Infant-directed speech is consistent with teaching

Baxter S. Eaves Jr
Rutgers University–Newark

Naomi H. Feldman
University of Maryland

Thomas L. Griffiths
University of California, Berkeley

Patrick Shafto
Rutgers University–Newark

Abstract

Infant-directed speech (IDS) has distinctive properties that differ from adult-directed speech (ADS). Why it has these properties – and whether they are intended to facilitate language learning – is matter of contention. We argue that much of this disagreement stems from a lack of a formal, guiding theory of how phonetic categories should best be taught to infant-like learners. In the absence of such a theory, researchers have relied on intuitions about learning to guide the argument. We use a formal theory of teaching, validated through experiments in other domains, as the basis for a detailed analysis of whether IDS is well-designed for teaching phonetic categories. Using the formal theory of teaching, we generate ideal data for teaching phonetic categories in English. We qualitatively compare the simulated teaching data with human IDS, finding that the teaching data exhibit many features of IDS, including some that have been taken as evidence IDS is not for teaching. The simulated data reveal potential pitfalls for experimentalists exploring the role of IDS in language learning. Focusing on different formants and phoneme sets leads to different conclusions, and the benefit of the teaching data to learners is not apparent until a sufficient number of examples have been provided. Finally, we investigate transfer of IDS to learning ADS. The teaching data improves classification of ADS data, but only for the learner they were generated to teach (the naive, infant-like learner) and not universally across all classes of learner. This research offers a theoretically-grounded framework which empowers experimentalists to systematically evaluate whether IDS is for teaching.

Keywords: Infant-directed speech, language acquisition, social learning, Bayesian model

Children learn language from input, but often the input children receive differs markedly from normal speech. Infant-directed speech (IDS, also known as “motherese”) is characterized by reduced speed, elevated pitch and affect, and unusual prosody. Infants are able to distinguish IDS from normal, adult-directed speech (ADS) and prefer IDS over ADS (Pegg, Werker, & McLeod,

This work was funded in part by NSF CAREER award DRL-1149116.

1992). Subsequently, researchers have sought to answer why it is that adults speak to children in this unusual way. Seminal work by Kuhl et al. (1997) found that IDS has unusual formant-level properties. Formants are the representative frequencies of vowel phonemes and manifest as peaks in the spectral envelope. The first formant is the lowest frequency peak, the second formant is the second lowest, and so on. IDS's corner vowels (/ɑ/, as in pot; /i/, as in beet; /u/, as in boot) are hyper-articulated, resulting in an increased vowel space. Intuitively speaking, hyper-articulation should improve the learnability of vowel categories. All things being equal, example clusters that are more distant are easier to identify. This sparked the idea that IDS is for teaching; an idea that after nearly two decades remains a matter of controversy among researchers.

Research suggests that corner vowel hyper-articulation is not simply an unintended consequence of highly-affectual speech. Corner vowel hyper-articulation is present in speech to infants but not speech to pets (Burnham, Kitamura, & Vollmer-Conna, 2002). Additionally, corner vowel hyper-articulation is found in speech to foreigners (Uther, Knoll, & Burnham, 2007), which, outwardly, sounds more like normal, adult speech. In fact, the social learning literature refers to IDS as an *ostensive cue*: a social cue that engages stricter learning mechanisms in its target (Gergely, Egyed, & Király, 2007). It would appear that IDS and its unique features are optimized to teach learners the vowel categories of their language.

However, recent work has discovered statistical features of IDS that are potentially detrimental to learning. Other, non-corner vowels are hypo-articulated (closer together) in IDS (Kirchhoff & Schimmel, 2005; Cristia & Seidl, 2013) and within-phoneme variability increases for some vowels (de Boer & Kuhl, 2003; McMurray, Kovack-Lesh, Goodwin, & McEchron, 2013). Hypo-articulation is argued to be detrimental to learning because clusters of examples become less distinct as they become nearer. Increased variability is argued to be detrimental because as clusters increase in size, their effective borders shrink or overlap, which makes them less discriminable. Additionally, Martin et al. (2015) found that temporally sequential pairs of vowel phonemes are less discriminable in IDS than in ADS. It would appear that IDS and its unique features may make learning phonetic categories more difficult.[1]

Over the course of the debate about the role of IDS in language learning, researchers have attempted to quantitatively evaluate the benefit of IDS to learners by comparing the outcome of different learning algorithms given IDS and ADS data (de Boer & Kuhl, 2003; Kirchhoff & Schimmel, 2005; McMurray et al., 2013). These studies have achieved mixed results. de Boer and Kuhl (2003) found that a mixture model trained using the expectation-maximization algorithm was better able to recover the means of IDS corner vowel categories from IDS data than it was to recover the means of ADS corner vowel categories from ADS data. Kirchhoff and Schimmel (2005) explored the usefulness of IDS to training Bayesian automatic speech recognition systems (ASR), finding that the IDS-trained ASR classified certain types of data more effectively than ADS-trained ASR and other types more poorly. McMurray et al. (2013) found that multinomial logistic regression trained on IDS data correctly classified fewer new IDS examples than its ADS-trained counterpart classified new ADS examples. Based on these results, the debate appears only to be farther from being resolved.

We argue that much of the disagreement in the literature with respect to whether IDS is optimized for teaching stems from a lack of a coherent theoretical framework for characterizing teaching. In the absence of such a framework, researchers have substituted intuitions about learning. This has three significant limitations. First, researchers have largely intuited which qualitative features are desirable and which are not. Second, existing computational approaches have attempted to assess teaching indirectly through improvements in learning using various, very different, computational models. Moreover, assessments of model performance have not focused on the key question: the implications of training on IDS for categorization of ADS. Third, the literature tends to focus attention on subsets of the data, both in terms of the vowels and the formants considered for any given analysis.

[1] Related but orthogonal work suggests that infant- and child-directed speech is less intelligible to adults (Bard & Anderson, 1983, 1994).

Each limitation potentially undermines interpretation. First, computational models are preferable to intuitive arguments precisely because intuition is fallible, especially when considering the kinds of interactions involved in teaching many categories in a low-dimensional space. Second, while we would expect teaching to lead to better learning, teaching is defined in terms of the intent of the speaker, thus improvements in learning are not a necessary implication—especially if the learner used for performance benchmarking solves a different problem than the learner for whom the teacher generates data. Moreover, given that learners ultimately need to acquire ADS, any improvements in learning should be in transfer between IDS and ADS. Third, because teaching involves considering not just the target vowel but also potentially confusable alternatives, any results derived from subsets of the data may lead to unrepresentative predictions. It is thus important to investigate whether these limitations do affect conclusions in the literature.

Our contribution to the debate is a formal theoretical analysis of how phonetic categories should optimally be taught to infant-like learners. This is the first work to directly address whether IDS is consistent with optimal teaching. We begin by defining the teaching and learning problems under a probabilistic framework. From this model, we generate data designed to teach. We address whether certain features of data are consistent with teaching by qualitatively comparing the features of the teaching data with those of IDS. We address whether IDS-like data are beneficial for learning normal (ADS) speech, and whether these effects generalize, by comparing learning transfer under the target learning model and under standard machine learning algorithms. We also identify some important caveats related to computational analyses based on subsets of data. We address the problems with looking at dimensional and categorical subsets of the data by comparing the features of, and learning outcomes given the original teaching data with those of the teaching data projected onto two-formant space, and we compare the effect of sample size (the number of IDS examples) on learning performance given ADS data and teaching data. We conclude by discussing limitations of the current work and future directions.

Teaching and learning

To simulate teaching, we must define the components of teaching. In this section we define, in mathematical terms, the components of the problem: the teacher, the learner, and the concept to be learned and taught. Mathematically defining the concept (the phonetic category model) is matter of applying a formalism that is sufficiently representative of the concept. Similarly, defining a learner requires applying a learning framework that is capable of learning the concept and does so in a psychologically-valid way. And, as we shall see, defining a teacher requires defining a data selection method that is intended to induce the defined concept in the defined learner. Throughout the paper, the words *teacher* and *learner* will be used to refer to the definitions in this section; we will make the necessary distinction when referring to human learners.

What is being taught and what is being learned

In their work on automatic speech recognition, Kirchhoff and Schimmel (2005) posed the question of what is being learned from IDS. If IDS is for teaching then what does IDS teach? While it is typically implied that the intent would be to teach normal speech, existing computational studies compare the effectiveness of IDS at teaching IDS with the effectiveness of ADS at teaching ADS (de Boer & Kuhl, 2003; McMurray et al., 2013). That is, these studies evaluate whether IDS is better at teaching an abnormal (non-adult) speech model than ADS is at teaching the normal speech model. Here, we assume that it is the intent of a teacher to teach the set of phonetic categories used in normal speech.

Building on previous research formalizing phonetic categories, we adopt a Gaussian mixture model (GMM) framework (de Boer & Kuhl, 2003; Vallabha, McClelland, Pons, Werker, & Amano, 2007; Feldman, Griffiths, Goldwater, & Morgan, 2013; McMurray, Aslin, & Toscano, 2009). Each phonetic category is represented as a multidimensional Gaussian in formant space. We focus on the first, second, and third formants, denoted F_1, F_2, and F_3, which we capture with 3-dimensional Gaussians.

