Psychological Review
2016, Vol. 123, No. 6, 758–771

© 2016 American Psychological Association
0033-295X/16/$12.00 http://dx.doi.org/10.1037/rev0000031

THEORETICAL NOTE

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

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 a matter of contention. We argue that much of this disagreement stems from lack of a formal, guiding theory of how phonetic categories should best be taught to infantlike 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 theory, 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 improve classification of ADS data but only for the learner they were generated to teach, not universally across all classes of learners. This research offers a theoretically grounded framework that 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, 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. When plotted in formant space, the vowels that form the “corners” of the space of possible vowels (/ɑ/, as in pot; /i/, as in beet; /u/, as in boot) are hyperarticulated, making them more different from one another. This results in an expansion of the vowel space. Intuitively speaking, hyperarticulation 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 has suggested that corner vowel hyperarticulation is not simply an unintended consequence of highly affectual speech. Corner vowel hyperarticulation is present in speech to infants but not speech to pets (Burnham, Kitamura, & Vollmer-Conna, 2002). Additionally, corner vowel hyperarticulation 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 appears that IDS and its unique features are optimized to teach learners the vowel categories of their language.

This article was published Online First April 18, 2016.

Baxter S. Eaves Jr., Department of Mathematics and Computer Science, Rutgers University–Newark; Naomi H. Feldman, Department of Linguistics and Institute for Advanced Computer Studies, University of Maryland; Thomas L. Griffiths, Institute of Cognitive and Brain Sciences, University of California, Berkeley; Patrick Shafto, Department of Mathematics and Computer Science, Rutgers University–Newark.

This work was funded in part by National Science Foundation CAREER Award DRL-1149116.

Correspondence concerning this article should be addressed to Baxter S. Eaves Jr., Department of Mathematics and Computer Science, Rutgers University–Newark, Smith Hall, Room 216, 101 Warren Street, Newark, NJ 07102. E-mail: baxter.eaves@rutgers.edu

However, recent work has discovered statistical features of IDS that are potentially detrimental to learning. Other, noncorner vowels are hypoarticulated (closer together) in IDS (Cristia & Seidl, 2013; Kirchhoff & Schimmel, 2005) and within-phoneme variability increases for some vowels (de Boer & Kuhl, 2003; McMurray, Kovack-Lesh, Goodwin, & McEchron, 2013). Hypoarticulation 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 appears 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 (ASR) systems, 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. On the basis of these results, the debate appears to be only 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, in terms of both the vowels and the formants considered for any given analysis.

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, although one 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 problem that is different from the one solved by 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 infantlike 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, one 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 a 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 will become clear, defining a teacher requires defining a data selection method that is intended to induce the defined concept in the defined learner. Throughout the article, the words *teacher* and *learner* are used to refer to the definitions in this section; we 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? Although 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 (nonadult) 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.

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

Building on previous research formalizing phonetic categories, we adopt a Gaussian mixture model (GMM) framework (de Boer & Kuhl, 2003; Feldman, Griffiths, Goldwater, & Morgan, 2013; McMurray, Aslin, & Toscano, 2009; Vallabha, McClelland, Pons, Werker, & Amano, 2007). 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, respectively, which we capture with three-dimensional Gaussians.

A GMM is defined by the probability density function

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

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 $\mathcal{N}(X \mid \mu, \Sigma)$ is the Normal (Gaussian) probability density function applied to the data X given μ and Σ.

