components. In practice the number of Gaussian components is learned from the data and much higher. (a) Model training: the learning algorithm extracts moderate-dimensional (d=39) descriptors 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 descriptors for 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 infants. 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 recent advances in the field of machine learning, and more specifically in *unsupervised representation learning* for speech technology, which have established that, given only raw, untranscribed, unsegmented speech recordings, it is possible to learn representations that accurately discriminate the phonetic categories of a language (53–70). The learning algorithms considered have been argued to be particularly relevant for modeling how infants learn in general, and learn language in particular (71). Among available *learning algorithms*, we select the one at the core of the winning entries in the Zerospeech 2015 and 2017 international competitions in unsupervised speech representation learning (58, 59, 69). Remarkably, it is based on a Gaussian mixture clustering mechanism—illustrated in Figure 1 (a)—that can straightforwardly be interpreted as a form of distributional learning (8, 10). A different *input representation* to the Gaussian mixture is used than in previously proposed implementations of distributional learning, however (9, 12, 14, 40, 42, 44, 45). Simple descriptors of the shape of the speech signal's short-term auditory spectrum sampled at regular points in time (every 10ms) (72) are used instead of traditional phonetic measurements obtained separately for each vowel and consonant, such as formant frequencies or harmonic amplitudes.[§] This type of input representation only assumes basic auditory abilities from infants, which are known to be fully operational shortly after birth (75), and has been proposed previously as a potential way to get around both the lack of invariance and the phonetic category segmentation problems in the context of adult word recognition (37). A second difference from previous implementations of distributional learning is in the *output representation*. Test stimuli are represented as sequences of posterior probability vectors (posteriorgrams) over K Gaussian components in the mixture (Figure 1 (b)), rather than simply being assigned to the most likely Gaussian component. These continuous representations have been shown to support accurate discrimination of native phonetic categories in the Zerospeech challenges.

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 (85)	Japanese	Read	19h33	9h40	96	47
Sp-Eng (86)	Am. English	Spont.	9h13	9h01	20	20
Sp-Jap (87)	Japanese	Spont.	9h11	8h57	20	20

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

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

[§] 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 novel predictions regarding the differences in speech discrimination abilities between these two populations. Because we do not assume that the outcome of infants' learning is adult-like knowledge, we can only rely on infant data for evaluation. The absence of specific assumptions *a priori* about what is going to be learned, and the sparsity of empirical data on infant discrimination, makes this challenging. The algorithm we consider outputs complex, high-dimensional representations (Figure 1 (b)) that are not easy to link to concrete predictions regarding infant discrimination abilities. Traditional signal detection theory models of discrimination tasks (88) cannot handle high-dimensional perceptual representations, while more elaborate (Bayesian) probabilistic models (89) have too many free parameters given the scarcity of available data from infant experiments. We rely instead on the *machine ABX* approach that we previously developed (90, 91). It consists of a simple model of a discrimination task, which can handle any representation format provided the user can provide a reasonable measure of (dis)similarity between representations (90, 91). This is not a detailed model of infant's performance in a specific experiment, but rather a simple and effectively parameterless way to systematically link the complex speech representations produced by our models to predicted discrimination patterns. For each trained model and each phonetic contrast of interest, we obtain an 'ABX error rate' such that 0% and 50% error indicate perfect and chance-level discrimination, respectively. This allows us to evaluate the qualitative match between the model's discrimination abilities and the available empirical record in infants (see Supplementary Discussion 3 for an extended discussion of our approach to interpreting the simulated discrimination errors and relating them to empirical observations, including why it would not be meaningful to seek a quantitative match at this point).

Finally, we investigate whether the learned Gaussian components correspond to phonetic categories. We first compare the number of Gaussians in a learned mixture to the number of phonemes in the training language (*category number* test): although a phonetic category can be more concrete than a phoneme, the number of phonetic categories documented in typical linguistic analyses remains on the same order of magnitude as the number of phonemes. We then administer two diagnostic tests based on the two defining characteristics identified above that any representation corresponding to phonetic categories should pass.[¶] The first characteristic is size/duration: a phonetic category is a phoneme-sized unit (i.e. the size of a vowel or a consonant). Our *duration* test probes this by measuring the average duration of activation of the learned Gaussian components (a component is taken to be 'active' when its posterior probability is higher than all other components), and comparing this to the average duration of activation of units in a baseline system trained to recognize phonemes with explicit supervision. The second characteristic is abstractness: although phonetic categories can depend on phonetic context[‖] and on non-linguistic properties of the speech signal—e.g. the speaker's gender—at a minimum, the central phone in the same word repeated several times by the same speaker is expected to be consistently realized as the same phonetic category. Our *acoustic (in)variance* test probes this by counting the number of distinct representations needed by our model to represent ten occurrences of the central frame of the central phone of the same word either repeated by the same speaker (within speaker condition) or by different speakers (across speaker condition). We use a generous correction to handle possible misalignment (see Materials and Methods). The last two tests can be related to the phonetic category segmentation and lack of invariance problems: solving the phonetic category segmentation problem involves finding units that would pass the *duration* test, while solving the lack of invariance problem involves finding units that would pass the *acoustic (in)variance* test. Given the laxity in the use of the concept of phonetic category in the literature, some might be tempted to challenge that even these diagnostic tests can be relied on. If they cannot, however, it is not clear to us how phonetic category accounts of early phonetic learning should be understood as scientifically refutable claims.