A GMM is defined by the probability density function

$$f(X|\pi_1, \ldots \pi_k, \mu_1, \ldots, \mu_k, \Sigma_1, \ldots \Sigma_k) = \sum_{i=1}^{k} \pi_i \mathcal{N}(X|\mu_i, \Sigma_i), \tag{1}$$

where $\{\pi_1, \ldots, \pi_k\}$ is a set of k components weights (real numbers between 0 and 1 inclusive and which sum to 1), $\{\mu_1, \ldots, \mu_k\}$ is a set of component means, $\{\Sigma_1, \ldots, \Sigma_k\}$ is the set of component covariance matrices, and $\mathcal{N}(X|\mu, \Sigma)$ is the Normal (Gaussian) probability density function applied to the data X given μ and Σ.

Importantly, we view the *whole system* of phonetic categories as being the object that is being taught. The best data for teaching a single phonetic category might be different from the best data for teaching that category in the context of a set of other categories. When learning a single category, data that are representative of that category are sufficient to communicate the relevant statistical information. When learning multiple categories, without a clear indication of what category each sound belongs to, the possible ambiguity of each sound interacts with the need to provide good information about the statistics of each category to create a much more complex problem.

Learning

Teaching data are by definition generated with the learner in mind (Shafto & Goodman, 2008; Shafto, Goodman, & Griffiths, 2014). A teacher chooses data to induce the correct belief in learners, hence we must define the learner.

Previous computational accounts of learning under IDS have evaluated learning in computational learners that know the correct number of categories (de Boer & Kuhl, 2003) or learn from labeled data (McMurray et al., 2013). These approaches miss an important difficulty of the learning problem infants face. Infants are not born knowing how many phonemes comprise their native language nor are they given veridical feedback as to which phonetic categories individual components of utterances belong to. In order to learn the locations (means, μ) and shapes (covariance matrices, Σ) of phonetic categories, infants must learn how many there are; all while inferring to which phonetic categories each example belongs.

Learning the nature and the number of categories simultaneously can be done using the Dirichlet process Gaussian Mixture Model (DPGMM) (J. Anderson, 1991; Escobar & West, 1995; Rasmussen, 2000; Sanborn, Griffiths, & Navarro, 2010). The basic idea is that when a learner cannot assume a fixed number of categories, she must allow for the possibility that there may be as many categories as there are data. This problem can be addressed by using a probabilistic process that determines which data are assigned to which categories (see Rasmussen, 2000). Rather than learning the weights of infinitely many categories, the learner learns an assignment, $Z = \{z_1, \ldots, z_n\}$ where z_i is an integer indicating to which component of the mixture the i^{th} datum belongs. Imagine that we have observed n examples to which we have attributed k categories. Assuming no upper bound on the number of categories, a new example may be assigned to one of the k existing categories or—if it is especially anomalous—may warrant creation of a new, singleton category (a category of which datum $n + 1$ is the only member). The mixture weights are then implicit in Z. Components with more assigned data have higher weights. We outline this approach in more detail in Appendix A.

Teaching

We employ an existing model of teaching that has been used successfully to capture human learning in a variety of scenarios (Shafto & Goodman, 2008; Bonawitz et al., 2011; Shafto et al., 2014; Gweon, Pelton, Konopka, & Schulz, 2014), under which optimal teaching data derive from the inverse of the learning process. Rather than sampling data randomly from the true distribution, optimal data for teaching are sampled from the distribution that leads learners to the correct inference. Thus teaching involves directing learners' inferences; not just toward the correct hypothesis, but away from alternatives.

Mathematically, the goal of the teacher is to maximize the posterior probability that the learner ends up with the correct hypothesis—in this case, the correct estimate of the category assignments Z and the mixture parameters $\boldsymbol{\mu}$ (all the means μ) and $\boldsymbol{\Sigma}$ (all the covariance matrices Σ). To express this idea—and allow for the fact that there will be some stochasticity in teaching—we define the probability that the optimal teacher generates data X to be proportional to the posterior probability of the correct hypothesis given that value of X. Formally,

$$P_{\text{opt}}(X|Z,\boldsymbol{\mu},\boldsymbol{\Sigma}) \propto \frac{P(Z,\boldsymbol{\mu},\boldsymbol{\Sigma}|X)}{\int_X P(Z,\boldsymbol{\mu},\boldsymbol{\Sigma}|X)dX} \tag{2}$$

where the denominator normalizes the distribution, ensuring that it sums to 1 over all X.

Recall that arguments for or against IDS as pedagogical input in existing research rely on the assumption that the pedagogical intent of data can be measured by its benefit to learners. To the contrary, as we shall see, the benefit of data to learners is not a strict indication of the pedagogical intent of data even in our ideal teacher-learner scenario. For example, if the target concept is complex, large amounts of data may be required before any benefit over random data (data generated directly from the target concept) becomes apparent. Alternatively, the adherence of some data to patterns consistent with pedagogically-selected data does provide evidence of pedagogical intent. But without a rigorous definition of pedagogical data selection one can only guess at what these patterns are.

The output of the teaching model is dependent on what is being taught and how it is being taught. Because our goal is to evaluate a claim in the literature, in keeping with the literature—which is framed in terms of learning phonemes from formants—we generate data to teach a subset of language (a specific phonetic category model derived from Hillenbrand, Getty, Clark, and Wheeler [1995]) by manipulating first, second, and third formant values. This is a significant simplification of the real-world problem and makes the teaching problem both easier and more difficult. It is easier because a less complicated model requires less computation to teach, and a teacher need not be concerned with which features are relevant to learners or whether learners must learn which features are relevant (we assume learners use F_1-F_3); and it is more difficult because we have reduced the information to the learner and reduced the number of manipulable dimensions for the teacher. Thus, the teaching output should be interpreted with care. Differences between our formalization of the problem and nature's will result in differences between the model output and empirical data. We expect the output to be qualitatively similar to human IDS, but do not expect all observed trends to match exactly.

Comparison with Human Infant-Directed Speech

To evaluate the predictions that this formal model makes about the optimal data for teaching a system of phonetic categories, we focus on twelve American English vowel phonemes and their first, second, and third formants, F_1, F_2, and F_3. Hillenbrand et al. (1995) provide 48 examples of each phoneme from female speakers. Examples with unmeasurable formant values were discarded, leaving several phonemes with fewer examples (see Table 1). The target model – the one that teachers should be trying to convey to learners – was derived from the means and covariance matrices calculated from each phoneme's examples (the full list of phonemes and their means and variances can be found in Table 1).

Using an algorithm outlined in Appendix A, we generated a total of 10,000 samples from the distribution defined in Equation 2, each consisting of one example of each of the 12 phonetic categories. We then analyzed these samples, comparing them to human ADS and IDS. Figure 1a shows the distributions of the ADS vowels and the model predictions for IDS along the first and second formants.

Table 1

List of Phonemes in International Phonetic Alphabet Transcription with Means and Variances Calculated from Hillenbrand, Getty, Clark, and Wheeler (1995).

			mean			variance			covariance		
IPA	e.g.	n	F_1	F_2	F_3	F_1	F_2	F_3	F_1-F_2	F_1-F_3	F_2-F_3
æ	bat	47	678.06	2332.47	2972.68	4627.84	25475.73	40006.61	-4247.73	-1274.09	21255.98
ɑ	pot	47	916.36	1525.83	2822.57	8449.84	15615.80	27556.25	4354.50	1197.37	448.93
ɔ	bought	47	801.02	1188.28	2819.21	5172.15	16614.68	44701.74	6057.43	128.67	99.29
ɛ	bet	48	726.67	2062.54	2952.35	5454.06	20402.51	36093.30	-854.33	3539.42	11775.23
e	bait	44	536.86	2517.09	3049.86	3807.70	24872.41	32855.10	-1656.22	-1608.30	19084.57
ɝ	Bert	40	526.60	1589.35	1929.85	2193.73	12356.90	17234.28	-402.32	989.35	10092.08
ɪ	bit	48	484.31	2369.10	3057.12	1181.03	22330.69	36138.92	-182.84	1726.00	19153.52
i	beet	45	435.47	2755.96	3372.76	1662.21	20746.41	56255.83	967.00	1010.07	18241.44
o	boat	48	555.46	1035.52	2828.29	6496.21	15020.30	35040.38	6953.69	-16.69	771.31
ʊ	put	48	518.65	1228.56	2829.44	1695.72	20907.53	33424.00	2399.33	232.84	1976.00
ʌ	but	48	760.19	1415.67	2900.92	3312.88	13318.10	29810.38	2538.87	3730.06	6977.70
u	boot	48	459.67	1105.52	2735.40	1496.06	42130.34	19576.20	-417.93	-57.95	2436.00

The model predicts that the simulated teaching data do not simply parrot the target distribution but modify it in ways that match infant-directed speech. Specifically, consistent with previous research (Kuhl et al., 1997; Cristia & Seidl, 2013; Burnham et al., 2002) the corner vowels are hyper-articulated. Additionally, features that researchers have used to argue against the potential pedagogical intent of IDS are present in the teaching data. Figure 2 shows the predicted change in Euclidean distance between all pairs of vowels. We chose Euclidean distance rather than a variance-based measure of intelligibility because hyperarticulation is defined in terms of movement; the intelligibility of individual phoneme pairs is misleading in the context of teaching to infants (it is well known that IDS is less intelligible to adults [Bard & Anderson, 1983, 1994]) because teaching has to do with conveying the entire category model. Most vowel pairs are hyper-articulated, but consistent with IDS, and contrary to previous arguments that IDS is not for teaching (Cristia & Seidl, 2013), the simulated teaching data include hypo-articulation of some vowel pairs. Figure 3 shows the predicted effects on within-category variability. Consistent with IDS (de Boer & Kuhl, 2003; Cristia & Seidl, 2013), but contra previous arguments (McMurray et al., 2013), the statistically optimal input includes increases in within-category variability for most categories. Of note is the the difference in behaviour between variances and covariances. Other than /ɑ/ in F_1 and /ɝ/ in F_3, each phoneme's variance increases. The covariance behavior is less uniform. Four of twelve phonemes decrease F_1-F_2 covariance, six of twelve decrease F_3-F_1 covariance, and four of twelve decrease F_3-F_2 covariance. This suggests that though the teaching data in general exhibit greater variance, orientation plays a role.