It is important to note that 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, one 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; 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, the person 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 ith 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 (Bonawitz et al., 2011; Gweon, Pelton, Konopka, & Schulz, 2014; Shafto & Goodman, 2008; Shafto et al., 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 μ (all the means μ) and Σ (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 (opt) generates data X to be proportional to the posterior probability of the correct hypothesis given that value of X. Formally,

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

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 have relied on the assumption that the pedagogical intent of data can be measured by its benefit to learners. To the contrary, as will be clear, the benefit of data to learners is not a strict indication of the pedagogical intent of data even in the 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, & Wheeler, 1995) by manipulating first, second, and third formant values. Formants are known to correlate with vowel identity (Hillenbrand et al., 1995; Peterson & Barney, 1952), though the dimensions that listeners use when storing and categorizing sounds may be more complex than absolute encoding of formant frequencies (Apfelbaum & McMurray, 2015; McMurray & Jongman, 2011; Monahan & Idsardi, 2010; Peterson, 1961). Listeners' reliance on perceptual dimensions may also change over the course of development (Jusczyk, 1992; Nittrouer, 2004). Thus, our characterization is a significant simplification of the real-world problem. It 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. On the other hand, the task 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 12 American English vowel phonemes and their first, second, and third formants, F_1, F_2, and F_3, respectively. Hillenbrand et al. (1995) provided 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.

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 (Burnham et al., 2002; Cristia & Seidl, 2013; Kuhl et al., 1997), the corner vowels are hyperarticulated. 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 and because teaching is meant to communicate the entire category model, not just pairs of phonemes. Most vowel pairs are hyperarticulated, but consistent with IDS, and contrary to previous arguments that IDS is not for teaching (Cristia & Seidl, 2013), the simulated teaching data include hypoarticulation of some vowel pairs. Figure 3 shows the predicted effects on within-category variability. Consistent with IDS (Cristia & Seidl, 2013; de Boer & Kuhl, 2003), but contrary to previous arguments (McMurray et al., 2013), the statistically optimal input includes increases in within-category variability for most categories. Of note is the difference in behavior 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 12 phonemes decrease F_1–F_2 covariance, six of 12 decrease F_3–F_1 covariance, and four of 12 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 hypoarticulation change when the three-formant data are flattened onto two dimensions (see Figure 2A and 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 hyperarticulated in both sets, but many of the pairs that are hyperarticulated in three-formant space show little change, or are hypoarticulated, in two-formant space. This demonstrates that measures (and thus

Table 1
List of Phonemes in International Phonetic Alphabet (IPA) Transcription With Means and Variances Calculated From Hillenbrand, Getty, Clark, and Wheeler (1995)

			Mean			Variance			Covariance		
IPA	Example	*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	2,332.47	2,972.68	4,627.84	25,475.73	40,006.61	−4,247.73	−1,274.09	21,255.98
ɑ	pot	47	916.36	1,525.83	2,822.57	8,449.84	15,615.80	27,556.25	4,354.50	1,197.37	448.93
ɔ	bought	47	801.02	1,188.28	2,819.21	5,172.15	16,614.68	44,701.74	6,057.43	128.67	99.29
δ	bet	48	726.67	2,062.54	2,952.35	5,454.06	20,402.51	36,093.30	−854.33	3,539.42	11,775.23
e	bait	44	536.86	2,517.09	3,049.86	3,807.70	24,872.41	32,855.10	−1,656.22	−1,608.30	19,084.57
ɝ	Bert	40	526.60	1,589.35	1,929.85	2,193.73	12,356.90	17,234.28	−402.32	989.35	10,092.08
I	bit	48	484.31	2,369.10	3,057.12	1,181.03	22,330.69	36,138.92	−182.84	1,726.00	19,153.52
i	beet	45	435.47	2,755.96	3,372.76	1,662.21	20,746.41	56,255.83	967.00	1,010.07	18,241.44
o	boat	48	555.46	1,035.52	2,828.29	6,496.21	15,020.30	35,040.38	6,953.69	−16.69	771.31
ʊ	put	48	518.65	1,228.56	2,829.44	1,695.72	20,907.53	33,424.00	2,399.33	232.84	1,976.00
Λ	but	48	760.19	1,415.67	2,900.92	3,312.88	13,318.10	29,810.38	2,538.87	3,730.06	6,977.70
u	boot	48	459.67	1,105.52	2,735.40	1,496.06	42,130.34	19,576.20	−417.93	−57.95	2,436.00

Note. F_1, F_2, and F_3 = first, second, and third formants, respectively.

