Evaluating computational models of infant phonetic learning across languages

Yevgen Matusevych (yevgen.matusevych@ed.ac.uk)
School of Informatics, University of Edinburgh

Thomas Schatz (tschatz@umd.edu)
Department of Linguistics & UMIACS, University of Maryland

Herman Kamper (kamperh@sun.ac.za)
E&E Engineering, Stellenbosch University

Naomi H. Feldman (nhf@umd.edu)
Department of Linguistics & UMIACS, University of Maryland

Sharon Goldwater (sgwater@inf.ed.ac.uk)
School of Informatics, University of Edinburgh

Abstract

In the first year of life, infants' speech perception becomes attuned to the sounds of their native language. Many accounts of this early phonetic learning exist, but computational models predicting the attunement patterns observed in infants from the speech input they hear have been lacking. A recent study presented the first such model, drawing on algorithms proposed for unsupervised learning from naturalistic speech, and tested it on a single phone contrast. Here we study five such algorithms, selected for their potential cognitive relevance. We simulate phonetic learning with each algorithm and perform tests on three phone contrasts from different languages, comparing the results to infants' discrimination patterns. The five models display varying degrees of agreement with empirical observations, showing that our approach can help decide between candidate mechanisms for early phonetic learning, and providing insight into which aspects of the models are critical for capturing infants' perceptual development.

Keywords: early phonetic learning; representation learning; phone discrimination; computational model

Introduction

Infants' speech perception changes in the first year of their life. For example, at the age of 6–8 months, English-learning and Japanese-learning infants are equally able to detect the difference between sounds [ɹ] (as in *rock*) and [l] (as in *lock*), whereas by the age of 10–12 months, the two groups diverge, showing attunement to the phonetic contrasts present in their input language (Kuhl et al., 2006). Similar results have been reported for many other languages (Werker & Tees, 1984; Bosch & Sebastián-Gallés, 2003; Tsao, Liu, & Kuhl, 2006, etc.). A number of theoretical accounts explaining such early phonetic learning have been proposed (e.g., Kuhl & Iverson, 1995; Best, 1994; Werker & Curtin, 2005), but until recently no computational models could explain how the specific speech input to which infants are exposed leads to the observed changes in those infants' discrimination of phonetic contrasts. Models *have* been evaluated on their ability to learn phonetic categories, rather than to predict patterns of discrimination (Vallabha et al., 2007; McMurray, Aslin, & Toscano, 2009), but none has yet succeeded in that task when learning from non-idealized natural speech input (Bion et al., 2013; Antetomaso et al., 2017).

In a recent study, Schatz et al. (2019) were the first to present such a computational model, which correctly predicted the documented cross-linguistic difference in infants' discrimination of [ɹ] and [l] after learning from unsegmented speech. They explicitly simulated the learning process for Japanese and American English infants by (separately) training their model on naturalistic speech recordings either in Japanese or in American English. They then measured the trained models' ability to discriminate [ɹ] and [l] with the machine ABX task, a flexible measure of discrimination that can be applied to model representations in essentially any format. To obtain a model capable of handling realistic input, they selected an algorithm for unsupervised learning from naturalistic speech that had been proposed in the context of engineering applications. The success of their model raises the question of whether other learning algorithms recently proposed in this context may lead to equally good, or even better, models of infants' early phonetic learning.

In this paper we apply the approach introduced in Schatz et al. (2019) to test a variety of models on multiple data sets of infant phone discrimination. Doing so allows us to gain insight into the kinds of representations and learning mechanisms that infants are likely to employ. We consider five different learning algorithms developed in the speech technology community, focusing on those that implement cognitively plausible learning mechanisms. We test the models on three crosslinguistic phone discrimination tasks grounded in infant studies from different languages. Our two goals are (1) to test whether Schatz et al.'s result is specific to their particular model applied to a particular American English phonetic contrast, and (2) to identify which computational models can best explain the existing infant data. We find that two models show infant-like crosslinguistic discrimination patterns for American English [ɹ]–[l] and another contrast in Mandarin Chinese, while three other models appear less successful as models of early phonetic learning. We perform additional

analyses to assess whether the two most successful models—which substantially differ in their learning mechanisms and representation formats—also differ in what they learn. The results suggest that it should be possible to find an empirical test to decide between these two models.