[¶] 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 $[t^h]$ (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 $[t^h]$.

Results

Overall discrimination. After having trained a separate model for each of the four possible combinations of language and register, we test whether the models' overall discrimination abilities, like those of infants (2, 3, 30), are specific to their 'native' (i.e. training) language. Specifically, for each corpus, we look at overall discrimination errors averaged over all consonant and vowel contrasts available in a held-out test set from that corpus (See Table 1). We tested each of the two American English-trained and each of the two Japanese-trained models on each of four test sets, yielding a total of 4×4 discrimination errors. We tabulated the average errors in terms of 4 conditions depending on the relation between the test set and the training background of the model: native versus non-native contrasts and same versus different register. The results are reported in Figure 2 (see also Figures S1, S4 for non-tabulated results). Panel (a) shows that discrimination performance is higher on average in matched-language conditions (in blue) than in mismatched-language conditions (in red). In contrast, register mismatch has no discernible impact on discrimination performance. A comparison with a supervised phoneme recognizer baseline (Figure S3) shows a similar pattern of results, but with a larger absolute cross-linguistic difference. If we interpret this supervised baseline as a proxy to the adult state, then our model suggests that infant's phonetic representations, while already language-specific, remain 'immature'.[**] Panel (b) shows the robustness of these results, with 81.7% of the 1295 distinct phonetic contrasts tested proving easier to discriminate on the basis of representations from a model trained on the matching language. Taken together, these results suggest that, similar to infants, our models acquire language-specific representations, and that these representations generalize across register.

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

[**] This is compatible with empirical evidence that phonetic learning continues into childhood well beyond the first year (see 92–94, for example).

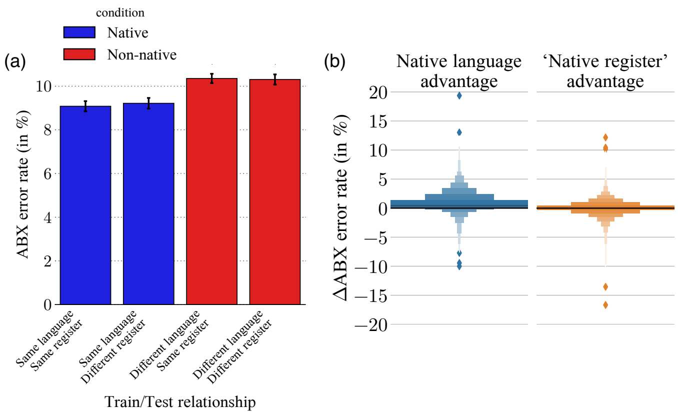

Fig. 2. (a) Average ABX error rates over all consonant and vowel contrasts obtained with our models as a function of the match between the training set and test set language and register. Error bars correspond to plus and minus one standard deviation of the errors across resampling of the test stimuli speakers. The 'Native' (blue) 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 performed similarly in discriminating this contrast, 10- to 12-month-old American English infants outperformed their Japanese peers. We compare the discrimination errors obtained with each of our four models for American English [ɹ]-[l] and for two controls: the American English [w]-[j] contrast (as in 'wet' versus 'yet'), for which we do not expect a gap in performance between American English and Japanese natives (96), and the average error over all the other consonant contrasts of American English. For each contrast and for each of the four models, we average discrimination errors obtained on each of the two American English held-out test sets, yielding 3×4 discrimination errors. We further average over models with the same 'native' language to obtain 3×2 discrimination errors. The results are shown in Figure 3 (see also Figures S2 and S6 for untabulated results and a test confirming our results with the synthetic stimuli used in the original infant experiment, respectively). In panel (a), we see that, similar to 10- to 12-month old infants, American English 'native' models (in blue) greatly outperform Japanese 'native' models (in red) in discriminating American English [ɹ]-[l]. Here again a supervised phoneme recognizer baseline yields a similar pattern of results, but with larger cross-linguistic differences (panel (c), see also Figure S5), again suggesting that the representations learned by the unsupervised models—like those of infants—remain somewhat 'immature'. In panel (b), we see results obtained by training ten different models on ten different subsets of the training set of each corpus, varying the sizes of the subsets (see Materials and Methods for more details). It reveals that one hour of input is sufficient for the divergence between the Japanese and English models to emerge robustly, and that this divergence increases with exposure to the native 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 nature of the learned representations and test whether what has been learned can be understood in terms of phonetic categories. Results are reported in Figure 4 (see also Figure S7 for comparisons with a different supervised baseline). First, looking at the *category number* criterion in Figure 4 (a), we see that our models learned more than ten times as many categories as the number of phonemes in the corresponding languages. Even allowing for notions of phonetic categories more granular than phonemes, we are not aware of any phonetic analysis ever reporting that many allophones in these languages. Second, looking at the *duration* criterion in Figure 4 (b), the learned Gaussian units appear to be activated on average for about a quarter the duration of a phoneme. This is shorter than any linguistically identified unit. It shows that the phonetic category segmentation problem has not been solved. Next, looking at the *acoustic (in)variance* criterion in Figure 4 (c) and (d)—for the within and across speakers conditions, respectively—we see that our models require on average around two distinct representations to represent ten tokens of the same phonetic category without speaker variability, and three distinct representations across different speakers. The supervised phoneme recognizer baseline establishes that our results cannot be explained by defective test stimuli. Instead, this result shows that the learned units are finer-grained than phonetic categories along the spectral axis, and that the lack of invariance problem has not been solved. Based on these tests, we can conclude that the learned units do not correspond to phonetic categories in any meaningful sense of the term.