It is important to note that trends in hyper- and hypo-articulation change when the three-formant data are flattened onto two dimensions (Figure 2a, b). Figure 2a shows the change in distance between each phoneme pair in three dimensions (F_1, F_2, F_3) and Figure 2b shows the change in distance in the same data within the F_1-F_2 plane. All corner vowel pairs are hyper-articulated in both sets, but many of the pairs that are hyper-articulated in three-formant space show little change, or are hypo-articulated, in two-formant space. This demonstrates that measures (and thus, conclusions) derived from a dimensional subset of teaching data may provide an incomplete view of the data. For example, it is not appropriate to argue that the data are not for teaching because the /o/-/u/ and /ɔ/-/ɝ/ pairs are hypo-articulated in the two-formant projection because the data were not generated to teach using only F_1 and F_2.

These results include some divergences from human IDS. IDS studies focus on different languages and dialects, and different interior vowels; because the model output is designed to teach an American English phonetic category model, we limit our discussion of systematic deviations to those

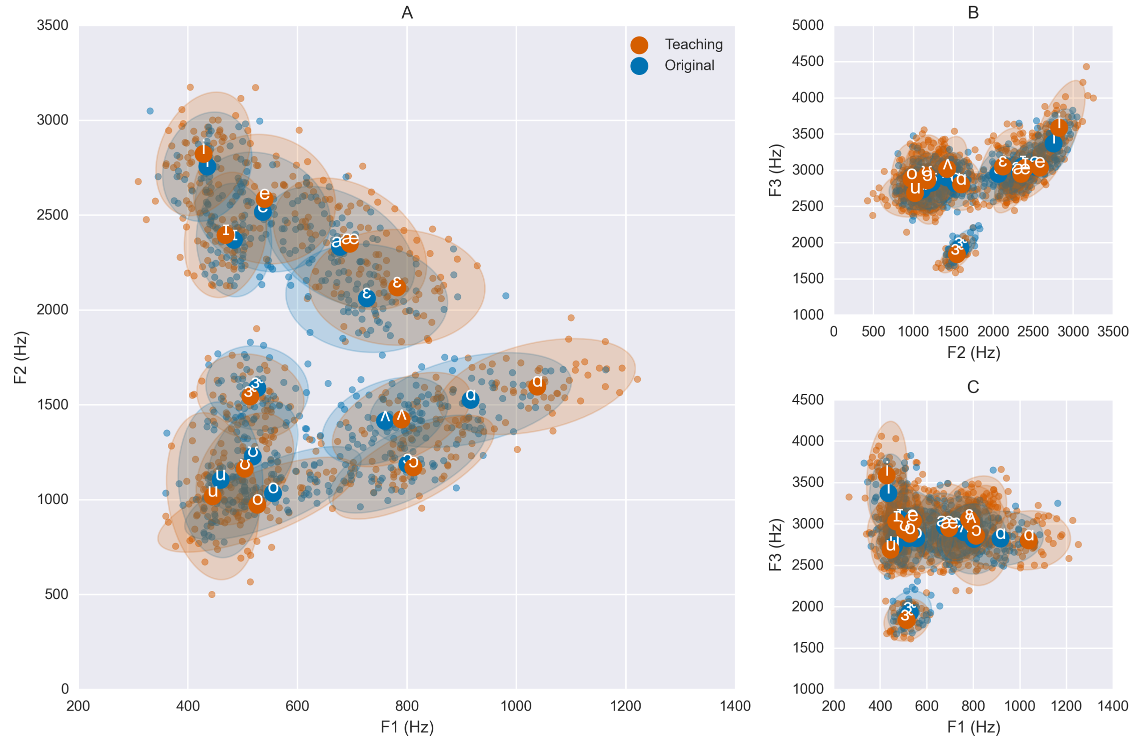

Figure 1. Distributions of vowels along first, second, and third formants (F_1, F_2, and F_3) in adult-directed speech (blue) and speech optimized for the learner (orange). Differences in distributions correspond to the properties of infant-directed speech. Labels are placed at each mean, ellipses represent covariance matrices, and points are a randomly-selected subset of samples from the teaching data and the full set of adult data. All of the original ADS data are represented while a random subset of the teaching data are represented.

between the model output and American English IDS. Though the corner vowels hyper-articulate in the teaching data, American English IDS corner vowels hyper-articulate more uniformly (see Kuhl et al., 1997; Cristia & Seidl, 2013) than the teaching data, which exhibit most hyper-articulation in /ɑ/. In general, the phonemes in the teaching data move away from the interior of the vowel space in the F_1-F_2 plane, while McMurray et al. (2013) observed that /ɝ/ and /æ/ moved toward the interior.[2] Cristia and Seidl (2013) observed that the F_1-F_2 distance between the /i/-/ɪ/ pair did not change (or hypo-articulated, depending on the measure) from ADS to IDS. Given these discrepancies, our analysis cannot be taken on its own to provide conclusive evidence that IDS is optimized for teaching. It does, however, motivate further investigation of previous findings in the literature that have been presented as evidence against IDS serving a teaching function.

Effect on learning

Earlier we argued that the benefit of teaching data is not a strict indication of its pedagogical intent—the implication being that finding that human IDS does or does not improve the performance of some learning algorithm is not, on its own, evidence that IDS is or is not meant to teach. This raises the question of why we should bother investigating learning at all. Certain patterns

[2] We assume McMurray et al. (2013) focused on native American English speakers though they only specify that participants were "from the Ripon, WI area" and "all were Caucasian and lived in homes where English was the primary language" (p. 366).

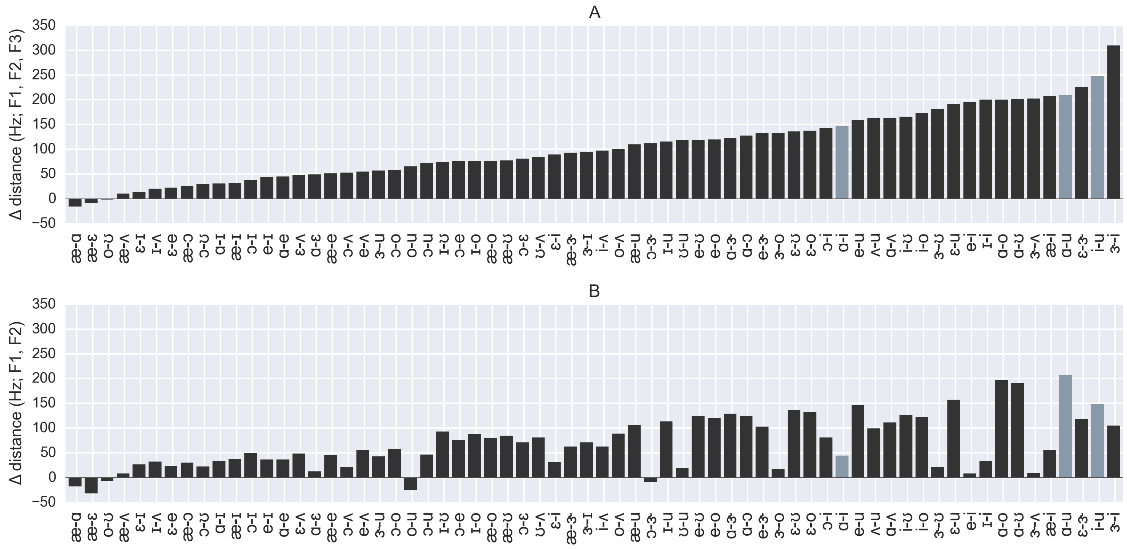

Figure 2. Change in Euclidean distance (Hz; vertical axis) between phonemes pairs (horizontal axis) from ADS to teaching data. Gray bars represent corner vowel pairs. *A*) Given the full, three-formant data. *B*) Given the three-formant data projected onto the F_1-F_2 plane.