Method

In a series of simulations, we train computational models on unsegmented speech data from different languages. Each group of simulations focuses on one phonetic contrast (such as American English [ɹ]–[l]) for which cross-linguistic phone discrimination data for infants exist. We test five different models on each contrast. For each model, we train two different versions: a 'native' model, which simulates a learner of the language from which the contrast is drawn (American English, in this example), and a 'non-native' model, which simulates a learner of another language that does not contain the relevant contrast (here, Japanese). Models are trained on corpora of natural speech, to simulate learning 'in the wild'. We then test each model by simulating a phone discrimination task, using real examples from the language where the contrast exists. To show an infant-like pattern, the 'native' trained version of the model should display better discrimination than the 'non-native' trained version of the model.

The training and test conditions for the three simulations are summarized in Table 1. The contrasts are chosen based on the strength of evidence regarding infants' behavior on this contrast and the availability of corresponding speech corpora for training computational models. **Simulation 1** tests models learning American English and Japanese on the English [ɹ]–[l] contrast, where English learners show better discrimination than Japanese learners by 12 months (Kuhl et al., 2006; Tsushima et al., 1994). **Simulation 2** tests models learning Mandarin Chinese and American English on the Mandarin [ɕ]–[tɕʰ] contrast, where Mandarin learners show better discrimination than English learners by 12 months (Tsao et al., 2006; Kuhl, Tsao, & Liu, 2003). **Simulation 3** tests Catalan- and Spanish-learning models on the Catalan [e]–[ɛ] contrast, where Catalan learners show better discrimination than Spanish learners by 8 months (Bosch & Sebastián-Gallés, 2003; Albareda-Castellot, Pons, & Sebastián-Gallés, 2011). In the experiments with infants, each phonetic contrast was tested in a particular phonetic context (e.g., [ɹa]–[la]). To have sufficient test data for our models, we report the results averaged over all phonetic contexts instead. However, we also tested the models in the restricted contexts that were actually used in the experiments, and the trends were in the same direction.

Simulating phone discrimination tasks

To test a model's ability to discriminate a phonetic contrast, similar to the tests carried out with infants such as conditioned head turn (Werker, Polka, & Pegg, 1997), we use the machine ABX task (Schatz et al., 2013).[1] In this task, A and X are two instances of the same phone (e.g., [l]), while B is a different phone (e.g., [ɹ]). If A and X are closer to each other in a model's representation space than B and X, the model's prediction is correct, otherwise it is not.[2] A model is evaluated by considering the proportion of ABX triplets for which it makes correct predictions: 0% error rate corresponds to perfect discrimination, and 50% to chance performance. Following Schatz et al. (2019), we sample ABX test triplets in such a way that all three phones—A, B, and X—appear in the same neighboring phonetic context and are pronounced by the same speaker.

Table 1: Training and test conditions.

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

To test whether the difference between the ABX error rates in a given pair of simulated listeners (native vs. non-native) is significant, we fit mixed-effects regressions to the error rates of the two models in question. Each regression includes main effects of simulated listener type (native vs. non-native) and data subset (as shown in Table 3 below) and random intercepts, to account for the variation among data subsets, speakers and phonetic contexts. Significance for the effect of simulated listener type was then determined using two-tailed ANOVA tests (with Satterthwaite degrees of freedom approximation) on the predicted values of the regressions.