Discussion

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

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

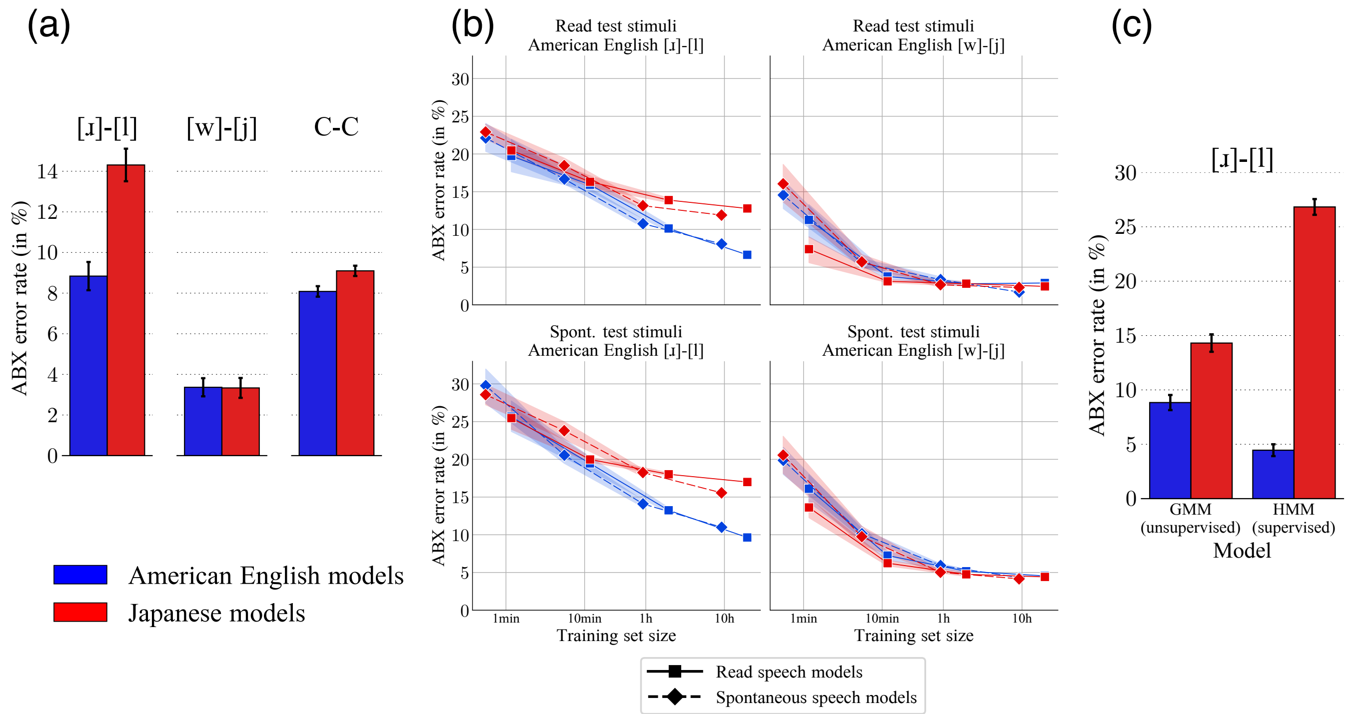

Fig. 3. (a) ABX error rates for the American English [ɹ]-[l] contrast and two controls: American English [w]-[j] and average over all American English consonant contrasts (C-C). Error rates are reported for two conditions: average over models trained on American English and average over models trained 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 [ɹ]-[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 [ɹ]-[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 [ɹ]-[l] discrimination for 'Japanese-native' models robustly emerges with as little as 1h of training data. (c) To give a sense of scale we compare the cross-linguistic difference obtained with the unsupervised Gaussian mixture models on American English [ɹ]-[l] (GMM, left) to the one obtained with supervised phoneme recognizer baselines (HMM, right). The larger cross-linguistic difference obtained with the supervised baselines suggests that the representations learned by our unsupervised models, similar to those observed in infants, remain somewhat immature.

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

A new account of early phonetic learning. Our results suggest an account of phonetic learning that substantially differs from existing ones. Whereas previous proposals have been primarily motivated through an *outcome-driven* perspective—starting from assumptions about what it is about language that is learned—the motivation for the proposed account comes from a *mechanism-driven* perspective—starting from assumptions about how learning might proceed from the infant's input. This contrast is readily apparent in the choice of the initial speech representation upon which the early phonetic learning process operates (the input representation). Previous accounts assumed speech to be represented innately through a set of universal (i.e. language-independent) phonetic feature detectors (5–7, 25, 31, 49–52). The influential phonetic category accounts furthermore assumed these features to be available phonetic segment by phonetic segment (i.e. for each vowel and consonant separately) (5–7, 25, 31). While these assumptions are attractive from an *outcome-driven* perspective—they connect transparently to phonological theories in linguistics and theories of adult speech perception that assume a decomposition of speech into phoneme-sized segments defined in terms of abstract phonological features—from a *mechanism-driven* perspective, both assumptions are difficult to reconcile with the continuous speech signal that infants hear. The lack of acoustic-phonetic invariance problem challenges the idea of phonetic feature detectors, and the phonetic category segmentation problem challenges the idea that the relevant features are segment-based (37–39). The proposed account does not assume either problem to be solved by infants at birth. Instead, it relies on basic auditory abilities that are available to neonates (75), using simple auditory descriptors of the speech spectrum obtained regularly along the time axis. This type of spectrogram-like representation is effective in speech technology applications (72) and can be seen as the output of a simple model of the peripheral auditory system (91, chap. 3), which is fully operational shortly after birth (75). Such representations have also been proposed before as an effective

