Input matters in the modeling of early phonetic learning

Ruolan Li (rlli@umd.edu)
Program in Neuroscience and Cognitive Science, Department of Linguistics, & UMIACS, University of Maryland
College Park, MD 20742, USA

Thomas Schatz (tschatz@umd.edu)
Department of Linguistics & UMIACS, University of Maryland
College Park, MD 20742, USA

Yevgen Matusevych (yevgen.matusevych@ed.ac.uk)
School of Informatics, University of Edinburgh, 10 Crichton St
Edinburgh EH8 9AB, UK

Sharon Goldwater (sgwater@inf.ed.ac.uk)
School of Informatics, University of Edinburgh, 10 Crichton St
Edinburgh EH8 9AB, UK

Naomi H. Feldman (nhf@umd.edu)
Department of Linguistics & UMIACS, University of Maryland
College Park, MD 20742, USA

Abstract

In acquiring language, differences in input can greatly affect learning outcomes, but which aspects of language learning are most sensitive to input variations, and which are robust, remains debated. A recent modeling study successfully reproduced a phenomenon empirically observed in early phonetic learning—learning about the sounds of the native language in the first year of life—despite using input that differed in quantity and speaker composition from what a typical infant would hear. In this paper, we carry out a direct test of that model's robustness to input variations. We find that, despite what the original result suggested, the learning outcomes are sensitive to properties of the input and that more plausible input leads to a better fit with empirical observations. This has implications for understanding early phonetic learning in infants and underscores the importance of using realistic input in models of language acquisition.

Keywords: early phonetic learning, computational modeling, input variation, speech perception

Introduction

Early experience with language can vary widely across children. The consequences of these variations for language acquisition remain under debate. Variability in the linguistic input that children are exposed to has been associated with differences in learning outcomes in some cases (Hart & Risley, 1995; Cristia, 2011; Hoff, 2006; Song, Demuth, & Morgan, 2018), but learning has been argued to be robust to input variations in others (Bergmann & Cristia, 2018; Cristia, 2018).

A recent modeling study by Schatz, Feldman, Goldwater, Cao, and Dupoux (2019) appears to support the idea that a specific aspect of language learning—early phonetic learning—is robust across a range of input variations. Schatz et al. focused on the finding that English-learning infants become better at discriminating [l]-[ɹ], as in "lock" and "rock," but Japanese infants—whose native language does not distinguish between these two sounds—do not (Kuhl et al., 2006; Werker & Tees, 1984). They used a distributional learning algorithm (Maye, Werker, & Gerken, 2002) operating on low-level auditory representations of unsegmented speech, and showed that models trained on English performed significantly better on discriminating the [l]-[ɹ] contrast than models trained on Japanese. This was true even though the training data consisted of a roughly even mix of speech from more than 20 speakers that was balanced across gender, and even when the model was trained on as little as one hour of speech. Those training characteristics do not match input conditions expected for a typical infant in the behavioral literature in terms of either the number of speakers (Bergmann & Cristia, 2018), the duration of input, or the composition of the input by speaker and by speaker gender (Bergelson et al., 2019). The model's successful prediction of infants' discrimination of [l] and [ɹ] despite this atypical input raises the possibility that early phonetic learning could be robustly supported across a wide range of input conditions.

In this paper, we directly examine the sensitivity of Schatz et al.'s (2019) learning model to variations in input, and find that while the model is robust to some variations in input, it is highly sensitive to others. Specifically, we look at the effect of varying the duration, speaker composition, and speaker gender of the input. To obtain enough speech data to manipulate these input parameters, we assemble a corpus based on public-domain audiobooks, which provide abundant data separable by speaker (up to 900 hours per speaker). The large size of public-domain audiobooks allows us to control the speaker distribution or select a single speaker to train the model, which makes it an ideal resource to answer questions about how the distribution of speakers in a model's input affects what the model learns. We find that differences in input affect the model learning outcome substantially, and that training on more plausible input leads to predicted learning outcomes that better fit empirical observations. Our study suggests that if the

model from Schatz et al. (2019) is an accurate model of early phonetic learning, then early phonetic learning may be more sensitive to input conditions than had been previously realized. More broadly, our results underscore the importance of critically assessing input assumptions when modeling language acquisition.