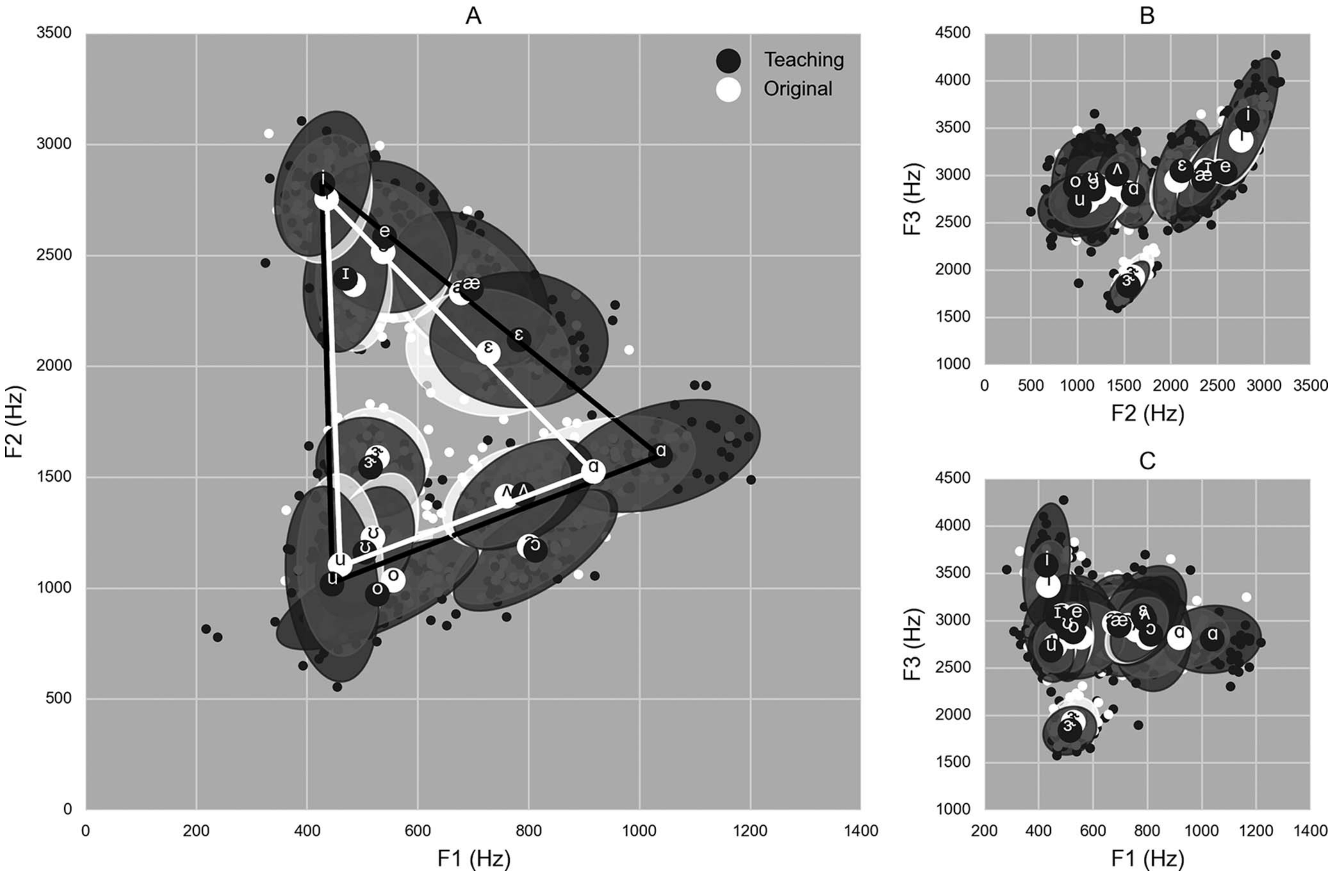

Figure 1. Distributions of vowels along first, second, and third formants (F_1, F_2, and F_3, respectively) in adult-directed speech (ADS; light) and speech optimized for the learner (dark). Differences in distributions correspond to the properties of infant-directed speech. Labels are placed at each 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, whereas a random subset of the teaching data is represented. The light and dark triangles represent the corner vowel triangles for adult-directed and teaching examples, respectively.