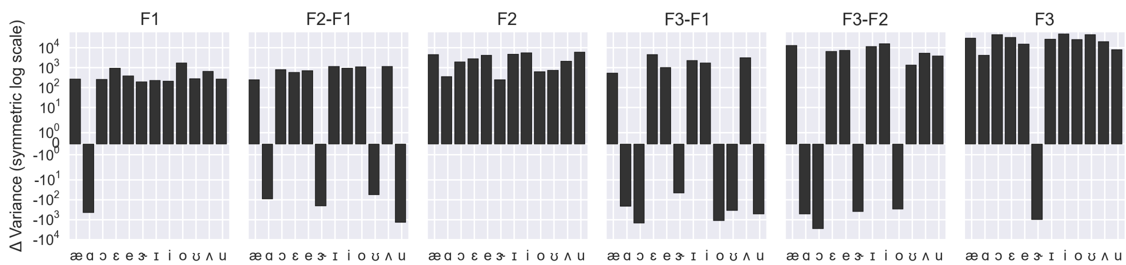

Figure 3. Change in variance, and covariance (symmetric log scale vertical axis) from ADS to teaching data for each phoneme (horizontal axis).

of learning behavior may be indicative of the presence or absence of pedagogical intent if they are consistent or inconsistent with the predictions of the theory. In this section we venture to identify such patterns. We explore the benefit of the simulated teaching data to several classes of learner, focusing on classification of IDS and ADS data, as well as the effect training on IDS data has on future classification of ADS data. We also investigate how learning performance changes when learning from specific subsets of formants and as a function of sample size.

We first evaluated whether the simulated teaching data, with their unintuitive pedagogical properties, are detrimental to learners' ability to classify example phonemes. We will first evaluate learning performance under several learning models: logistic regression (McMurray et al., 2013), support vector machines (SVM) with linear kernels, expectation-maximization on Gaussian mixture models (GMM) (de Boer & Kuhl, 2003), and the Dirichlet process Gaussian Mixture model (DPGMM; the learner model outlined above, and used as the basis for generating the teaching data). We used the scikit-learn (Pedregosa et al., 2011) implementation for each algorithm except DPGMM, which we implemented using the standard sequential Gibbs sampling algorithm (Neal, 2000, Algorithm 3) coupled with intermittent split-merge transitions (Jain & Neal, 2004), which improves mixing by allowing the Markov Chain to more easily move between modes in the probability distribution.

Each algorithm classified, in batch, random subsets of the teaching data and sets of ADS data randomly generated from the empirical distribution.[3] Each set of data consisted of 500 examples of each phoneme (6000 data points total). Each algorithm classified 500 sets of ADS data and 500 sets of teaching data. Logistic regression and SVM, which must first fit a model to labeled data, were provided an identically sized set of different training data and the GMM was provided with the correct number of categories. The DPGMM's prior distribution was identical to the teacher's. The choice of prior is important; the patterns of movement (hyper- and hypo-articulation and variance increase) depend on the prior assumed by the teacher (the teacher chooses data to teach a learner with a certain prior), hence the benefit of patterns of movement to the learner depend on the level of agreement between the teacher' assumed prior and the learner's prior. We evaluated the DPGMM based on its inferred assignment at the 500^{th} simulation step. We also evaluated the transfer of learning from teaching data to ADS by having each algorithm classify ADS data after having learned a model from teaching data. This *transfer condition* can be thought of as a simulation of the transfer of IDS to ADS. While this has not been evaluated in previous analyses of IDS, it is the critical condition for determining whether IDS helps learners acquire normal speech.

Similarity between each algorithm's inferred category assignments and the correct category assignments was evaluated via the adjusted Rand Index (ARI, see Hubert & Arabie, 1985). The ARI offers a measure of similarity between categorizations in circumstances in which it does not make sense to count the number of correct categorizations (i.e. to count the number of times items with label z are assigned to category z). It makes sense to use counting with logistic regression and SVM because these algorithms fit models given labeled training data and are then used to explicitly label new examples. The GMM, however, is only provided with the number of categories and does not care about their labels; a GMM can label k categories $k!$ different ways. And in addition to not caring about labels, the DPGMM is not guaranteed to have the same number of categories as the true distributions. We use ARI to evaluate all four models.

ARI is provided two partitions of data into categories: the true partition, which is part of the target model; and the inferred partition, which is generated by the learning algorithm. As an example, the partition $[1, 2, 3, 3]$, of four data into three categories implies that datum one belongs to category one, datum two belongs to category two, and data three and four belong to category three. ARI takes on values from -1 to 1 with expected value 0, and assumes the value 1 when the two partitions of stimuli into categories are identical (disregarding labels). For two partitions $\mathbf{U}$ and $\mathbf{V}$ of N data points into i and j categories, ARI is computed as follows:

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - [\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}]/\binom{N}{2}}{\frac{1}{2}[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}] - [\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}]/\binom{N}{2}}. \quad (3)$$

where n_{ij} is the number of datapoints assigned to i in $\mathbf{U}$ and j in $\mathbf{V}$, a_i is the sum $\sum_j n_{ij}$, and b_j is the sum $\sum_i n_{ij}$. ARI is an adjusted-for-chance version of the Rand Index (Rand, 1971), which is a normalized sum of the number of pairs of data points that are assigned to the same category in $\mathbf{U}$ and the same category in $\mathbf{V}$, and the number of data points that are assigned to different categories in $\mathbf{U}$ and different categories in $\mathbf{V}$.

Figure 4 (top row) shows that the teaching data (orange) lead to improved classification over ADS (blue) data in each of the algorithms we tested. Of the four algorithms, DPGMM performs the worst on the ADS data. This is unsurprising because of the four algorithms, DPGMM has the most to learn. However, DPGMM outperforms GMM on the teaching data. On the full, three-formant data, Logistic regression, SVM, and GMM all perform worst in the transfer condition (green) compared

[3]As researchers, we acknowledge that human learning does not happen in batch, but over time from sequential examples. Sequential Monte Carlo (SMC; see Sanborn et al., 2010) algorithms are designed to handle exactly these problems, but to evaluate sequential learning we must make assumptions about the sequence in which examples arrive. In the absence of a reasonable assumption about the order of examples we must marginalize (enumerate and average) over the $N!$ possible orders, which is computationally intractable.

with the ADS-only and teaching-data-only conditions, while the target learner (DPGMM) classifies ADS data better after having learned from the teaching data. These results show that the teaching data are themselves more classifiable than ADS and improve classification of ADS, in this case, only for the class of learner for which they were intended: the class of learner which must learn the number of phonetic categories. The transfer result is of particular importance and suggests that data that are statistically very different from data generated directly by the true concept can improve learning of the true concept. The real-world implication of this finding is that early learning from IDS may improve future ADS comprehension.

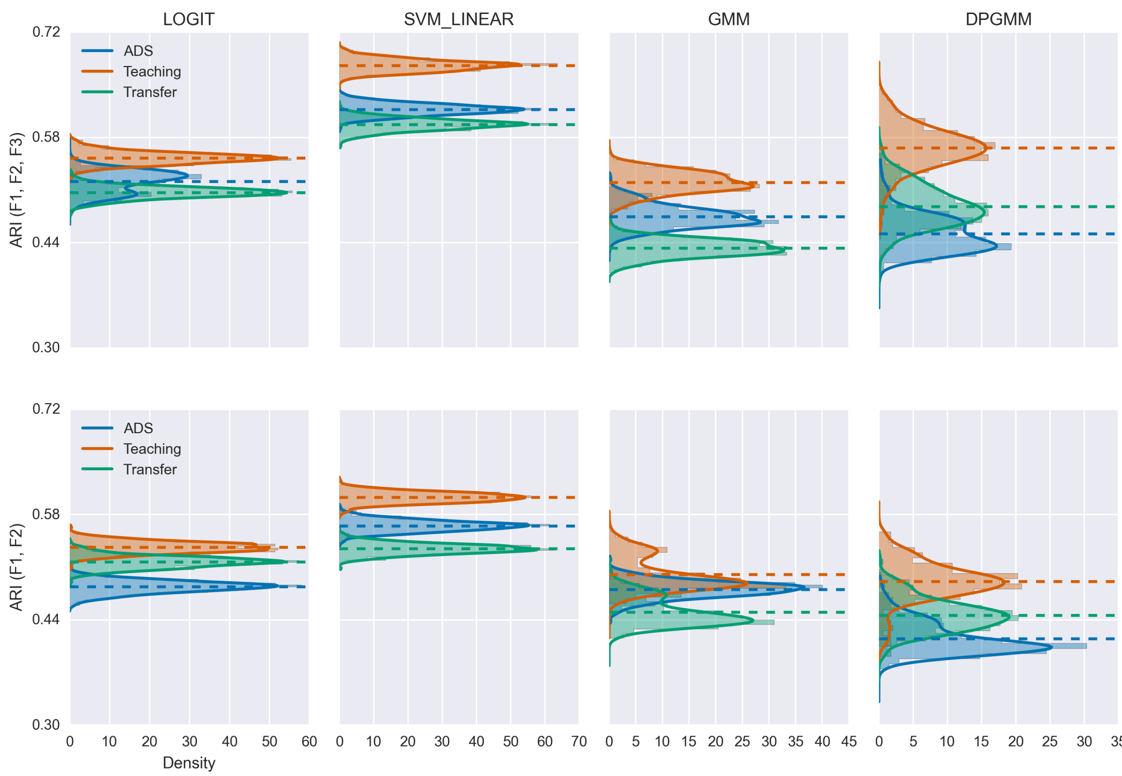

Figure 4. Distributions of ARI for four categorization algorithms (Logistic regression, support vector machine with linear kernel, finite Gaussian mixture model using expectation-maximization, and Dirichlet process Gaussian mixture model) given ADS data generated from the empirical distribution (blue), simulated teaching data (orange), and ADS after having learned from teaching data (transfer; green). *Top row*) ARI given the original, three-dimensional data. *Bottom row*) ARI given the data with the third formant removed.