Methods

We simulate the phonetic learning process from corpora of unsegmented raw speech and examine the ability of the resulting model representations to discriminate between pairs of speech sounds in the 'native' (i.e. training) language or in a 'foreign' language. To examine the learning outcomes predicted by our models under different input data, we use two kinds of input to train speech learning models. The first type of input consists of speech from 10+ individuals, similar to the input in Schatz et al. (2019); the second type of input comes from an individual speaker, such that each model learns from one speaker's speech. We focus on whether the amount and speaker gender of the input speech affects the learning outcome, using the cross-linguistic difference between models' discrimination performance as the measure of learning outcomes. Specifically, we train models on English and Japanese, representing infants learning English or Japanese as their native languages. Then, we test them on English contrasts. Based on the empirical results from Kuhl et al. (2006), we expect English models to perform better as they are trained on more data[1], but do not expect Japanese models to improve. We also expect the difference between English and Japanese performance to increase as the models are trained on more data.

Corpus construction

We assembled an audiobooks corpus containing read speech in English and Japanese using (untranscribed) speech recordings from LibriVox (`https://librivox.org`) and the Internet Archive (`https://archive.org/`), two public domain media repositories. For each language, the corpus contains speech from 6 males and 5 females. The total duration of the corpus is more than 2000 hours of speech, and the duration for each speaker ranges between 2.9 and 918 hours. Code to assemble the corpus can be found at `https://osf.io/9fnxa/`.

Input representation

Following Schatz et al. (2019), we use Mel-frequency cepstral coefficients (MFCCs) as the input representation to our models (Mermelstein, 1976). They take the form of 13-dimensional descriptors of the short-term auditory spectrum of the speech input, with added first and second derivatives. They are obtained at regular intervals of 10ms along the speech stream and are computed over 25ms-long (overlapping) stretches of signal. See Schatz et al. (2019) for further motivation for using MFCCs as input to models of early phonetic learning.

[1] We do not assume an exact mapping between the model training duration and the actual number of hours of speech that an infant hears, but we do assume that the increase in training duration reflects the learning trajectory as infants have access to more speech data.

Learning model

Following Schatz et al. (2019), we train Dirichlet process Gaussian mixture models on the input MFCCs. This generative model assumes that each datapoint was generated from a multidimensional Gaussian cluster, and learns the number of clusters as well as the parameters (weight, mean and covariance matrix) of each cluster. The means and covariance matrices are assumed to be generated from a normal-inverse-Wishart prior. We set the hyperparameters in the Dirichlet process as follows: α to 1, μ_0 and λ_0 to the average and inverse of the covariance of all input MFCCs, λ to 1, and ν to 42. The Dirichlet process (Ferguson, 1973) allows the model to learn the number of Gaussian clusters from the data.

The model is trained through an efficient parallel MCMC sampling scheme (Chang & Fisher III, 2013). Each model is trained for 1501 iterations, sampling through every datapoint in the training data in each iteration. From the number of Gaussian categories and the test error rate, it appears that most models converge before the training is finished: the number of categories usually stabilizes between 100 and 1000 iterations, and the error rate in testing also stabilizes at about the same point in training. The two models (out of 84) that failed to converge according to these metrics are excluded from the analysis and results.

Discrimination test

After training, we extract the models' representations of the test stimuli, which is transformed into MFCCs like the training data. We take the posterior probability distributions over the learned Gaussian clusters as the models' representations at each time point of the speech stimuli. The posterior probability distributions are calculated based on Bayes' Rule, where G_n denotes the nth Gaussian category, and y_t denotes the datapoint at time t,

$$p(G_n \mid y_t) \propto p(y_t \mid G_n)p(G_n) \tag{1}$$

As $p(G_n \mid y)$ is calculated for every Gaussian category, the posterior of each datapoint is a vector $[p(G_1 \mid y), p(G_1 \mid y), ..., p(G_N \mid y)]$, where N is the number of Gaussians.