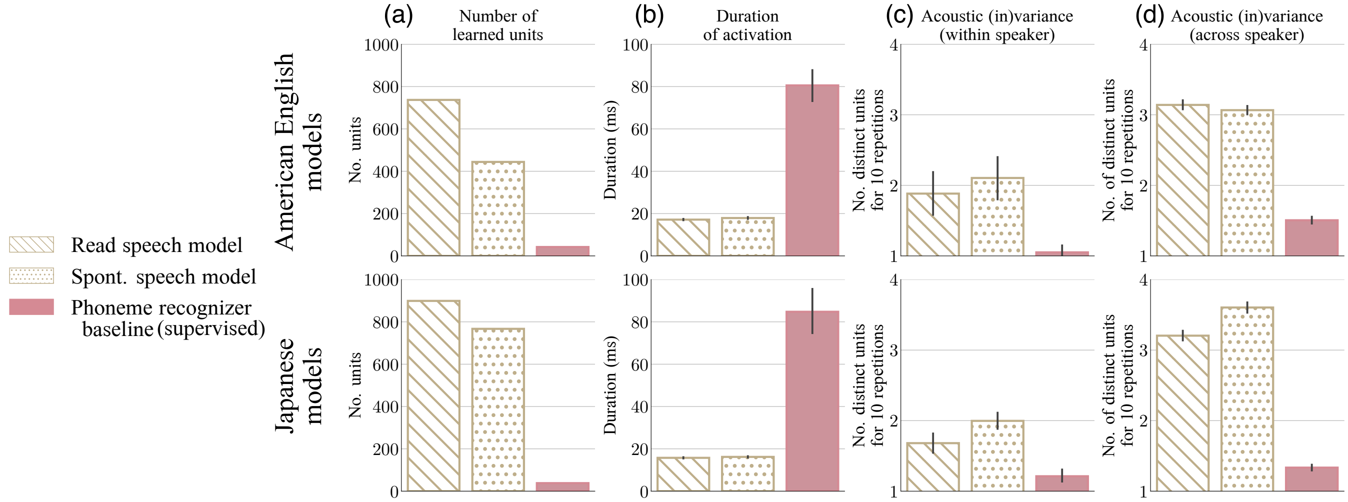

Fig. 4. Diagnostic test results for our four unsupervised Gaussian mixture models (in beige) and phoneme recogniser baselines trained with explicit supervision (in pink). Top row: American English 'native' models. Bottom row: Japanese 'native' models. Models are tested on read speech in their 'native' language. (a) Number of units learned by the models. Gaussian mixtures discover ten to twenty 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 each unit is computed and the average and standard deviation of 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 correspond to phonetic categories. (c) Average number of distinct representations for the central frame of the central phone for ten repetitions of a same word by the same speaker, corrected for possible misalignment. The number of distinct representations is computed for each word type with sufficient repetitions in the test set and the average and standard deviation of the resulting distribution over word types are shown. The phoneme recogniser baseline reliably identifies the ten tokens as exemplars from a common phonetic category, whereas our Gaussian mixture models typically maintain on the order of two distinct representations, indicating representations too fine-grained to be phonetic categories. (d) As in (c) but with repetitions of a same word by ten speakers, showing that the learned Gaussian units are not speaker-independent.

way to get around both the lack of invariance and the phonetic category segmentation problems in the context of adult word recognition (37) and can outperform representations based on traditional phonetic measurements (like formant frequencies) as predictors of adult speech perception (106–110).

While the input representation is different, the learning mechanism in the proposed account—distributional learning—is similar to what had originally been proposed in phonetic category accounts. Infants' abilities, both in the lab (8, 35) and in ecological conditions (34), are consistent with such a learning mechanism. Moreover, when applied to the input representation considered in this paper, distributional learning is adaptive in that it yields speech representations that can support remarkably accurate discrimination of the phonetic categories of the training language, outperforming a number of alternatives that have been proposed for unsupervised speech representation learning (58, 59, 69).

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

Phonetic categories are often assumed as precursors in accounts of phenomena occurring later in the course of language acquisition. Our account does not necessarily conflict with this view, as phonetic categories may be learned later in development, before phonological acquisition. Alternatively, the influential *PRIMIR* account of early language acquisition (7) proposes that infants learn in parallel about the phonetics, word-forms, and phonology of their native language, but do not develop abstract phonemic representations until well into their second year of life. Although *PRIMIR* explicitly assumes phonetic learning to be *phonetic category* learning, other aspects of their proposed framework do not depend on that assumption, and our framework may be able to stand in for the phonetic learning process they assume.