Computational models

We consider five models: the one used in Schatz et al. (2019) and four neural network models inspired by existing work in unsupervised speech representation learning. These models show high performance in low-resource speech technology applications, making them a good starting point for modeling unsupervised infant learning. In high-level terms, the models differ along several dimensions, as summarized in Table 2. Three of the models learn representations at the level of speech frames (i.e., 25-ms-long chunks of speech commonly used in automatic speech recognition), while two learn to encode word-sized units of variable length as vector representations of fixed length (somewhat similar to semantic word embeddings). Note, however, that at test time all models provide a way to compute distances between speech sequences (in this case,

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

[2] Following earlier studies, we use Kullback–Leibler divergence to measure distances in the representations for one of the models (DPGMM, see Table 2), and angular cosine distance for the other models. For models using frame-level representations (see Table 2), we align the frames in each pair of phones using dynamic time warping (Vintsyuk, 1968), and compute the distance as an average over the framewise distances.

phones) of any duration. Finally, three models are strictly unsupervised, while two others rely on top-down guidance from known wordforms—based on evidence that even 6–8-month-olds can segment and recognize some wordforms (Bortfeld et al., 2005; Jusczyk & Aslin, 1995; Jusczyk, Houston, & Newsome, 1999). In all cases, we use existing implementations that were developed for processing speech with minimal supervision (Schatz et al., 2019; Kamper et al., 2015; Kamper, 2019), adopt the previously used training options and do not retune hyperparameters.

Dirichlet process Gaussian mixture model (DPGMM, Chen et al., 2015) with parallel MCMC sampling (Chang & Fisher III, 2013) is the model used by Schatz et al. (2019). It is a probabilistic generative model with the number of Gaussian components (clusters) derived from the data. It learns in a fully unsupervised bottom-up manner, by soft-clustering individual speech frames. Using the learned mixture, each frame in the test data is represented as a posterior probability vector, consisting of the probabilities of this frame being generated by each component. We use the implementation based on Chang & Fisher III (2013)[3] and parameter settings from Schatz et al. (2019): the model is initialized with 10 clusters and is trained for 1501 iterations.

Autoencoder (AE, Kramer, 1991) is an unsupervised feedforward neural network. It learns by reconstructing (i.e., encoding and decoding) a given input, which results in the emergence of latent representations in its hidden layers. In this case, the model reconstructs individual speech frames one-by-one. We use a deep implementation with 8 hidden layers (7×100 and 1×39 units), as described in Kamper et al. (2015),[4] and pretrain it for 5 epochs per layer plus 5 epochs of final fine-tuning, without early stopping, using Adadelta optimization with adaptive learning rate (Zeiler, 2012) with decay 0.95. At test time, the 39-dimensional second-to-last hidden layer is used to encode individual frames from the test data into the model's representation space.

Correspondence autoencoder (CAE, Kamper et al., 2015) is a deep neural network similar to the AE, but it uses weak word-level supervision. Instead of trying to encode and reconstruct each input frame to itself, as is done in the AE, it is given pairs of spoken instances of the same word (aligned at the frame level), and tries to reconstruct each frame in one instance of a word from the aligned frame in the other instance—so the encoded representation must focus on linguistically meaningful information and abstract away from other variation between the aligned frames. Following Kamper et al. (2015), we initialize the CAE using the AE, and train the CAE with the same architecture as in the AE for 120 epochs.[4]

Autoencoding recurrent neural network (AE-RNN, Chung et al., 2016), or sequence-to-sequence autoencoder, is a type of AE in which both the encoder and the decoder are recurrent neural networks (RNNs). RNNs are commonly used in language modeling (Linzen, 2019), as they can process an input sequence as a whole. In our case, the model is given a random word-sized chunk of speech, encodes it into a vector of fixed dimensionality, and then uses this vector to reconstruct the same chunk sequentially, frame-by-frame. We use the implementation by Kamper (2019),[5] with 3 hidden layers (400 gated recurrent units each) in both the decoder and the encoder, and an embedding dimensionality of 130. We train it for 15 epochs without early stopping using Adam optimization (Kingma & Ba, 2015) with a learning rate of 0.001, and then use its encoder to convert each test chunk into a fixed-dimensional representation.

Table 2: Models of early phonetic learning used in the study.

Model	Top-down guidance	Representation type
DPGMM	No	Frames
AE	No	Frames
CAE	Yes	Frames
AE-RNN	No	Word-sized
CAE-RNN	Yes	Word-sized

Correspondence-autoencoding recurrent neural network (CAE-RNN, Kamper, 2019) is similar to the AE-RNN, but instead of training on random chunks of speech, it is trained on pairs of instances of the same word—i.e., like the CAE, it also relies on weak top-down supervision. Again, following previous work (Kamper, 2019),[5] we use the AE-RNN to initialize the parameters of the CAE-RNN and then train it (with parameters analogous to those of the AE-RNN) for 3 epochs.