Model testing uses the machine ABX task (Schatz et al., 2013; Schatz, 2016). This task evaluates a model's ability to discriminate two sounds. Specifically, it takes occurrences[2] of two sound categories in an annotated and time-aligned speech corpus, and compares triplets of three tokens at a time (A, B, and X, where A and X are taken from the same phonetic category, e.g., [ɪ], and B is taken from the other category, e.g., [l]) to see whether the model correctly predicts a greater similarity between A and X, or incorrectly predicts a greater similarity between B and X. The similarity for each token pair is calculated as the average KL-divergence between the two representations aligned by dynamic time warping. KL-divergence is a measure of the difference between two probability distributions, and the dynamic time warping algorithm allows representations of two speech segments of different temporal

[2] Occurrences are controlled by phonetic context and subsampled, in the same way as Schatz et al. (2019).

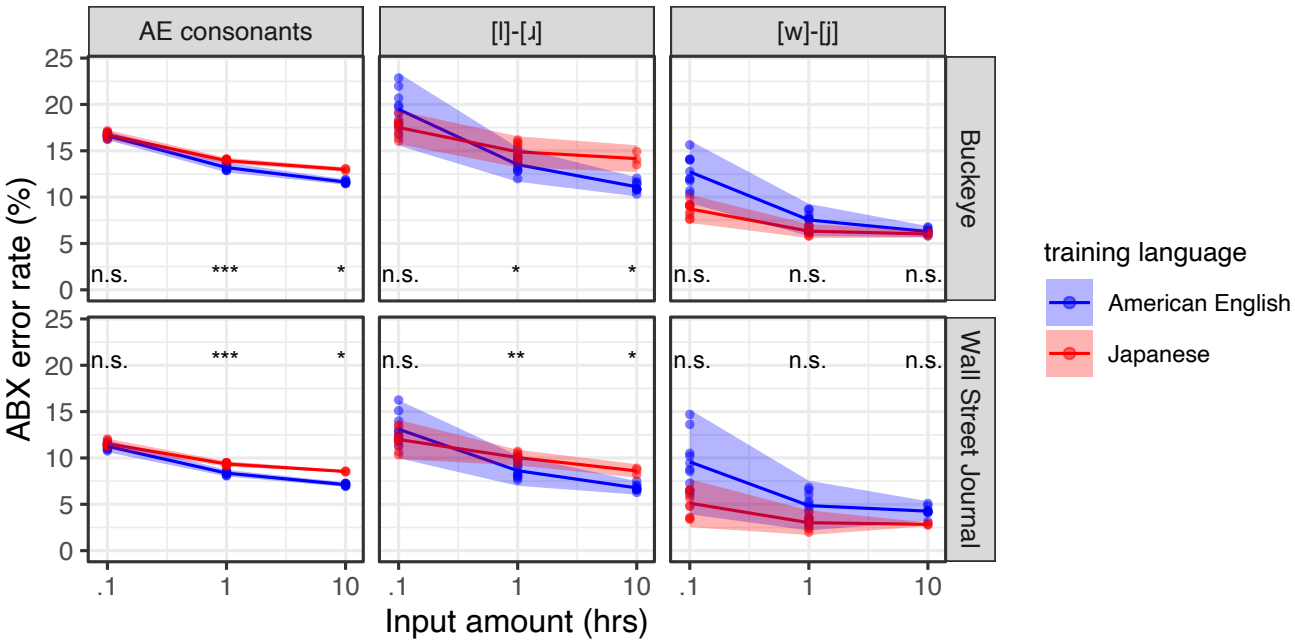

Figure 1: The effect of input amount on ABX discrimination errors, for models trained with multiple speakers. For each duration, 10 models are trained. Each dot represents a model; the lines represent means of the ten models, and the error bands represent 95% confidence intervals. The two rows represent the two testing corpora. Statistical significance is obtained from permutation tests, in which the training language is permuted 10,000 times for each training duration, test corpus and consonant pair. One-tailed tests are performed against the null hypothesis that English models do not have smaller error rates than Japanese models. Significance is assessed with Benjamini-Yekutieli correction for multiple comparisons (Benjamini & Yekutieli, 2001). One, two and three stars indicate $p < 0.05$, $p < 0.01$ and $p < 0.001$ respectively, n.s. stands for non-significant.

lengths to be compared (Vintsyuk, 1968). After the model's choice (correct vs. incorrect) is obtained via the method above, we collapse across all triplets and all environments (phonetic neighbors) to obtain an overall error rate for the pair of sound categories, e.g., [l]-[ɹ]. We denote this measure as the ABX error rate. The lower the ABX error rate is, the better the model is able to discriminate between the target sound pair, where chance performance is 50%.