To sum up, we introduced and motivated a new account of early phonetic learning and showed that it is feasible under realistic learning conditions, which cannot be said of any other account at this time. Importantly, this does not constitute decisive evidence for our account over alternatives. Our primary focus has been on modeling cross-linguistic differences in the perception of one contrast, [ɹ]-[l]; further work is necessary to determine to what extent our results extend to other contrasts and languages (111). Furthermore, an absence of feasibility proof does not amount to a proof of infeasibility. While we have preliminary evidence that simply forcing the model to learn fewer categories is unlikely to be sufficient (Figures S9 and S10), recently proposed partial solutions to the phonetic category segmentation problem (e.g. (112–114)) and to the lack of invariance problem (115) (see also Supplementary Discussion 2 regarding the choice of model initialization) might yet lead to a feasible phonetic category-based account, for example. In addition, a number of other representation learning algorithms proposed in the context of unsupervised speech technologies and building on recent developments in the field of machine learning have yet to be investigated (53–70). They might provide concrete implementations of previously proposed accounts of early phonetic learning or suggest new ones altogether. This leaves us with a large space of appealing theoretical possibilities, making it premature to commit to a particular account. Candidate accounts should instead be evaluated on their ability to predict empirical data on early phonetic learning, which brings us to the second main contribution of this article.

Toward predictive theories of early phonetic learning. Almost since the original empirical observation of early phonetic learning (1), a number of theoretical accounts of the phenomenon have co-existed (6, 25, 49, 50). This theoretical under-determination has typically been thought to result from the scarcity of empirical data from infant experiments. We argue instead that the main limiting factor on our understanding of early phonetic learning might have been the lack—on the theory side—of a practical method to link proposed accounts of phonetic learning with concrete, systematic predictions regarding the empirical discrimination data they seek to explain. Establishing such a systematic link has been challenging due to the necessity of dealing with the actual speech signal, with all its associated complexity. The modeling framework we introduce provides, for the first time, a practical and scalable way to overcome these challenges and obtain the desired link for phonetic learning theories—a major methodological advance, given the fundamental epistemological importance of linking *explanandum* and *explanans* in scientific theories (116).

Our mechanism-driven approach to obtaining predictions—which can be applied to any phonetic learning model implemented in our framework—consists first of explicitly simulating the early phonetic learning process as it happens outside of the lab, which results in a trained model capable of mapping any speech input to a model representation for that input. The measurement of infants' perceptual abilities in laboratory settings—including their discrimination of any phonetic contrast—can then be simulated on the basis of the model's representations of the relevant experimental stimuli. Finally, phonetic contrasts for which a significant cross-linguistic difference is robustly predicted can be identified through a careful statistical analysis of the simulated discrimination judgments (see Supplementary Materials and Methods 4). As an illustration of how such predictions can be generated, we report specific predictions made by our distributional learning model in Table S1 (see also Supplementary Discussion 5).

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

Our framework readily generates novel, empirically testable, predictions regarding infants' discrimination, yet further computational modeling is called for before we return to experiments. Indeed, existing data—collected over more than three decades of research (2, 3, 21, 30)—might already suffice to distinguish between different learning mechanisms. To make that determination, and to decide which contrasts would be most useful to test next in case more data are needed, many more learning mechanisms and training/test language pairs will need to be studied. Even for a specified learning mechanism and training/test datasets, multiple implementations should ideally be compared (e.g. testing different parameter settings for the input representations or the clustering algorithm), as implementational choices that weren't initially considered to be important might nevertheless have an effect on the resulting predictions and thus need to be included in our theories. Conversely, features of the model that may seem important *a priori* (e.g. the type of clustering algorithm used) might turn out to have little effect on the learning outcomes in practice.

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

Materials and Methods

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

Signal processing, models and inference. The raw speech signal is decomposed into a sequence of overlapping 25ms-long frames sampled every 10ms and moderate-dimensional (d=39) descriptors of the spectral shape of each frame are then extracted, describing how energy in the signal spreads across different frequency channels. The descriptors are comprised of 13 mel-frequency cepstral coefficients (MFCC) with their first and second time derivatives. These

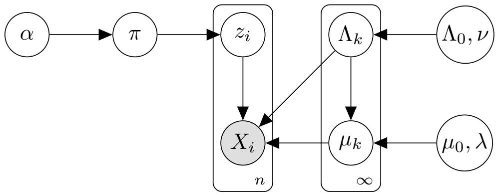

Fig. 5. Generative Gaussian mixture model with Dirichlet process prior with normal-inverse-Wishart base measure, represented as a graphical model in plate notation based on the stick-breaking construction of Dirichlet processes.

coefficients correspond approximately to the principal components of spectral slices in a log-spectrogram of the signal, where the spectrogram frequency channels are selected on a mel frequency scale (linear for lower frequency and logarithmic for higher frequencies, matching the frequency selectivity of the human ear).

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 Normal-inverse-Wishart base measure. The generative model is depicted as a graphical model in plate notation in Figure 5, where n is 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 hyper-parameters. The depicted variables have the following conditional distributions:

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

for any $1 \leq i \leq n$, for any $k \in \{1, 2, ...\}$, with $\mathcal{N}$ the multivariate Gaussian distribution, $\mathcal{W}$ the Wishart distribution, *Multi* the generalisation of the usual multinomial probability distribution to an infinite discrete support and *SB*, the mixing weights generating distribution from the stick-breaking representation of Dirichlet processes (120). Mixture parameters with high posterior probability given the observed input features vectors and the prior are found using an efficient parallel Markov chain Monte Carlo sampler (121). Following previous work (61, 66), model initialization is performed by partitioning training points uniformly at random into ten clusters and the hyperparameters are set as follows: α to 1, μ_0 to the average of all input features vectors, λ to 1, λ_0 to the inverse of the covariance of all input feature vectors and ν to 42 (i.e. the spectral shape descriptors dimension plus three). We additionally train a model on each of the ten $1/10^{th}$, $1/100^{th}$ and $1/1000^{th}$ training subsets of each of the four corpora, following the same procedure.