Input to the models

To prepare input to the models from unsegmented speech data, we follow a standard approach in speech processing, also adopted by Schatz et al. (2019): we discretize the speech data into 25-ms-long frames (sampled every 10 ms) and extract mel-frequency cepstral coefficients (MFCCs), together with their first and second time derivatives, from each frame using Kaldi (Povey et al., 2011). Representing speech using its auditory spectrum—as MFCCs do—is grounded in human auditory processing and is different from traditional accounts of phonetic learning, which assume phonetic feature detectors (see Schatz et al., 2019, for further discussion).

Additionally, we need to obtain speech alignments, i.e., labels that map chunks of speech to their corresponding phones or words. For training the models with top-down guidance, we obtain word-level alignments, and for testing all the models, we obtain phone-level alignments, using the Montreal Forced Aligner (McAuliffe et al., 2017). When running the alignments, we noticed that the Catalan dictionary did not always contain the correct entries for our target Catalan contrast, [e] vs. [ɛ], and we replaced such entries with standard Catalan pronunciations available in Wiktionary[6]. For Simu-

[3] http://people.csail.mit.edu/jchang7/code/subclusters/dpmm_subclusters_2014-08-06.zip

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

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

[6] http://wiktionary.org

Table 3: Corpus samples used in the simulations.

A. Training data.

Sim. #	Language	Corpus	Register	Amount of data	No. of spk.
1	EN	WSJ[1]	Rd[7]	19:30	96
	JA	GlobalPhone[2]	Rd	19:33	96
	EN	Buckeye[3]	Sp[7]	9:13	20
	JA	CSJ[4]	Sp	9:11	20
2	ZH	AIShell[5]	Rd	58:59	166
	EN	WSJ	Rd	58:49	166
	ZH	GlobalPhone	Rd	11:51	48
	EN	WSJ	Rd	11:49	48
3	CA	Glissando[6]	Rd+Sp	7:41	26
	ES	Glissando	Rd+Sp	7:41	26
	CA	Glissando	Rd+Sp	7:02	17
	ES	Glissando	Rd+Sp	7:03	17

B. Test data.

1	EN	WSJ	Rd	9:39	47
		Buckeye	Sp	9:01	20
2	ZH	AIShell	Rd	58:45	165
		GlobalPhone	Rd	11:51	48
3	CA	Glissando	Rd+Sp	1:15	2
		Glissando	Sp	2:19	11

[1] The Wall Street Journal CSR corpus (Paul & Baker, 1992).
[2] The multilingual text and speech database (Schultz, 2002).
[3] The Buckeye corpus of conversational speech (Pitt et al., 2005).
[4] Corpus of spontaneous Japanese (Maekawa, 2003).
[5] An open-source Mandarin speech corpus (Bu et al., 2017).
[6] A corpus for multidisciplinary prosodic studies in Spanish and Catalan (Garrido et al., 2013).
[7] Rd and Sp stand for read and spontaneous speech, respectively.

lation 1, we obtained the existing alignments (Schatz et al., 2019) generated with Kaldi (Povey et al., 2011).

In each case we train and test models on two different subsets of speech data per language, in order to ensure that the results for each model are robust across various data sets. Ideally, each subset should come from a different corpus, and the corpora should represent two different speech registers: spontaneous and read. In practice, our choices are limited to the available speech corpora, so that in Simulation 2 we use corpora of read speech only, and in Simulation 3 all our data come from the same bilingual corpus (Table 3). To further reduce potential variability across corpora, we sample the audio signal in each corpus at 16 kHz and balance the speakers' gender within each corpus sample.