Training and Testing

For each speaker, models are trained on speech that varies in duration along a log 10 scale, starting at 6 minutes (6 minutes, 1 hour, 10 hours, 100 hours, etc.). Up to 10 models are trained on each training duration when enough speech data are available. Additionally, one model is trained on the full duration of the speech data (e.g., for the speaker with 8.5 hours of data, in addition to the 8 models that are trained on 1 hour of speech, one model is trained on all 8.5 hours of speech). This is to ensure that we have models trained on the full range of input durations without sacrificing statistical power for shorter durations.

The testing is carried out on the Buckeye Corpus (Pitt, Johnson, Hume, Kiesling, & Raymond, 2005), which consists of conversational, spontaneous speech, and the Wall Street Journal Corpus (Paul & Baker, 1992), which consists of read speech. The ABX error rate is obtained for each test speaker, and then averaged in the main results. Following Schatz et al. (2019), we look at [l]-[ɹ] and two controls: [w]-[j], which is contrastive in both Japanese and American English, and the average over all American English consonants. While [l]-[ɹ] is critical for the observation of perceptual attunement, the average across American English consonants offers an overview of how the model performs in general on the ABX task; and the [w]-[j] pair is used as a control pair, which the model is expected to discriminate well regardless of its training language. For [l]-[ɹ], if the gap between the English and Japanese models' performance increases as the models are trained on more data, it would suggest that the models are showing cross-linguistic differences in discrimination that result from exposure to speech input in their native language, and in this way behaving like infants.

Results

Models trained on multiple speakers

We first replicate Schatz et al. (2019) in order to ensure that possible confounds in our audiobooks, such as lower-quality microphones or uncontrolled recording environments, do not impact the results. Following their gender-balanced input distribution of speakers with a roughly equal amount of speech per speaker, we look at the results of models trained on an equal distribution of 11 speakers.

The replication is successful: On all American English consonants as well as [l]-[ɹ], the Japanese models perform worse than English models, and this difference increases with the