Given a trained Gaussian mixture with K components, mixing weights $(\pi_1, \pi_2, ..., \pi_K)$, means $(\mu_1, \mu_2, ..., \mu_K)$ and covariance matrices $(\Sigma_1, \Sigma_2, ..., \Sigma_K)$, we extract a test stimulus representation from the sequence $(x_1, x_2, ..., x_m)$ of spectral-shape descriptors for that stimulus, as the sequence of posterior probability vectors $(p_1, p_2, ..., p_m)$ where for any frame i, $1 \leq i \leq m$, $p_i = (p_{i1}, p_{i2}, ..., p_{iK})$, with, for any $1 \leq k \leq K$:

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

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 representations for phonetic contrasts of interest is assessed using machine ABX discrimination errors (90, 91). Discrimination is assessed *in context*, defined as the preceding and following sound and the identity of the speaker. For example, discrimination of American English [u] versus [i] is assessed in each available context independently, yielding—for instance—a separate discrimination error rate for test stimuli in [b]_[t] phonetic context, as in 'boot' versus 'beet', as spoken by a specified speaker. Other possible factors of variability, such as word boundaries or syllable position are not controlled. For each model, each test corpus and each phonemic contrast in that test corpus (as specified by the corpus' phonemic transcriptions), we obtain a discrimination error for each context in which the contrasted phonemes occur at least twice in the test corpus' test set. 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.

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_δ tokens of phonetic category δ and n_ξ tokens of phonetic category ξ, the *non-symmetrized Machine ABX discrimination error* between δ and ξ is then estimated as the proportion of representation triplets a, b, x, with a and x taken from Δ and b taken from Ξ, such that x is closer to b than to a, i.e.:

$$\hat{e}(\Delta, \Xi) := \frac{1}{n_\delta(n_\delta - 1)n_\xi} \sum_{a=1}^{n_\delta} \sum_{b=1}^{n_\xi} \sum_{\substack{x=1 \\ x \neq a}}^{n_\delta} \Big[\mathbb{1}_{d(\xi_b, \delta_x) < d(\delta_a, \delta_x)} + \frac{1}{2} \mathbb{1}_{d(\xi_b, \delta_x) = d(\delta_a, \delta_x)} \Big],$$

where $\mathbb{1}$ 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 δ and ξ is then obtained as:

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

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

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

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.

The acoustic (in)variance of the learned units is probed by looking at multiple repetitions of a single word and testing whether the dominant unit at the central frame of the central phone of the word remains the same for all repetitions. Specifically, we count the number of distinct dominant units occurring at the central frame of the central phone for ten repetitions of the same word. To compensate for possible misalignment of the central phones' central frames (e.g. due to slightly different time courses in the acoustic realization of the phonetic segment and/or small errors in the forced alignment), we allow the dominant unit at the central frame to be replaced by any unit that was dominant at some point within the previous or following 46ms (thus covering a 92ms slice of time corresponding to the average duration of a phoneme in our read speech test sets), provided it can bring down the overall count of distinct dominant units for the ten occurrences (see Supplementary Materials and Methods 3 for more information). We consider two conditions: in the *within-speaker* condition, the test stimuli are uttered by the same speaker ten times; in the *across-speaker* condition, they are uttered by ten different speakers one time. See Supplementary Materials and Methods 3 for more information on the stimulus selection procedure.

Data and code availability. The datasets analysed in this study are publicly available from the commercial vendors and research institutions holding their copyrights (84–87). Datasets generated during the course of the study are available from the corresponding author upon reasonable request. Code to reproduce the results will be made available at https://github.com/Thomas-Schatz/perceptual-tuning-pnas upon publication.

ACKNOWLEDGMENTS. We thank the editor, anonymous reviewers, and Yevgen Matusevych for their helpful comments on the manuscript. X-NC and ED's contributions at Cognitive Machine Learning were supported by the Agence Nationale pour la Recherche (ANR-17-EURE-0017 Frontcog, ANR-10-IDEX-0001-02 PSL*, ANR-19-P3IA-0001 PRAIRIE 3IA Institute), by a grant from Facebook AI Research and from CIFAR (LMB). TS and NF's contributions were supported by NSF BCS-1734245, and SG's contributions by ESRC ES/R006660/1 and JSMF 220020374.