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 hypoarticulated in the two-formant projection, because the data were not generated to teach using only F_1 and F_2. More broadly, the absolute formant values that are typically analyzed in IDS research differ from the relative formant encodings hypothesized in many perceptual theories. It has been suggested, for example, that listeners rely on ratios among formants (Miller, 1989; Monahan & Idsardi, 2010; Peterson, 1961) or on comparisons of formants among different vowels from the same speaker (Cole, Linebaugh, Munson, & McMurray, 2010; Gerstman, 1968; Lobanov, 1971). Inaccurate assumptions about perceptual dimensions can potentially lead to incorrect conclusions about whether IDS is for teaching.

The simulated teaching data include some divergences from human IDS. IDS studies have focused on different languages and dialects and different interior vowels, but because the model output is designed to teach an American English phonetic category model, we limit our discussion of systematic deviations to those between the model output and American English IDS. Though the corner vowels hyperarticulate in the teaching data, American English IDS corner vowels hyperarticulate more uniformly (see Cristia & Seidl, 2013; Kuhl et al., 1997) than do the teaching data, which exhibit most hyperarticulation 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, whereas 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/–/I/ pair did not change (or hypoarticulated, 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 pre-

[2] We assume McMurray, Kovack-Lesh, Goodwin, and McEchron (2013) focused on native American English speakers, though they specified only that participants were “from the Ripon, WI area” and “all were Caucasian and lived in homes where English was the primary language” (p. 366).

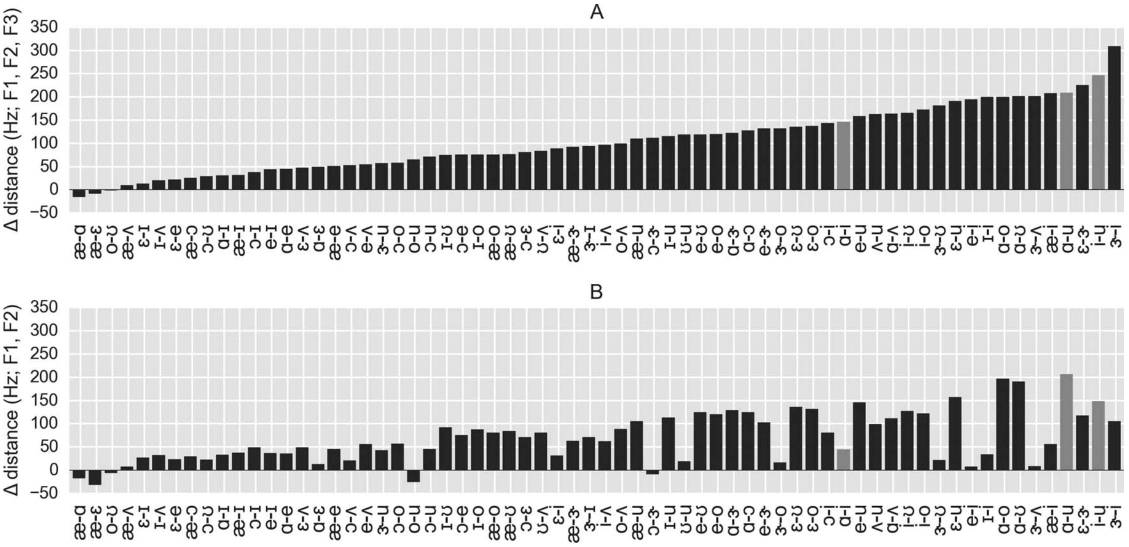

Figure 2. Change in euclidean distance (Hz; vertical axis) between phonemes pairs (horizontal axis) from adult-directed speech to teaching data: given the full, three-formant data (Panel A) and given the three-formant data projected onto the F_1–F_2 plane (Panel B). Gray bars represent corner vowel pairs. F_1, F_2, and F_3 = first, second, and third formants, respectively.