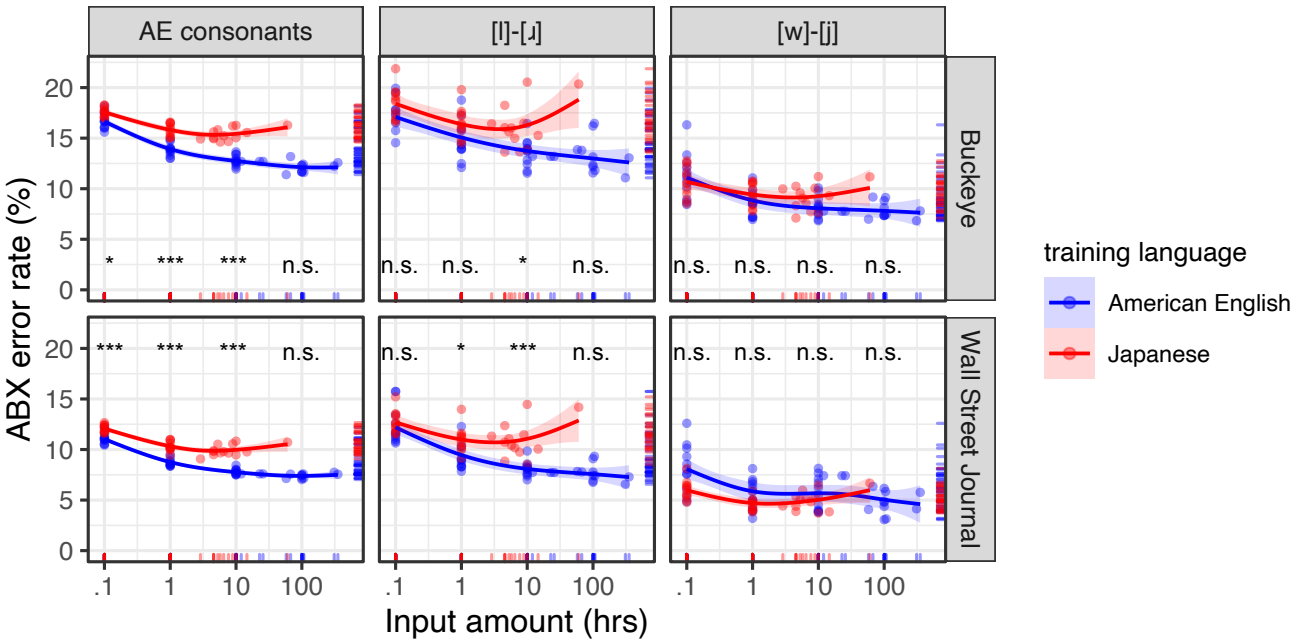

Figure 2: The effect of training duration on ABX discrimination errors, for models trained on a single speaker. Each blue dot shows one American English model, and each red dot shows one Japanese model. The colored lines correspond to smoothing splines fitted to the results of models trained on each language, and the error bands around each colored line represent 95% confidence intervals. For significance testing, the continuum of training durations is divided into four blocks: [0, 6min], (6min, 1hr], (1hr, 10hrs], (10hrs, 100hrs]. Models trained on more than 100hrs of data ($n = 3$) are not included in a block, since they are all English models and lack a Japanese counterpart to compare with. These English models are still shown in the figure for visualization purposes. For each block, permutation tests are performed to test statistical significance, in the same way as in Figure 1.

amount of input (Figure 1). While we observe no significant cross-linguistic difference on [l]-[ɹ] when the models are trained on 0.1 hours of data, there are significant differences with 1 hour and 10 hours of input data, which indicates that the models show a cross-linguistic difference with greater input amounts. The effect size also increases with more input amounts; this increase in the effect along the learning trajectory is similar to the patterns observed in infants (Kuhl et al., 2006). On the [w]-[j] pair, no significant difference is found.[3]

Models trained on single speakers

Figure 2 gives results for the models trained on single-speaker data. We can observe that the simulated English infants have lower error, i.e., better performance on American English consonants overall, than the simulated Japanese infants. As expected, while models trained on more input showed significant cross-linguistic difference between [l] and [ɹ], there is no significant cross-linguistic difference in the discrimination of [w]-[j] for any amount of training data.

Discrimination of 'native' contrasts (here for 'American English' models) appears to improve as the amount of input increases, even when the input amounts are already relatively large (e.g., 10 vs. 100 hours). Since the amount of relevant linguistic input to a child might be even larger than even the largest training sets considered here (Bergelson et al., 2019), this suggests that large databases of naturalistic recordings might prove necessary to model early phonetic learning.

In all panels of Figure 2, the ABX error rate of the Japanese models increases after an initial decrease. However, this observation appears to be driven by the single datapoint with 67 hours of input, which is the only model trained on more than 15 hours of Japanese data. After the removal of this datapoint, this observation disappears.

Due to the small amount of Japanese speech data, there are only two models trained on 10–100 hours input range, and no model is trained on more than 100 hours of Japanese data. This made the statistical tests on 10–100 hours unstable and thus unlikely to be informative.

Overall, our results on single speaker models are similar to those on multiple speaker models in Figure 1. As expected, the models demonstrate cross-linguistic differences in discrimination on all American English consonants and [l]-[ɹ], whereas there is no significant difference on [w]-[j].

[3]There is a trend in the other direction in Figure 1, such that the blue (English) line lies above the red (Japanese) line. However, a two-tailed test on [w]-[j] results did not find a significant difference between languages in either direction. The trend may be attributable to the recording settings: a large part of the Japanese corpus came from a professional recording, while all English speakers except one are LibriVox contributors who record by themselves, using whatever equipment is available to them. This may have allowed the Japanese models to learn the [w]-[j] contrast better due to possibly clearer recordings. This point would only strengthen our observation on the [l]-[ɹ] contrast, as the models are able to capture the cross-linguistic effects despite the input data for English being harder to learn from.