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).
2. JF Werker, RC Tees, Influences on infant speech processing: Toward a new synthesis. *Annu. review psychology* **50**, 509–535 (1999).
3. S Tsuji, A Cristia, Perceptual attunement in vowels: A meta-analysis. *Dev. psychobiology* **56**, 179–191 (2014).
4. PK Kuhl, et al., Infants show a facilitation effect for native language phonetic perception between 6 and 12 months. *Dev. science* **9**, F13–F21 (2006).
5. 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).
6. CT Best, , et al., The emergence of native-language phonological influences in infants: A perceptual assimilation model. *The development speech perception: The transition from speech sounds to spoken words* **167**, 233–277 (1994).
7. JF Werker, S Curtin, Primir: A developmental framework of infant speech processing. *Lang. learning development* **1**, 197–234 (2005).
8. J Maye, JF Werker, L Gerken, Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition* **82**, B101–B111 (2002).
9. F Adriaans, D Swingley, Distributional learning of vowel categories is supported by prosody in infant-directed speech in *Proc. COGSCI.* (2012).
10. JF Werker, HH Yeung, KA Yoshida, How do infants become experts at native-speech perception? *Curr. Dir. Psychol. Sci.* **21**, 221–226 (2012).
11. O Räsänen, Computational modeling of phonetic and lexical learning in early language acquisition: existing models and future directions. *Speech Commun.* **54**, 975–997 (2012).
12. NH Feldman, TL Griffiths, S Goldwater, JL Morgan, A role for the developing lexicon in phonetic category acquisition. *Psychol. review* **120**, 751 (2013).
13. 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).
14. S Frank, N Feldman, S Goldwater, Weak semantic context helps phonetic learning in a model of infant language acquisition in *Proc. ACL.* (2014).
15. F Adriaans, D Swingley, Prosodic exaggeration within infant-directed speech: Consequences 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).
18. H Goto, Auditory perception by normal Japanese adults of the sounds "L" and "R". *Neuropsychologia* **9**, 317–323 (1971).
19. K Miyawaki, et al., An effect of linguistic experience: The discrimination of [r] and [l] by native speakers of Japanese and English. *Percept. & Psychophys.* **18**, 331–340 (1975).
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).
21. W Strange, *Speech perception and linguistic experience: Issues in cross-language research.* (York Press), (1995).
22. JS Logan, SE Lively, DB Pisoni, Training japanese listeners to identify english/r/and/l: A first 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 without french language experience. *J. Phonetics* **36**, 141–157 (2008).
25. PK Kuhl, KA Williams, F Lacerda, KN Stevens, B Lindblom, Linguistic experience alters 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. 233–277 (1995).
27. CT Best, *A direct realist view of cross-language speech perception.* (York Press), pp. 171–206 (1995).
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 native language magnet theory in *Developmental neurocognition: Speech and face processing in the first year of life.* (Springer), pp. 259–274 (1993).
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).
34. A Cristia, Fine-grained variation in caregivers'/s/predicts their infants'/s/category. *The J. Acoust. Soc. Am.* **129**, 3271–3280 (2011).
35. A Cristia, Can infants learn phonology in the lab? a meta-analytic answer. *Cognition* **170**, 312–327 (2018).
36. D Swingley, Contributions of infant word learning to language development. *Philos. Transactions Royal Soc. B: Biol. Sci.* **364**, 3617–3632 (2009).
37. DH Klatt, *Speech perception: A model of acoustic-phonetic analysis and lexical access.* pp. 243–288 (1980).
38. D Shankweiler, W Strange, R Verbrugge, *Speech and the problem of perceptual constancy.* pp. 315–345 (1977).
39. I Appelbaum, The lack of invariance problem and the goal of speech perception in *Proc. ICSLP.* (1996).
40. 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).
42. GK Vallabha, JL McClelland, F Pons, JF Werker, S Amano, Unsupervised learning of vowel 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).
44. B McMurray, RN Aslin, JC Toscano, Statistical learning of phonetic categories: insights from a computational approach. *Dev. science* **12**, 369–378 (2009).
45. C Jones, F Meakins, S Muawiyath, Learning vowel categories from maternal speech in gurindji kriol. *Lang. Learn.* **62**, 1052–1078 (2012).
46. B Dillon, E Dunbar, W Idsardi, A single-stage approach to learning phonological categories: Insights from inuktitut. *Cogn. Sci.* **37**, 344–377 (2013).
47. 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).
48. 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).
49. 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).
50. PW Jusczyk, *Developing Phonological Categories from the Speech Signal.* (York, Timonium, MD), pp. 17–64 (1992).
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).
53. B Varadarajan, S Khudanpur, E Dupoux, Unsupervised learning of acoustic sub-word units in *Proc. ACL.* (2008).
54. AS Park, JR Glass, Unsupervised pattern discovery in speech. *IEEE Transactions on Audio, Speech, Lang. Process.* **16**, 186–197 (2008).
55. Cy Lee, J Glass, A nonparametric bayesian approach to acoustic model discovery in *Proc. ACL.* (2012).
56. A Jansen, et al., A summary of the 2012 jhu clsp workshop on zero resource speech technologies 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: Proposed approaches and results. *Procedia Comput. Sci.* **81**, 67–72 (2016).
60. L Ondel, L Burget, J Černockỳ, Variational inference for acoustic unit discovery. *Procedia Comput. Sci.* **81**, 80–86 (2016).
61. H Chen, CC Leung, L Xie, B Ma, H Li, Parallel inference of dirichlet process gaussian mix-