Results

Crosslinguistic patterns of ABX discrimination

The ABX error rates of each model across languages are shown in Figure 1, together with the average performance of each model across all consonant (for English and Mandarin Chinese) or vowel (for Catalan) contrasts (red lines in the figure). In this figure, results are averaged over multiple ABX triplets,

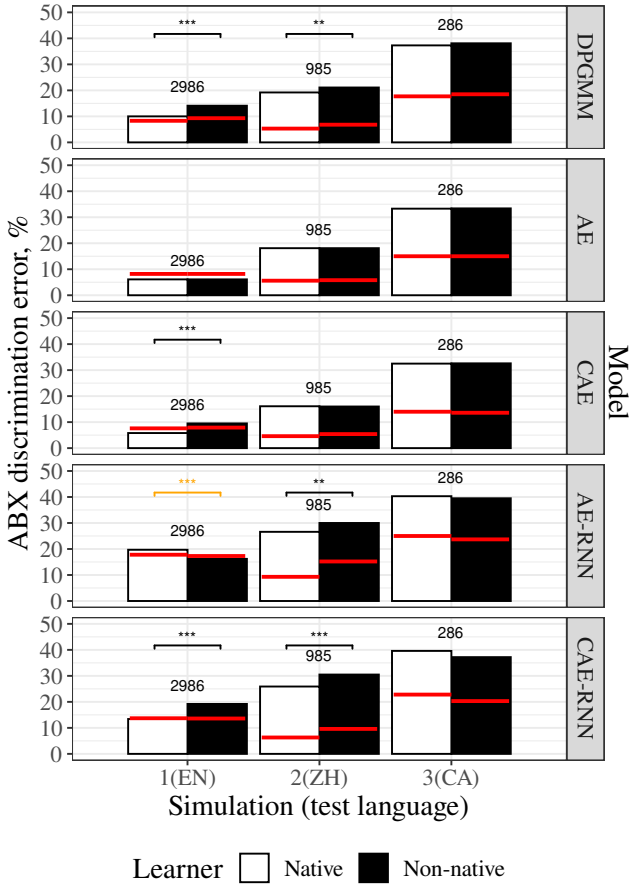

Figure 1: ABX error rates of the five native and non-native models in the three discrimination tasks (EN [ɹ]–[l], ZH [ɕ]–[tɕʰ] and CA [e]–[ɛ]). To match the infant pattern of discrimination, the native model in each pair must show significantly lower error rates than the non-native model: out of 7 patterns with a significant difference, 6 are in the predicted direction (black brackets) and 1 is in the wrong direction (orange bracket). The number of data pairs in each test set is shown on top of each bar. Red lines indicate models' error rates averaged over all consonant (for EN and ZH) or all vowel (for CA) contrasts.

speakers, neighboring phonetic contexts, and subsets of the corpus, although the significance values take into account all of these variables (as previously discussed). The reported patterns are consistent over the two corpus subsets in all cases, except the AE-RNN on the Mandarin contrast (where the difference between the 'native' and the 'non-native' model does not reach significance when trained on the smaller subset). In what follows, we compare the performance of each simulated 'native' listener to its corresponding simulated 'non-native' listener. Note that comparing the results *across* the simulations is not meaningful, as the amount of training data and number of speakers differed depending on the simulation.

In Simulation 1, which tests the models on the English [ɹ]–[l] contrast, three out of five models (DPGMM, CAE and CAE-RNN) correctly predict the discrimination pattern ob-

served in infants: the error rate of the simulated native listener is significantly lower compared to the simulated non-native listener. Two models do not predict this pattern correctly: the AE shows no significant difference between the two types of simulated learners, and the AE-RNN predicts a significant difference in the wrong direction (i.e., lower error rate in the non-native listener). Across all models (native and non-native), the discrimination error on this contrast ranges from 5.8–19.7%, comparable to the average error on all English consonant contrasts (red lines, 7.6–17.8%). In other words, the [ɪ]–[l] contrast is easy to discriminate for all models.