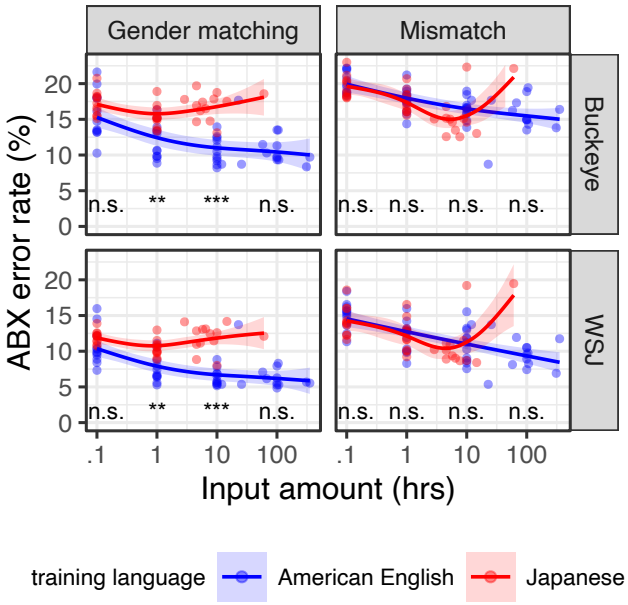

Figure 3: The effect of gender match or mismatch between training and test speakers on [l]-[ɹ] ABX discrimination error. Each blue dot shows one American English model, and each red dot shows one Japanese model. The colored lines correspond to smoothing splines fitted to the results of models trained on each language, and the colored bands represent 95% confidence intervals. Permutation tests are performed in the same way as Figure 1.

Effect of speaker gender

Schatz et al. (2019) used input data from a gender-balanced set of multiple speakers. As a result, their models encountered different genders in training. However, the current single speaker models do not have access to this variation between speakers during training, and that lack of variation in model input might lead the models to generalize not as well as model trained on data with greater variations. Here, by analyzing the effect of speaker gender, we examine whether the model's generalization ability is affected by this difference in the input data. Using the models trained on single speakers from the previous section, we separately analyze performance on test speakers who have the same gender as the training speaker (gender-matching) and test speakers who do not (gender-mismatch).

Figure 3 shows the difference between these two conditions. The cross-linguistic difference is not significant at any training duration in the gender mismatch condition, while this difference is significant in the gender matching condition, and the difference between native and nonnative models increases with longer input amounts. We used one-tailed permutation tests (one per test corpus and duration block) to compare these data against the null hypothesis that the average cross-linguistic difference in [l]-[ɹ] discrimination errors between 'Japanese' and 'American English' models is not larger in the gender-matching condition than in the gender-mismatch condition, and found a significant difference in all cases,[4] showing that matching gender facilitates the cross-linguistic effect. Interestingly, the crosslinguistic differences in discrimination abilities observed in the gender matching condition in Figure 3 are quite larger than those observed with models trained on multiple speakers (Figure 1). In addition, the same change in learning trajectory observed in Figure 2, where Japanese models' discrimination gets worse between intermediate and large quantities of training data, appears on Figure 3. In the gender-mismatch group, the change is still driven by the 67-hour outlier, and permutation tests permuting the input amount (10 vs. more than 10 hours of input) on Japanese speakers does not show a reliable change in learning trajectory ($p = 1$ for both testing corpora). In the gender-matching group, however, there is a reliable increase in error rate when comparing models with 10 hours of input and those with more than 10 hours of input, with permutation tests yielding p-values of 0.0251 and 0.0234 for Buckeye and Wall Street Journal as the testing corpus, respectively.[5]. This result suggests that in the gender-matching condition, while the performance on native sound discrimination improves further with more input, the performance on nonnative sound discrimination deteriorates with more input: Japanese models, in successfully learning properties of their "native" language, become worse at discriminating the nonnative contrasts of [l]-[ɹ].

Discussion

In this study, we asked how outcomes of phonetic learning models are affected by differences in the input, and contributed a large public-domain corpus in American English and Japanese that is suitable for addressing this question. We first replicated earlier results of phonetic learning models (Schatz et al., 2019) on much larger, single speaker data. We found that models trained on single speaker inputs and models trained on multiple speakers both reproduce cross-linguistic patterns of phone discrimination similar to those observed in infants. While this result suggests that different (multi- and single-speaker) input could lead to similar learning outcomes, a subsequent gender-specific analysis showed that the similarity in learning outcomes can be modulated by additional factors. Specifically, we found that for single-speaker models, the expected cross-linguistic difference is found only when the gender of the test speakers matches the gender of the training speaker. This illustrates the importance of critically assessing input assumptions when modeling language acquisition.

Our results also provide evidence that models trained on more naturalistic input lead to a tighter fit to the learning outcomes observed empirically, underscoring the benefits of using naturalistic input in phonetic learning models. Compared to the conditions considered in (Schatz et al., 2019), a typical

[4] Using Benjamini-Yekutieli correction for multiple comparisons; p-values fell between < 0.0001 and 0.018.