We see that many of the induced ARI distributions in Figure 4 are multimodal. Two-sample Kolmogorov-Smirnov (KS) tests indicates that the distribution of ARI given three-formant ADS and teaching data differ under each algorithm; the statistic for each is significant at the $p < 10^{-40}$ level (see Table 2).[4] The categorization outcome differs when the three-formant data are projected onto the F_1-F_2 plane (see Figure 4 bottom row). Categorization performance generally decreases when

[4]We use the notation $KS_{LOGIT}(500, 500) = 0.668$ to denote that the resulting statistic of a two-sample Kolmogorov-Smirnov test on two samples, both containing 500 data points, equals 0.668

F_3 is removed. More features (dimensions) provide learners with more information by which they can form categories. For example, in Figure 1b and c we see that locating and categorizing /ɝ/ (as in Bert) becomes trivial given F_3.

Table 2
Uncorrected Kolmogorov-Smirnov Test Statistics for Figure 4. Note: p values range from $\approx 10^{-220}$ *to* $\approx 10^{-41}$.

		F_1, F_2, F_3		F_1, F_2	
Algorithm	Comparison	KS	p	KS	p
Logit	ADS-Teaching	0.894	$\ll 0.0001$	0.998	$\ll 0.0001$
	ADS-Transfer	0.584	$\ll 0.0001$	0.972	$\ll 0.0001$
	Teaching-Transfer	0.996	$\ll 0.0001$	0.828	$\ll 0.0001$
SVM (linear)	ADS-Teaching	1.0	$\ll 0.0001$	0.994	$\ll 0.0001$
	ADS-Transfer	0.822	$\ll 0.0001$	0.976	$\ll 0.0001$
	Teaching-Transfer	1.0	$\ll 0.0001$	1.0	$\ll 0.0001$
GMM	ADS-Teaching	0.872	$\ll 0.0001$	0.434	$\ll 0.0001$
	ADS-Transfer	0.932	$\ll 0.0001$	0.69	$\ll 0.0001$
	Teaching-Transfer	1.0	$\ll 0.0001$	0.830	$\ll 0.0001$
DPGMM	ADS-Teaching	0.946	$\ll 0.0001$	0.886	$\ll 0.0001$
	ADS-Transfer	0.54	$\ll 0.0001$	0.596	$\ll 0.0001$
	Teaching-Transfer	0.858	$\ll 0.0001$	0.726	$\ll 0.0001$

In the previous paragraphs we demonstrated that the simulated teaching data are indeed beneficial to several classes of learners. It is important to note that these learners benefited from sets of data consisting of a fixed number (500) of examples per phoneme. Here we investigate how this benefit changes as the number of examples increases or decreases by investigating the effect of the number of examples per phoneme on the classification ability of the target learner (DPGMM). The DPGMM classified 128 random sets of data comprising $2, 4, 8, 16, \ldots, 2048$ examples of each phoneme. The results can be seen in Figure 5. The behavior induced in the DPGMM by the ADS (blue) and Teaching (orange) data differ. Adding ADS data appears not to benefit the learner between about 32 and 256 examples per phoneme while adding teaching data continues to improve categorization at an approximately logarithmic rate. This suggest that the benefits of IDS to learners may not be apparent from a small number of data points and that researchers may benefit from comparing learning performance as a function of the number of data points. Learning under ADS begins to improve again after 512 examples, while the benefit of adding ADS examples decreases; and at 2048 examples per phoneme the transfer of IDS results in mean performance similar to ADS. Teaching data are intended to be efficient, thus they should improve learning over random data given a smaller number of examples. If the number of examples is too small, learning is difficult regardless of the data's origin; if the number of examples is sufficiently large, teaching data offer no benefit over random data.

Hypoarticulation and increasing variance to teach

It may be obvious why a teacher would hyper-articulate examples, but the pedagogical usefulness of hypo-articulation and variance increase deserves discussion. Keep in mind that the teacher seeks to increase the likelihood of a globally correct inference. Hypo-articulation can improve categorization when it is the result of disambiguating movement—that is, movement of one cluster away from another cluster it may be mistaken with. Increased variability can be used to mitigate any negative affects of hypo-articulation by making close or overlapping clusters more distinguishable from each other. Imagine two very closely overlapping, circular clusters: examples from these clus-

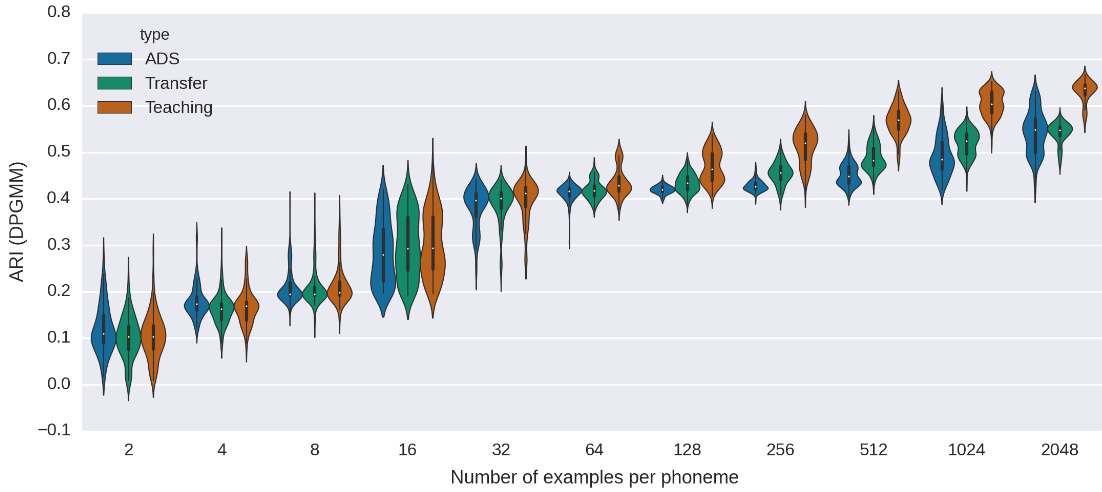

Figure 5. ARI as a function of the number of examples per phoneme for the Dirichlet process mixture model (DPGMM) given ADS data (blue), teaching data (orange), and ADS data after learning from teaching data (*transfer*; green).

ters may appear to come from one large cluster. If we wish to express that there are two clusters we could stretch each cluster perpendicularly so the resulting data manifest as an 'X' rather than a single Gaussian blob; indeed, the teaching model produces this behavior.

The teaching data offer similar examples of how hypo-articulation and increased variability, when employed systematically, do not necessarily reduce learning. For purposes of clarity, we shall look only at the F_1-F_2 plane (Figure 1a). The phonemes (/ɝ/; /u/; /ʊ/, as in put; /o/, as in boat) are difficult to distinguish in AD speech. In the teaching data, /u/, /ʊ/, and /o/ are pressed into each other (hypo-articulated) which makes /ɝ/ more distinguishable. The corner vowel /u/ greatly increases its F_2 variance and decreases its F_1-F_2 covariance and /o/ greatly increases its F_1 variance. This causes /o/ and /u/ to overlap through each other. Their tails then emerge conspicuously from the main mass of examples which makes them more identifiable. The hypo-articulation and directional changes in variance reduce the muddling effect of general increases in within-phoneme variance. Looking at the categorization performance of this subset of the flattened data shows that different algorithms come to different conclusions as to which data are better for learning (we chose categorization results on 500 examples per phoneme). SVM performs better on the ADS data ($M_{ADS} = 0.431, M_{Teach} = 0.403; KS(500, 500) = 0.716, p < 0.001; d = 2.019$) and logistic regression performs similarly on ADS and teaching data ($M_{ADS} = 0.294, M_{Teach} = 0.292; KS(500, 500) = 0.070, p = 0.166; d = 0.109$). GMM performs better on the teaching data ($M_{ADS} = 0.347, M_{Teach} = 0.353; KS(500, 500) = 0.184, p < 0.001; d = -0.301$), as does DPGMM ($M_{ADS} = 0.275, M_{Teach} = 0.283; KS(500, 500) = 0.14, p < 0.001; d = -0.231$). These result show first, that hypo-articulation and increased variance do not necessarily damage local inferences in the target model (DPGMM); and second, that looking at categorical subsets of teaching data may lead to conflicting conclusions from different learning algorithms with respect to the benefit of data to learners.