In Simulation 2, with the Mandarin [ɕ]–[tɕʰ] contrast, a different set of three models (DPGMM, *AE-RNN* and CAE-RNN) correctly predict the infants' discrimination pattern, and two models (AE and CAE) predict no significant difference between the simulated native and the non-native listener. Over all models, the discrimination error on this contrast ranges from 16.0–30.5%—noticeably higher than the average error over all Mandarin consonant contrasts (red lines, 4.6–9.6%). In other words, the [ɕ]–[tɕʰ] discrimination is difficult for the models. This may be due to the kinds of phones in this contrast: one of them, [tɕʰ], is a combination of a short [t] followed by a 'breathy' version of [ɕ], the other phone in the contrast (somewhat similar to the distinction between the first phones in *cheap* vs. *sheep*). As a result, one of the phones, [ɕ], is almost a 'subchunk' of the other phone, [tɕʰ], a distinction potentially difficult to learn for our models. Nevertheless, three 'native' models show lower error rate than the corresponding 'non-native' models.

In Simulation 3, no model predicts a significant difference for the Catalan [e]–[ɛ] contrast. In general, this contrast is more difficult for all models than an average Catalan vowel contrast, with error rates ranging from 32.5–40.3% (vs. the average of 14.0–23.7%). Further analysis (not shown) revealed that this may be because some [e] and [ɛ] vowels in the test data had very short duration (unlike the lab stimuli used with infants). Filtering out very short ($<$ 80 ms) phones from the test data reduced the overall error rates, but still yielded similar performance between the 'native' and 'non-native' models. Note that the models' average error rates on Catalan are generally high, suggesting that the models could benefit from additional training data. At the same time, speaker idiosyncrasies in the test data are unlikely to affect the results, as we observe no meaningful differences in the discrimination error across the two test samples (consisting of data from 2 vs. 11 speakers). Thus, all models struggle to discriminate the Catalan [e]–[ɛ] contrast, as well as to reproduce empirically observed cross-linguistic differences in its discrimination.

While no model captures all three crosslinguistic discrimination patterns, the DPGMM and the CAE-RNN correctly predict two patterns out of three. The CAE and the AE-RNN only predict one pattern each, while the AE makes no correct predictions. The two models that perform best use very different learning algorithms and representation formats, effectively presenting two alternative hypotheses about early phonetic learning, yet they make identical predictions regarding the crosslinguistic discrimination of three phone contrasts. This raises the question of whether these two models, despite their differences, may behave similarly on discrimination tasks in general. We address this question in the next section.

Table 4: Phone contrasts with the largest difference in ABX discrimination between the two models in terms of standard scores (or z-scores: how many standard deviations the error rate of a particular model is higher or lower than the mean error rate of that model). Positive difference indicates that the contrast is easier for the CAE-RNN than for the DPGMM.

		Standard score		
Contrast	Language	DPGMM	CAE-RNN	Difference
d–z	JA	−0.2	−1.0	+0.8
i–iː	JA	−0.4	−1.2	+0.8
a–aː	JA	−0.4	−1.1	+0.7
pʰ–tʰ	ZH	−0.2	−0.8	+0.6
r–z	JA	−0.8	−1.4	+0.6
ɕ–ɕː	JA	+0.2	−0.3	+0.5
h–θ	EN	−1.4	0.0	−1.4
g–ŋ	EN	−0.9	+0.4	−1.3
u–ʊ	EN	−0.3	+0.9	−1.2
f–x	ES	−0.4	+0.6	−1.0
t–θ	ES	−0.7	+0.3	−1.0
t–tː	JA	−0.7	+0.3	−1.0

Comparing predictions of the two models

In this section, we identify phone contrasts for which the two models—DPGMM and CAE-RNN—make different predictions in the discrimination task. For simplicity, we only focus on the 'native' listeners—i.e., trained and tested on the same language. We use 5 simulated listeners trained on different languages: English and Japanese from Simulation 1, Mandarin from Simulation 2, Catalan and Spanish from Simulation 3. We test each model on a variety of contrasts from its training language. Among hundreds of phone contrasts that could be tested, many of them are trivial to discriminate (e.g., a vowel vs. a consonant), and we only consider those where the two phones differ on one distinctive phonological feature: e.g., Japanese [i] and [iː] differ in *length* (i.e., duration), and Mandarin [p] and [pʰ] differ in *aspiration* (i.e., 'breathiness').