[5] The training speaker leading to the outlier is excluded in this test to ensure the outliers are not driving the statistical significance. The results are significant ($p < 0.05$) when this training speaker is included in the testing as well.

infant is likely to get a larger amount of input (Bergelson et al., 2019), and that input is likely to come mostly from a limited number of speakers with a skewed gender balance (Bergmann & Cristia, 2018). The large, single speaker training sets we consider are thus likely to better reflect naturalistic learning conditions than the balanced multi-speaker training sets in (Schatz et al., 2019). We find that this more naturalistic input predicts larger crosslinguistic differences in discrimination abilities, and—similar to infants (Tsushima et al., 1994)—a worsening of [l]-[ɹ] discrimination for models trained on Japanese, which was not found when training on the multi-speaker input.

While it is virtually impossible to test the entire range of inputs that infants could be learning from, since infants receive a wide range of inputs depending on the culture, SES, and even gender of the child (Hoff, 2006), our work represents a step in that direction. Admittedly, audiobooks are not the most naturalistic type of data for modeling infant phonetic learning, since infants primarily hear spontaneous speech, and since in many cultures, infant-directed speech has different acoustic properties from adult conversational speech (Fernald et al., 1989). In this study, we used audiobook data instead of infant-directed speech because no existing corpus of infant-directed speech is both long enough for training the current models, and labeled by speaker to allow us to control the distribution of speakers in the models' input. However, since the audiobooks data we used were naturalistic in many ways—we made direct use of speech recordings, unlabeled and unsegmented—the learning outcomes from audiobook data should be more generalizable to infant phonetic learning outcomes than previous modeling studies (De Boer & Kuhl, 2003; Feldman, Griffiths, Goldwater, & Morgan, 2013; McMurray, Aslin, & Toscano, 2009; Vallabha, McClelland, Pons, Werker, & Amano, 2007).

Our replication of the cross-linguistic effect originally found in multiple speakers (Figure 1) on models trained on single speakers (Figure 2) is compatible with the empirical literature. Bergmann and Cristia (2018) studied the relationship between number of talkers in infants' speech input and the infants' ability to discriminate native phonetic contrasts. Their overall results do not show an effect of the number of input speakers on the infants' ability to discriminate vowels in their native language. This is supported by our modeling results showing that models trained on one or 11 speakers are both successful in demonstrating phonetic learning of their native language.

Regarding infants' ability to generalize speech learning outcomes to different speakers, the literature does not give a conclusive answer. Kuhl (1979) and Kuhl (1983) found that 6-month-olds trained on one gender and tested on another can successfully generalize the task to the other gender on a phonetic discrimination task, whereas Houston and Jusczyk (2000) found that infants' ability to generalize across genders on a word segmentation task improved substantially between 7.5 and 10.5 months. While Kuhl (1979) and Kuhl (1983) examined vowel discriminability, which is closer to the current work than the word segmentation task in Houston and Jusczyk (2000), the vowel discrimination studies did not control for, or collect data on, the speaker distribution in the infants' speech learning environment. It is difficult to tell from these data whether sensitivity to phonetic contrasts would be modulated by speaker gender match between infants' input and the test stimuli used in laboratory studies.

If our model of how infants learn is correct, then in the setting that American children are most often in (Bergelson et al., 2019)—few speakers and mostly female—one might expect to see the same learning outcomes found in the gender-mismatch models, which is a lack of cross-linguistic difference in discriminating ability. However, the effect might not be as strong as it was in our models, since infants typically hear more than one speaker, and encounter speech from males at least sometimes. A direct behavioral test of whether infants' phone discrimination performance is affected by whether the gender of the speaker who recorded the experiment stimuli matches the gender of the primary speakers in their environment could be carried out. The presence of such an effect would suggest that early phonetic learning is more sensitive to input conditions than had been previously realized. Its absence might indicate that infants have strong enough generalization abilities capable of correcting for variability due to gender in the speech input, and that our phonetic learning model should be augmented with speaker normalization algorithms that reduce the between-speaker variability of the acoustic input.

In conclusion, our work shows that the properties in the speech input matter in training models of early phonetic learning. This highlights the importance of using naturalistic input for such models, and of testing different types of input when simulating the way in which children learn.