Discussion

In this paper we have explored the question of whether IDS is for teaching. We rigorously defined both the learning and teaching problems in a psychologically-valid, probabilistic theory. Using this theory, we generated data designed to teach a subset of the phonetic category model of adult speech to naive, infant-like learners using the F_1, F_2, and F_3 formants. In the process, we have identified, concretely demonstrated, and provided possible solutions to a number of issues in the existing literature. We address each in turn. We then conclude by noting the positive results of our analysis, limitations of our results, and recommendations for future research.

First, the existing literature has relied on intuitive arguments regarding which features of IDS may or may not be desirable. Hyper-articulation (expansion) of the corner vowels has been identified as a feature that would facilitate learning. However, hypo-articulation such as observed between /ɪ/ and /i/ by Cristia and Seidl (2013), and increases in variance of categories such as /æ/ and /ɝ/ observed by McMurray et al. (2013), have been argued to impede learning. Our results show that, when considered in aggregate, hypo-articulation and increases in variance are indeed consistent with teaching. Our analysis leads to predictions about when and why one may see these surprising properties. Hypo-articulation appears when vowels move away from more confusable alternatives. To compensate for this, hypo-articulated categories appear in conjunction with hyper-articulation on other formant dimensions (F_3) and/or increases in (co)variance that would facilitate the learner's inference that there is more than one category present. /o/ and /u/ are hypo-articulated in $F_1 \times F_2$, but are hyper-articulated in $F_1 \times F_2 \times F_3$. Both of these phonemes increase their F_1 and F_2 variance, but /o/ increases its F_1-F_2 covariance while /u/ decreases its $F_1 - F_2$ covariance, which causes the two phonemes to become more conspicuous by overlapping through each other. Thus, our results show that researchers' intuitive theories of which features of IDS are beneficial for teaching are contradicted by a more precise, computational analysis of teaching phoneme categories.

Second, existing computational approaches have attempted to assess teaching indirectly through improvements in learning using various, very different, computational models and have assessed the benefits of learning from IDS with transfer to IDS. We have argued that the existing models make unreasonable assumptions about the problem faced by the learner. Specifically, models assume that infants either know the number of phonemes in their language *a priori* (de Boer & Kuhl, 2003) or that the data they receive is accompanied by correct labels (McMurray et al., 2013). Prima facie, these assumptions are too strong. The problem the learner faces includes learning the number of categories. Analyses based on this problem lead to consequential differences in results. Learners who face the problem of learning the number of categories show positive effects of transfer from the simulated teaching data to ADS, while algorithms that assume labeled data or a known number of categories do not (see Figure 4). Our results based on more realistic assumptions about the learning problem contradict previous conclusions in the literature.

Third, the literature tends to focus attention on subsets of the data, both in terms of the vowels and the formants considered for any given analysis. Both empirical and computational analyses tend to focus on subsets of IDS. Rather than measuring F_1, F_2 and F_3, many analyses rely only on F1 and F2. Similarly, rather than recording data for all vowel categories, results tend to focus on subsets that are relevant to intuitively derived qualitative predictions. Our results show that predictions for teaching depend on knowledge of both of these aspects of context, and thus interpretation of empirical results do as well. As illustrated in Figure 2, hypo-articulation cannot be determined from F_1 and F_2 alone; the vowels may be separated on F_3. In fact, rhotic vowels such as /ɝ/ and /ɑr/ (as in start) are characterized by low F_3 frequencies. Similarly, hypo-articulation may be accompanied by increases in variance, which optimize the learner's ability to infer the existence of more than one category. Thus, our results show that more comprehensive data are necessary to develop accurate computational models and interpret empirical results.

Our results are based on the Hillenbrand et al. (1995) data, which do not include many of the interior and rhotic vowels use in other studies (McMurray et al., 2013; Cristia & Seidl, 2013). Because our results show that quantitative predictions are sensitive to the specifics of context, we do not expect a perfect match to the behavioral data. As we noted, the trends in the simulated teaching data did not exactly match trends others have reported in human IDS. The vowels /ɝ/ and /æ/ did not exhibit the interior movement reported by McMurray et al. (2013), nor did /i/ and /ɪ/ exhibit F_1-F_2 hypoarticulation as reported by Cristia and Seidl (2013). The qualitative implications of our analysis are more powerful as a consequence: these points illustrate the need for more comprehensive data sets to ensure progress in the debate.

Building on previous computational models of teaching, we have introduced an approach that may allow direct assessment of whether IDS is intended to teach. The analyses presented here suggest that surprising features identified by researchers are indeed predicted by the model and that IDS is indeed effective for teaching ADS categories provided one assumes a realistic model of learning. Our results also highlight challenges for research investigating the purpose of IDS.

Implicit in this problem is thus a dependence of teaching data on assumptions of what is being taught. Indeed, this dependence on the set of alternatives is likely what makes desirable features tricky to intuit. If IDS is only for teaching phonetic categories, a more complete set of phonemic data is necessary. Though we derived our target phonetic category model from a fairly extensive data set, we hardly encompass the full category model of American English.[5] We lack many of the interior vowels investigated by other researchers (see Cristia & Seidl, 2013; McMurray et al., 2013). However, it possible that IDS may be optimized for teaching a larger subset of language. Indeed, research has shown that IDS improves word segmentation (Thiessen, Hill, & Saffran, 2005), word recognition (Singh, Nestor, Parikh, & Yull, 2009), and label learning (Graf Estes & Hurley, 2012). Though daunting, our results highlight the need to systematically consider these alternatives. Our approach, in which we consider categories defined over F_1 and F_2 versus F_1, F_2 and F_3, can be viewed as a modest start in that direction. With such computational models in hand, it becomes an empirical question, albeit one that requires more comprehensive data than we currently have available.

Another concern that has not yet been addressed in the literature is differences in learning from individual caregivers and from aggregated data from multiple caregivers. Computational research has sought to answer the question of how people solve inference problems that are computationally intractable, positing that people use approximations (Sanborn et al., 2010). If this is the case, it is reasonable to assume that different caregivers will arrive at different solutions through stochastic search (e.g. Markov chain Monte Carlo). The distribution of teaching data is highly multi-modal and Markov Chains often find themselves stuck in local maxima. Pilot research suggest data from single chains is far more beneficial to learners than the data aggregated over chains—perhaps due to lower within-phoneme variability compared with aggregated data. We use the aggregated data because it represents the correct probabilistic solution, however because infants are exposed to only a few primary speakers, the literature's tendency to make comparisons over many individuals may misrepresent the problem (see Kleinschmidt & Jaeger, 2015, for a detailed discussion on how language learners may handle inter-speaker variability).

This work is also relevant to the articulation literature, where the theoretical underpinning of speakers' speech manipulations are under debate (see Buz & Jaeger, FIXME). The teaching model, coupled with a temporal model of articulation, could predict hyper- or hypo-articulation, and duration increases or decreases. Temporal effects that are explained in terms of a number of heuristics such as planning economy, phonetic neighborhood density, or binary-feature-based addressee-driven attenuation (Lindblom, 1990; Munson & Solomon, 2004; Galati & Brennan, 2010), may in fact be consistent with pedagogical manipulation. However, until the scaling of the teaching model is improved, the problem of temporal articulation will be unapproachable.

Conclusion

Increasingly, research has highlighted ways in which other people may affect learning (Gergely, Bekkering, & Király, 2002; Koenig & Harris, 2005; Bonawitz et al., 2011; Gweon et al., 2014). The problem of language, viewed as statistical learning, is in principle no different. Research has shown that people systematically vary their speech to different targets, with infant directed speech being

[5] Additionally, phonemes in Hillenbrand et al. (1995) were measured only from words beginning with an 'h' and ending with a 'd' e.g., /ɑ/, /i/, and /u/ were taken only from the words 'hod', 'heed', and 'who'd' respectively.

a canonical example. It is natural to ask, why. Is it for teaching? We have argued that precise formalization of these hypotheses is a necessary step toward the answer. Building off work in social learning, our computational model of teaching phonemes illustrates limitations in the existing literature. Our approach also points a way forward, through collection of more comprehensive datasets, and development of computational accounts that more accurately reflect the problems faced by learners and hypotheses posited by researchers.

References

Anderson, J. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*(3), 409.

Bard, E. G. & Anderson, A. H. (1983). The unintelligibility of speech to children. *Journal of Child Language*, *10*(02), 265–292.

Bard, E. G. & Anderson, A. H. (1994). The unintelligibility of speech to children: effects of referent availability. *Journal of Child Language*, *21*(03), 623–648.

Bonawitz, E., Shafto, P., Gweon, H., Goodman, N. D., Spelke, E., & Schulz, L. (2011, September). The double-edged sword of pedagogy: Instruction limits spontaneous exploration and discovery. *Cognition*, *120*(3), 322–30. doi:10.1016/j.cognition.2010.10.001

Burnham, D., Kitamura, C., & Vollmer-Conna, U. (2002, May). What's new, pussycat? On talking to babies and animals. *Science (New York, N.Y.) 296*(5572), 1435. doi:10.1126/science.1069587

Buz, E. & Jaeger, T. F. (FIXME). The (in)dependence of articulation and lexical planning during isolated word production. *FIXME*.

Cristia, A. & Seidl, A. (2013). The hyperarticulation hypothesis of infant-directed speech. *Journal of Child Language*, 1–22.

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. doi:10.1121/1.1613311

Escobar, M. D. & West, M. (1995). Bayesian density estimation and inference using mixtures. *90*(430), 577–588.