Table 4 shows the contrasts for which the discrimination performance of the DPGMM and the CAE-RNN differs the most. As the average error rates (and their distributions) vary across the two models, we report differences in terms of standard scores. One pattern we see is that the easiest contrasts for the CAE-RNN relative to the DPGMM (upper part of Table 4) are those where the two phones are continuous and differ only in their length: Japanese [i]–[iː] and [a]–[aː], Mandarin [ɕ]–[ɕː]. The DPGMM performs best relative to the CAE-RNN (lower part of Table 4) on such contrasts as [t]–[tː] and [u]–[ʊ], where one phone (here, [ʊ] and [t]) is similar to a 'subchunk' of the other phone ([u] and [tː]), so that the difference between the two is only observable within a short

time window: e.g., [tː] is a silence followed by [t] (and see Hillenbrand et al., 1995, regarding the vowel contrast). This may be because the CAE-RNN 'compresses' the time dimension into a fixed-dimensional representation, losing the information about individual speech frames. It is less clear why contrasts involving [θ] are difficult for the CAE-RNN, though Jongman (1989) shows that listeners need to hear the full duration of [θ] to successfully discriminate it from similar sounds.

We also looked at the models' error rates on *types* of contrasts (e.g., all contrasts where the two phones differ on a particular feature), to see whether the discrimination along a particular phonological feature (e.g., aspiration) is easier for one model than for the other. The most salient pattern is that length contrasts (e.g., short [a] vs. long [aː], cf. Table 4) are on average easier to discriminate for the CAE-RNN (compared to other types of contrasts) than for the DPGMM, in line with our analysis for the individual contrasts.

Conclusion

Using computational modeling on realistic input, we compared possible mechanisms for early phonetic learning in their ability to predict the changes in discrimination empirically observed in infants. We tested five models on three phone contrasts from different languages, using a phone discrimination task. In our study, Schatz et al.'s (2019) DPGMM model showed the infant-like crosslinguistic pattern of discrimination both for English [ɹ]–[l] and for Mandarin [ɕ]–[tɕʰ], demonstrating that the earlier result was not specific to a particular English contrast. Moreover, a second model, the CAE-RNN, also made correct predictions on the same two contrasts. Although no model predicted the correct pattern for the Catalan vowel contrast [e]–[ɛ], the fact that two of them make correct predictions on the other two contrasts is promising. This result supports the idea that models learning representations directly from unsegmented natural speech can correctly predict some of the infant phone discrimination data. Based on the results of this study, the DPGMM and the CAE-RNN show some promise as models of early phonetic learning, although their failure to capture the effect on the Catalan contrast warrants further investigation to determine whether the failure is a problem with the models or with the amount and/or quality of the training and the test data.

Here we are making predictions about infants' phone discrimination. Adults show many cross-linguistic discrimination differences as well, and one question for future research is how the presented models of infant speech perception relate to predictions one might make about adult perception. There are likely to be additional learning mechanisms later in childhood beyond those we have modeled here, but what changes over development, and how, remains an open question.

The DPGMM and the CAE-RNN both aim to model primarily unsupervised infant phonetic learning, although they make very different assumptions about the learning mechanism and the types of representations: the DPGMM is an unsupervised bottom-up model that learns by clustering frame-level representations, while the CAE-RNN encodes chunks of speech holistically, learns by sequentially reconstructing an acoustic word, and uses weak top-down guidance from the word level (but see Kamper, 2019, for a fully unsupervised alternative). Importantly, our analysis shows that the two models do not always make identical predictions about discriminability: for example, vowel length contrasts are generally easier to discriminate for the CAE-RNN than for the DPGMM. Such differences between the models allow us to make predictions—both for native and non-native listeners—which can then be tested in experimental studies with infants to differentiate between the models.

Acknowledgements

This work is based on research supported in part by an NSF award BCS-1734245 and an ESRC-SBE award ES/R006660/1. We thank the anonymous reviewers, the members of Maryland Phonology Circle, and Kate McCurdy for helpful feedback.