ture 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. INTERSPEECH.* (2015).
63. H Kamper, M Elsner, A Jansen, S Goldwater, Unsupervised neural network based feature 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 for unsupervised representation learning on the zero resource speech challenge in *Proc. INTERSPEECH.* (2015).
65. N Zeghidour, G Synnaeve, M Versteegh, E Dupoux, A deep scattering spectrum—deep siamese network pipeline for unsupervised acoustic modeling in *Proc. ICASSP.* (2016).
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).
67. M Heck, S Sakti, S Nakamura, Feature optimized dpgmm clustering for unsupervised subword modeling: A contribution to zerospeech 2017 in *Proc. ASRU.* (2017).
68. WN Hsu, Y Zhang, J Glass, Unsupervised learning of disentangled and interpretable representations from sequential data in *Proc. NEURIPS.* (2017).
69. E Dunbar, et al., The zero resource speech challenge 2017 in *Proc. ASRU.* (2017).
70. J Chorowski, RJ Weiss, S Bengio, Avd Oord, Unsupervised speech representation learning using wavenet autoencoders. *CoRR* **abs/1901.08810** (2019).
71. E Dupoux, Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner. *Cognition* **173**, 43–59 (2018).
72. P Mermelstein, Distance measures for speech recognition, psychological and instrumental. *Pattern recognition artificial intelligence* **116**, 91–103 (1976).
73. K Miyazawa, H Kikuchi, R Mazuka, Unsupervised learning of vowels from continuous speech based on self-organized phoneme acquisition model in *Proc. INTERSPEECH.* (2010).
74. K Miyazawa, H Miura, H Kikuchi, R Mazuka, The multi timescale phoneme acquisition model of the self-organizing based on the dynamic features in *Proc. INTERSPEECH.* (2011).
75. JR Saffran, JF Werker, LA Werner, *The infant's auditory world: Hearing, speech, and the beginnings of language.* (Wiley Online Library), (2006).
76. PK Kuhl, et al., Cross-language analysis of phonetic units in language addressed to infants. *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).
80. 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).
89. 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).
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).
91. T Schatz, Ph.D. thesis (Université Paris 6) (2016).
92. DK Burnham, Developmental loss of speech perception: Exposure to and experience with 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).
96. T Tsushima, et al., Discrimination of english/rl/and/wy/by japanese infants at 6-12 months: language-specific developmental changes in speech perception abilities in *Proc. ICSLP.* (1994).
97. PK Kuhl, FM Tsao, HM Liu, Foreign-language experience in infancy: Effects of short-term exposure and social interaction on phonetic learning. *Proc. Natl. Acad. Sci.* **100**, 9096–9101 (2003).
98. T Teinonen, RN Aslin, P Alku, G Csibra, Visual speech contributes to phonetic learning in 6-month-old infants. *Cognition* **108**, 850–855 (2008).
99. HH Yeung, JF Werker, Learning words' sounds before learning how words sound: 9-month-olds use distinct objects as cues to categorize speech information. *Cognition* **113**, 234–243 (2009).
100. NH Feldman, EB Myers, KS White, TL Griffiths, JL Morgan, Word-level information influences phonetic learning in adults and infants. *Cognition* **127**, 427–438 (2013).
101. N Mani, S Schneider, Speaker identity supports phonetic category learning. *J. Exp. Psychol. Hum. Percept. Perform.* **39**, 623 (2013).
102. HH Yeung, T Nazzi, Object labeling influences infant phonetic learning and generalization. *Cognition* **132**, 151–163 (2014).
103. HH Yeung, LM Chen, JF Werker, Referential labeling can facilitate phonetic learning in infancy. *Child development* **85**, 1036–1049 (2014).
104. C Bergmann, L Ten Bosch, P Fikkert, L Boves, A computational model to investigate assumptions in the headturn preference procedure. *Front. psychology* **4**, 676 (2013).
105. CA Thorburn, NH Feldman, T Schatz, A quantitative model of the language familiarity effect in infancy in *Proc. CCN.* (2019).
106. JL Schwartz, LJ Boë, N Vallée, C Abry, The dispersion-focalization theory of vowel systems. *J. phonetics* **25**, 255–286 (1997).
107. SA Zahorian, AJ Jagharghi, Spectral-shape features versus formants as acoustic correlates for vowels. *The J. Acoust. Soc. Am.* **94**, 1966–1982 (1993).
108. M Ito, J Tsuchida, M Yano, On the effectiveness of whole spectral shape for vowel perception. *The J. Acoust. Soc. Am.* **110**, 1141–1149 (2001).
109. MR Molis, Evaluating models of vowel perception. *The J. Acoust. Soc. Am.* **111**, 2433–2434 (2005).
110. JM Hillenbrand, RA Houde, RT Gayvert, Speech perception based on spectral peaks versus spectral shape. *The J. Acoust. Soc. Am.* **119**, 4041–4054 (2006).
111. Y Matusevych, T Schatz, H Kamper, NH Feldman, S Goldwater, Evaluating computational 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 segmentation in *Proc. MWSCAS.* (2001).
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 prediction errors in *Proc. ACL.* (2017).
115. E Hermann, S Goldwater, Multilingual bottleneck features for subword modeling in zero-resource languages in *Proc. Interspeech.* (2018).
116. CG Hempel, P Oppenheim, Studies in the logic of explanation. *Philos. science* **15**, 135–175 (1948).
117. M VanDam, et al., Homebank: An online repository of daylong child-centered audio recordings in *Seminars in speech and language.* (2016).
118. MC Frank, et al., A collaborative approach to infant research: Promoting reproducibility, best practices, and theory-building. *Infancy* **22**, 421–435 (2017).
119. C Bergmann, et al., Promoting replicability in developmental research through meta-analyses: 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).
121. J Chang, JW Fisher III, Parallel sampling of dp mixture models using sub-cluster splits in *Proc. NEURIPS.* (2013).
122. TK Vintsyuk, Speech discrimination by dynamic programming. *Cybern. Syst. Analysis* **4**, 52–57 (1968).