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.

Galati, A. & Brennan, S. E. (2010). Attenuating information in spoken communication: for the speaker, or for the addressee? *Journal of Memory and Language*, *62*(1), 35–51.

Gergely, G., Bekkering, H., & Király, I. (2002, February). Rational imitation in preverbal infants. *Nature*, *415*(6873), 755. doi:10.1038/415755a

Gergely, G., Egyed, K., & Király, I. (2007, January). On pedagogy. *Developmental science*, *10*(1), 139–46. doi:10.1111/j.1467-7687.2007.00576.x

Graf Estes, K. & Hurley, K. (2012, November). Infant-Directed Prosody Helps Infants Map Sounds to Meanings. *Infancy*, (499), n/a–n/a. doi:10.1111/infa.12006

Gweon, H., Pelton, H., Konopka, J. a., & Schulz, L. E. (2014, September). Sins of omission: children selectively explore when teachers are under-informative. *Cognition*, *132*(3), 335–41. doi:10.1016/j.cognition.2014.04.013

Hastings, W. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, *57*(1), 97–109.

Hillenbrand, J., Getty, L. a., Clark, M. J., & Wheeler, K. (1995, May). Acoustic characteristics of American English vowels. *The Journal of the Acoustical Society of America*, *97*(5 Pt 1), 3099–111.

Hubert, L. & Arabie, P. (1985). Comparing partitions. *Journal of classification*, *2*(1), 193–218.

Jain, S. & Neal, R. M. (2004). A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model. *Journal of Computational and Graphical Statistics*, *13*(1).

Kirchhoff, K. & Schimmel, S. (2005). Statistical properties of infant-directed versus adult-directed speech: Insights from speech recognition. *The Journal of the Acoustical Society of America*, *117*(4), 2238. doi:10.1121/1.1869172

Kleinschmidt, D. F. & Jaeger, T. F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review*, *122*(2), 148–203. doi:10.1037/a0038695

Koenig, M. & Harris, P. L. (2005). Preschoolers mistrust ignorant and inaccurate speakers. *Child development*, *76*(6), 1261–77. doi:10.1111/j.1467-8624.2005.00849.x

Kuhl, P. K., Andruski, J. E., Christovich, I. A., Christovich, L. A., Kozhevinkova, E. V., Ryskina, V. L., … Lacerda, F. (1997, August). Cross-Language Analysis of Phonetic Units in Language Addressed to Infants. *Science*, *277*(5326), 684–686. doi:10.1126/science.277.5326.684

Lindblom, B. (1990). Explaining phonetic variation: a sketch of the H&H theory. *Speech Production and Speech Modelling*, 403–439.

Martin, A., Schatz, T., Versteegh, M., Miyazawa, K., Mazuka, R., Dupoux, E., & Cristia, A. (2015). Mothers speak less clearly to infants: A comprehensive test of the hyperarticulation hypothesis. *Psychological science*, 1–7.

McMurray, B., Aslin, R. N., & Toscano, J. C. (2009, April). Statistical learning of phonetic categories: insights from a computational approach. *Developmental science*, *12*(3), 369–78. doi:10.1111/j.1467-7687.2009.00822.x

McMurray, B., Kovack-Lesh, K., Goodwin, D., & McEchron, W. (2013, November). Infant directed speech and the development of speech perception: enhancing development or an unintended consequence? *Cognition*, *129*(2), 362–78. doi:10.1016/j.cognition.2013.07.015

Munson, B. & Solomon, N. P. (2004). The effect of phonological neighborhood density on vowel articulation. *Journal of speech, language, and hearing research*, *47*(5), 1048–1058.

Murphy, K. P. (2007). *Conjugate Bayesian analysis of the Gaussian distribution*. University of British Columbia.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of computational and graphical statistics*, *9*(2), 249–265.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, E. (2011). Scikit-learn: machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Pegg, J., Werker, J., & McLeod, P. (1992). Preference for Infant-directed over adult-directed speech: Evidence from 7-week-old infants. *Infant Behavior and Development*, *15*, 325–345.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, *66*(336), 846–850.

Rasmussen, C. (2000). The infinite Gaussian mixture model. *Advances in neural information processing*, (11), 554–560.

Roberts, G. O., Gelman, a., & Gilks, W. R. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *Annals of Applied Probability*, *7*(1), 110–120. doi:10.1214/aoap/1034625254

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational models: alternative algorithms for category learning. *Psychological review*, *117*(4), 1144–1167. doi:10.1037/a0020511

Shafto, P. & Goodman, N. D. (2008). Teaching games: Statistical sampling assumptions for learning in pedagogical situations. In *Proceedings of the thirtieth annual conference of the cognitive science society*.

Shafto, P., Goodman, N. D., & Griffiths, T. L. (2014, March). A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cognitive psychology*, *71C*, 55–89. doi:10.1016/j.cogpsych.2013.12.004

Singh, L., Nestor, S., Parikh, C., & Yull, A. (2009, November). Influences of Infant-Directed Speech on Early Word Recognition. *Infancy*, *14*(6), 654–666. doi:10.1080/15250000903263973

Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2006). Hierarchical dirichlet processes. *Journal of the american statistical association*, *101*(476).

Thiessen, E. D., Hill, E. a., & Saffran, J. R. (2005, January). Infant-Directed Speech Facilitates Word Segmentation. *Infancy*, *7*(1), 53–71. doi:10.1207/s15327078in0701{_}5

Uther, M., Knoll, M., & Burnham, D. (2007, January). Do you speak E-NG-L-I-SH? A comparison of foreigner- and infant-directed speech. *Speech Communication*, *49*(1), 2–7. doi:10.1016/j.specom.2006.10.003

Vallabha, G. K., McClelland, J. L., Pons, F., Werker, J. F., & Amano, S. (2007, August). Unsupervised learning of vowel categories from infant-directed speech. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(33), 13273–8. doi:10.1073/pnas.0705369104

Appendix A
Details of model

Here we describe the mathematical details of the model. We construct the teaching model from the learning model.

Learner model

We formalize phonetic category acquisition as learning an infinite Gaussian mixture model (GMM; see Rasmussen, 2000; J. Anderson, 1991). A Gaussian mixture model comprises a set of k multidimensional Gaussian components $\theta = \{\{\mu_1, \Sigma_1\}, \dots, \{\mu_k, \Sigma_k\}\}$, where μ_j and Σ_j are the mean and covariance matrix of the j^{th} mixture component; and an k-length vector of mixture weights $\pi = \{\pi_1, \dots, \pi_k\}$, where each π_j is a positive real number and the set π sums to 1. The likelihood of some data, $X = \{x_i, \dots, x_n\}$, under a GMM is the product of weighted sums,

$$P(X|\theta, \pi) = \prod_{i=1}^{n} \sum_{j=1}^{k} \pi_j \mathcal{N}(x_i|\mu_i, \Sigma_i), \tag{4}$$

where $\mathcal{N}(x|\mu, \Sigma)$ is the Gaussian probability density function applied to x given μ and Σ.

We are concerned with the case where the learner infers the assignment of data to categories rather than the component weights. We introduce a length n assignment vector $Z = [z_1, \dots, z_n]$ where z_i is an integer in $1, \dots, k$ representing to which component datum i is assigned. Because the assignment is explicit, we no longer sum over each component. The likelihood is then,

$$P(X|\theta, Z) = \prod_{i=1}^{n} \sum_{j=1}^{k} \mathcal{N}(x_i|\mu_i, \Sigma_i)\delta_{z_i,j}, \tag{5}$$

where $\delta_{z_i,j}$ is the Kronecker delta function, which takes the value 1 if $z_i = j$ (data point x_i is assigned to the j^{th} category) and the value 0 otherwise.

Learning is then a problem of inferring θ and Z. Prior distributions on individual components, $\{\mu_j, \Sigma_j\}$, correspond to a learner's prior beliefs about the general location (μ), and the size and shape (Σ) of categories. For mathematical convenience, we assume that μ_j and Σ_j are distributed according to Normal Inverse-Wishart (denoted $\mathcal{NIW}$):

$$\mu_j, \Sigma_j \sim \mathcal{NIW}(\mu_0, \Lambda_0, \kappa_0, \nu_0) \quad \forall\, j \in \{1, \dots, k\}, \tag{6}$$

which implies

$$\Sigma_j \sim \text{Inverse-Wishart}_{\nu_0}(\Lambda_0^{-1}), \tag{7}$$

$$\mu_j|\Sigma_j \sim \mathcal{N}(\mu_0, \Sigma_k/\kappa_0) \quad \forall\, j \in \{1, \dots, k\}, \tag{8}$$

where Λ_0 is the prior scale matrix, μ_0 is the prior mean, ν_0 is the prior degrees of freedom, and κ_0 is the number of prior observations. For simulations, we chose vague prior parameters derived from the data.

$$\nu_0 = 3, \tag{9}$$

$$\kappa_0 = 1, \tag{10}$$

$$\mu_0 = \frac{1}{N}\sum_{i=1}^{N} X_i, \tag{11}$$

$$\Lambda_0 = \frac{1}{K}\sum_{k=1}^{K} \Sigma\left(X_k\right), \tag{12}$$

where $\Sigma\left(X_k\right)$ is the empirical covariance matrix of the adult data belonging to category k. The prior mean, μ_0, is the mean over the entire data set, and the prior covariance matrix, Λ_0, is the average of each category's covariance matrix (see Table 1).