References

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

Antetomaso, S., Miyazawa, K., Feldman, N., Elsner, M., Hitczenko, K., & Mazuka, R. (2017). Modeling phonetic category learning from natural acoustic data. In *Proceedings of the BUCLD.*

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

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

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

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

Bu, H., Du, J., Na, X., Wu, B., & Zheng, H. (2017). AISHELL-1: An open-source Mandarin speech corpus and a speech recognition baseline. In *Proceedings of O-COCOSDA.*

Chang, J., & Fisher III, J. W. (2013). Parallel sampling of DP mixture models using sub-cluster splits. In *Advances in Neural Information Processing Systems.*

Chen, H., Leung, C.-C., Xie, L., Ma, B., & Li, H. (2015). Parallel inference of Dirichlet process Gaussian mixture models for unsupervised acoustic modeling: A feasibility study. In *Proceedings of Interspeech.*

Chung, Y.-A., Wu, C.-C., Shen, C.-H., Lee, H.-Y., & Lee, L.-S. (2016). Audio word2vec: Unsupervised learning of audio segment representations using sequence-to-sequence autoencoder. In *Proceedings of Interspeech.*

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

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *The Journal of the Acoustical Society of America*, *97*, 3099–3111.

Jongman, A. (1989). Duration of frication noise required for identification of English fricatives. *The Journal of the Acoustical Society of America*, *85*, 1718–1725.

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

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

Kamper, H. (2019). Truly unsupervised acoustic word embeddings using weak top-down constraints in encoder-decoder models. In *Proceedings of ICASSP.*

Kamper, H., Elsner, M., Jansen, A., & Goldwater, S. (2015). Unsupervised neural network based feature extraction using weak top-down constraints. In *Proceedings of ICASSP.*

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In *Proceedings of ICLR.*

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

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

Kuhl, P. K., Stevens, E., Hayashi, A., Deguchi, T., Kiritani, S., & Iverson, P. (2006). Infants show a facilitation effect for native language phonetic perception between 6 and 12 months. *Developmental Science*, *9*, F13–F21.

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

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

Maekawa, K. (2003). Corpus of Spontaneous Japanese: Its design and evaluation. In *Proceedings of SSPR Workshop.*

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal Forced Aligner: Trainable text-speech alignment using Kaldi. In *Procedings of Interspeech.*

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

Paul, D. B., & Baker, J. M. (1992). The design for the Wall Street Journal-based CSR corpus. In *Proceedings of the Workshop on Speech and Natural Language.*

Pitt, M. A., Johnson, K., Hume, E., Kiesling, S., & Raymond, W. (2005). The Buckeye corpus of conversational speech: Labeling conventions and a test of transcriber reliability. *Speech Communication*, *45*, 89–95.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., ... Veselý, K. (2011). The Kaldi speech recognition toolkit. In *Proceedings of ASRU Workshop.*

Schatz, T., Feldman, N. H., Goldwater, S., Cao, X.-N., & Dupoux, E. (2019). Early phonetic learning without phonetic categories—insights from large-scale simulations on realistic input. *PsyArXiv*.

Schatz, T., Peddinti, V., Bach, F., Jansen, A., Hermansky, H., & Dupoux, E. (2013). Evaluating speech features with the minimal-pair ABX task: Analysis of the classical MFC/PLP pipeline. In *Proceedings of Interspeech.*

Schultz, T. (2002). GlobalPhone: a multilingual speech and text database developed at Karlsruhe University. In *Proceedings of ICSLP.*

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

Tsushima, T., Takizawa, O., Sasaki, M., Shiraki, S., Nishi, K., Kohno, M., ... Best, C. (1994). Discrimination of English /r-l/ and /w-y/ by Japanese infants at 6-12 months: Language-specific developmental changes in speech perception abilities. In *Proceedings of ICSLP.*

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

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

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

Werker, J. F., Polka, L., & Pegg, J. E. (1997). The conditioned head turn procedure as a method for testing infant speech perception. *Infant and Child Development*, *6*, 171–178.

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

Zeiler, M. D. (2012). ADADELTA: An adaptive learning rate method. *arXiv*.