Acknowledgments

This research was supported by the National Science Foundation (NSF BCS-1734245) and the Economic and Social Research Council (ESRC-SBE ES/R006660/1). We also thank the three anonymous reviewers for their feedback and suggestions.

References

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 1165–1188.

Bergelson, E., Casillas, M., Soderstrom, M., Seidl, A., Warlaumont, A. S., & Amatuni, A. (2019). What do North American babies hear? A large-scale cross-corpus analysis. *Developmental Science*, *22*(1), 1–12. doi: 10.1111/desc.12724

Bergmann, C., & Cristia, A. (2018). Environmental influences on infants' native vowel discrimination: The case of talker number in daily life. *Infancy*, *23*(4), 484–501. doi: 10.1111/infa.12232

Chang, J., & Fisher III, J. W. (2013). Parallel sampling of DP mixture models using sub-cluster splits. In *Advances in Neural Information Processing Systems* (pp. 620–628).

Cristia, A. (2011). Fine-grained variation in caregivers' /s/ predicts their infants' /s/ category. *Journal of the Acoustical Society of America*, *129*(5), 3271-3280.

Cristia, A. (2018). Language input and outcome variation as a test of theory plausibility: The case of early phonological acquisition.

De Boer, B., & Kuhl, P. K. (2003). Investigating the role of infant-directed speech with a computer model. *Acoustics Research Letters Online*, *4*(4), 129–134.

Feldman, N. H., Griffiths, T. L., Goldwater, S., & Morgan, J. L. (2013). A role for the developing lexicon in phonetic category acquisition. *Psychological Review*, *120*(4), 751.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, *1*(2), 209-230.

Fernald, A., Taeschner, T., Dunn, J., Papousek, M., de Boysson-Bardies, B., & Fukui, I. (1989). A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants. *Journal of Child Language*, *16*(3), 477–501.

Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday experience of young american children*. Paul H Brookes Publishing.

Hoff, E. (2006). How social contexts support and shape language development. *Developmental Review*, *26*(1), 55–88.

Houston, D. M., & Jusczyk, P. W. (2000). The role of talker-specific information in word segmentation by infants. *Journal of Experimental Psychology: Human Perception and Performance*, *26*(5), 1570.

Kuhl, P. K. (1979). Speech perception in early infancy: Perceptual constancy for spectrally dissimilar vowel categories. *Journal of the Acoustical Society of America*, *66*(6), 1668-1679.

Kuhl, P. K. (1983). Perception of auditory equivalence classes for speech in early infancy. *Infant Behavior and Development*, *6*, 263-285.

Kuhl, P. K., Stevens, E., Hayashi, A., Deguchi, T., Kiritani, S., & Iverson, P. (2006). Infants show a facilitation effect for native language phonetic perception between 6 and 12 months. *Developmental Science*, *9*(2), F13–F21.

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82*(3), B101–B111.

McMurray, B., Aslin, R. N., & Toscano, J. C. (2009). Statistical learning of phonetic categories: Insights from a computational approach. *Developmental Science*, *12*(3), 369–378. doi: 10.1111/j.1467-7687.2009.00822.x

Mermelstein, P. (1976). Distance measures for speech recognition, psychological and instrumental. *Pattern Recognition and Artificial Intelligence*, *116*, 91-103.

Paul, D. B., & Baker, J. M. (1992). The design for the Wall Street Journal-based CSR corpus. In *Proc. SNL*.

Pitt, M. A., Johnson, K., Hume, E., Kiesling, S., & Raymond, W. (2005). The Buckeye corpus of conversational speech: Labeling conventions and a test of transcriber reliability. *Speech Communication*, *45*(1), 89–95.

Schatz, T. (2016). *ABX-discriminability measures and applications*. Unpublished doctoral dissertation.

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. Retrieved from `psyarxiv.com/fc4wh`

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

Song, J. Y., Demuth, K., & Morgan, J. (2018). Input and processing factors affecting infants' vocabulary size at 19 and 25 months. *Frontiers in Psychology*, *9*, 2398.

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 *Third International Conference on Spoken Language Processing*.

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

Vintsyuk, T. K. (1968). Speech discrimination by dynamic programming. *Cybernetics*, *4*(1), 52–57.

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*(1), 49–63.