To formalize inference over the number of categories, we introduce a prior on the partitioning of data points into components via the Chinese Restaurant Process (Teh, Jordan, Beal, & Blei, 2006), denoted CRP(α), where the parameter α affects the probability of new components. Higher α creates a higher bias toward new components. Data points are assigned to components as follows:

$$P(z_i = j | Z^{-i}, \alpha) = \begin{cases} \frac{n_j}{n-1+\alpha} & \text{if } j \in 1 \ldots k \\ \frac{\alpha}{n-1+\alpha} & \text{if } j = k+1 \end{cases}, \tag{13}$$

where Z^{-i} is Z less entry i, k is the current number of components and n_j is the number of data points assigned to component j. One is a minimally informative value of α corresponding to a uniform weight over components.

The standard learning problem involves recovering the true model, defined by θ and Z, from the data, X, (give any prior beliefs) according to Bayes' theorem,

$$P(\theta, Z | X, \mu_0, \Lambda_0, \kappa_0, \nu_0, \alpha) = \frac{P(Z|\alpha) P(\theta|\mu_0, \Lambda_0, \kappa_0, \nu_0) P(X|\theta, Z)}{P(X|\mu_0, \Lambda_0, \kappa_0, \nu_0, \alpha)}. \tag{14}$$

The Normal Inverse-Wishart prior allows us to calculate the marginal likelihood, $P(X|\mu_0, \Lambda_0, \kappa_0, \nu_0, \alpha)$, analytically (Murphy, 2007), thus, for a small number of data points (the specific number being limited by one's computing power and patience; in our case, the number being thirteen or fewer) we can exactly calculate the above quantity via enumeration. Expanding the terms, the numerator is,

$$P(Z|\alpha)\left(\prod_{j=1}^{k} \mathcal{NIW}(\mu_j, \Sigma_j | \mu_0, \Lambda_0, \kappa_0, \nu_0)\right) \prod_{j=1}^{k} \mathcal{N}(\{x_i \in X : Z_i = j\} | \mu_j, \Sigma_j), \tag{15}$$

where the first term, $P(Z|\alpha)$, is the probability of Z under CRP(α); the second term is the prior probability of the parameters in each component under Normal Inverse-Wishart; and the third term is the (normal) likelihood of the data in each component given the component parameters.

The denominator of Equation 14 is calculable by summing over all possible assignment vectors, $\{Z \in \mathfrak{Z}\}$, and integrating over all possible component parameters,

$$P(X|\mu_0,\Lambda_0,\kappa_0,\nu_0,\alpha) = \sum_{Z\in\mathfrak{Z}} P(Z|\alpha)\prod_{j=1}^{k_Z}\iint_\theta \mathcal{N}(\{x_i\in X : Z_i=j\}|\theta)\mathcal{NIW}(\theta|\mu_0,\Lambda_0,\kappa_0,\nu_0)d\theta \quad (16)$$

$$= \sum_{Z\in\mathfrak{Z}} P(Z|\alpha)\prod_{j=1}^{k_Z} P(\{x_i\in X : Z_i=j\}|\mu_0,\Lambda_0,\kappa_0,\nu_0), \quad (17)$$

where k_Z is the number of components in the assignment Z and $P(\{x_i \in X : Z_i = j\}|\mu_0,\Lambda_0,\kappa_0,\nu_0)$ is the marginal likelihood of the set of data points in X assigned to component j in Z under a Normal likelihood with Normal Inverse-Wishart prior (this quantity is calculable in closed-form).

Teacher model

Optimal data for teaching are sampled from the distribution that leads learners to the correct inference and away from incorrect inferences (Shafto & Goodman, 2008; Shafto et al., 2014). The teacher must consider the learner's inferences given all possible choices of data. Thus, we normalize over all possible data X,

$$P_{\text{opt}}(X|\theta,Z,\mu_0,\Lambda_0,\kappa_0,\nu_0,\alpha) \propto \frac{P(\theta,Z|X,\mu_0,\Lambda_0,\kappa_0,\nu_0,\alpha)}{\int_X P(\theta,Z|X,\mu_0,\Lambda_0,\kappa_0,\nu_0,\alpha)dX}, \quad (18)$$

$$= \frac{\frac{P(Z|\alpha)P(X|\theta,Z)P(\theta|\mu_0,\Lambda_0,\kappa_0,\nu_0)}{P(X|\mu_0,\Lambda_0,\kappa_0,\nu_0,\alpha)}}{\int_X \frac{P(X|\theta,Z)P(\theta|\mu_0,\Lambda_0,\kappa_0,\nu_0)P(Z|\alpha)}{P(X|\mu_0,\Lambda_0,\kappa_0,\nu_0,\alpha)}dX}. \quad (19)$$

The term,

$$P(\theta,Z|X,\mu_0,\Lambda_0,\kappa_0,\nu_0,\alpha) = \frac{P(X|\theta,Z)P(\theta|\mu_0,\Lambda_0,\kappa_0,\nu_0)P(Z|\alpha)}{P(X|\mu_0,\Lambda_0,\kappa_0,\nu_0,\alpha)}, \quad (20)$$

is the posterior probability of the true hypothesis given the data—the learner's inference. The learner's inference over alternative hypotheses is captured by the marginal likelihood of the data, $P(X|\mu_0,\Lambda_0,\kappa_0,\nu_0,\alpha)$. The teacher's optimization of the choice of data is captured by the normalizing constant,

$$\int_X P(\theta,Z|X,\mu_0,\Lambda_0,\kappa_0,\nu_0,\alpha)dX. \quad (21)$$

We avoid the need to calculate this quantity directly by sampling from P_{opt} using the Metropolis algorithm (Hastings [1970], see Appendix B) according to the acceptance probability,

$$A(X'|X) = \min\left[1, \frac{P(X'|\theta,Z)P(X|\mu_0,\Lambda_0,\kappa_0,\nu_0,\alpha)}{P(X|\theta,Z)P(X'|\mu_0,\Lambda_0,\kappa_0,\nu_0,\alpha)}\right]. \quad (22)$$

Appendix B
Algorithm for generating samples

The normalizing constant in Equation 2 (also Equation 21 in Appendix A) is analytically intractable. We use the Metropolis-Hastings algorithm to sample from the distribution of teaching data without having to calculate the normalizing constant (Hastings, 1970). The Metropolis-Hastings algorithm can be applied to draw samples from a probability distribution with density $p : x \to \mathbb{R}^+$ when p can be calculated up to a constant. That is, when there exists a function $f(x)$, where $p(x) = cf(x)$ and c is a constant. A proposal distribution, $q(x'|x)$, is defined that proposes new samples, x', given the current sample, x. Beginning with a sample, x, a proposed sample, x', is drawn from q. The acceptance ratio, A, is calculated from f and q,

$$A = \frac{f(x')q(x|x')}{f(x)q(x'|x)}. \tag{23}$$

It is easy to see that

$$\frac{f(x')q(x|x')}{f(x)q(x'|x)} = \frac{cf(x')q(x|x')}{cf(x)q(x'|x)} = \frac{p(x')q(x|x')}{p(x)q(x'|x)}. \tag{24}$$

If q is symmetric, that is $q(x'|x) = q(x|x')$ for all x, x', then $\frac{q(x|x')}{q(x'|x)}$ (the Hastings ratio) cancels from the equation, leaving,

$$A = \frac{f(x')}{f(x)}, \tag{25}$$

from which we calculate the probability with which x' is accepted,

$$P(x'|x) = \min\left[1, A\right]. \tag{26}$$

To sample from the distribution of teaching data using the Metropolis algorithm, we calculate the numerator of Equation 2 exactly via enumeration and propose symmetric Gaussian perturbations to resample data. The acceptance probability is thus,

$$P(X'|X) = \min\left[1, \frac{P(X'|Z, \boldsymbol{\mu}, \boldsymbol{\Sigma})P(X)}{P(X|Z, \boldsymbol{\mu}, \boldsymbol{\Sigma})P(X')}\right]. \tag{27}$$

For the simulations, the sampler simulated one datapoint for each phoneme (twelve total). X comprised twelve data points, one for each phoneme. X was initialized by sampling data from the prior parameters, that is $X_0 \sim N(\mu_0, \Lambda_0/\kappa_0)$ (see Appendix A). At each iteration, new data, X', were generated from X by adding Gaussian noise distributed $N(0, 40)$. This proposal distribution was chosen so that the acceptance rate of X' was near the optimal value of 0.23 (Roberts, Gelman, & Gilks, 1997). X' was then accepted according to Equation 27.

The final data comprise samples from 10 independent runs to the sampler. The first 500 samples of each run were discarded, then each 20^{th} sample was collected until 1000 samples had been collected. The full set of data thus contains 10,000 total samples of twelve data points each (one for each of the twelve phonemes) for a total of 120,000 examples. Aggregating data over speakers is common practice in the IDS literature; we conduct analyses on data aggregated over independent runs of the sampler.