vious 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 one should bother investigating learning at all. Certain patterns 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 learners, 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 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 earlier 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 (6,000 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 hypoarticulation 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's assumed prior and the learner's prior. We evaluated the DPGMM on the basis of its inferred assignment at the 500th 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. Although 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

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

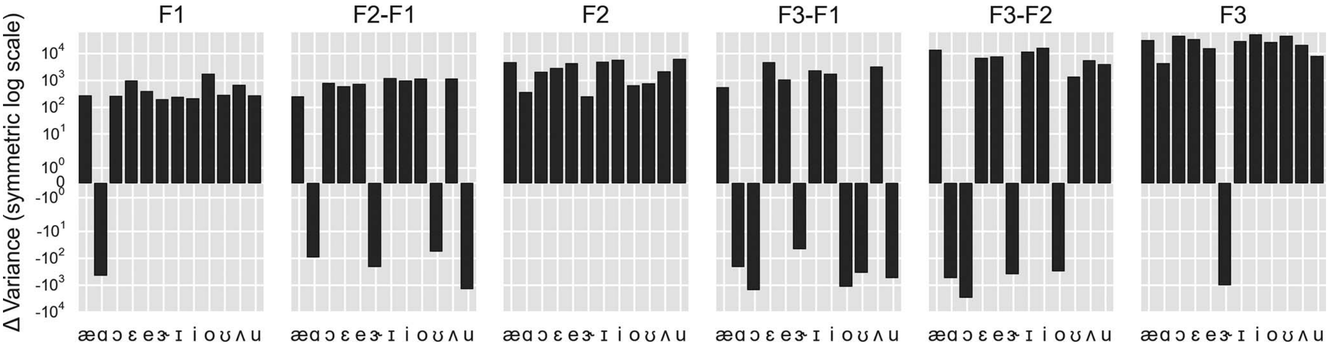

Figure 3. Change in variance, and covariance (symmetric log scale vertical axis) from adult-directed speech to teaching data for each phoneme (horizontal axis). F_1, F_2, and F_3 = first, second, and third formants, respectively.

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 provided with only the number of categories and does not care about their labels; a GMM can label k categories $k!$ different ways. And the DPGMM, in addition to not caring about labels, is not guaranteed to have the same number of categories as the true distributions. We used 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 1 belongs to Category 1, Datum 2 belongs to Category 2, and Data 3 and 4 belong to Category 3. 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 **U** and **V** of N data points into i and j categories, ARI is computed as follows:

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] \Big/ \binom{N}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] \Big/ \binom{N}{2}}, \qquad (3)$$

where n_{ij} is the number of data points assigned to i in **U** and j in **V**, a_i is the sum $\sum n_{ij}$, and b_j is the sum $\sum n_{ij}$. 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 **U** and the same category in **V**, and the number of data points that are assigned to different categories in **U** and different categories in **V**.

Figure 4 (top row) shows that the teaching data (dark) lead to improved classification over ADS (light) data in each of the algorithms we tested. Of the four algorithms, DPGMM performed the worst on the ADS data. This is unsurprising because, of the four algorithms, DPGMM has the most to learn. However, DPGMM outperformed GMM on the teaching data. On the full, three-formant data, logistic regression, SVM, and GMM all performed worst in the transfer condition (gray) compared with the ADS-only and teaching-data-only conditions, whereas the target learner (DPGMM) classified ADS data better after having learned from the teaching data. These results show that the teaching data are themselves more classifiable than are ADS and improve classification of ADS, in this case, only for the class of learner for which they were intended: the class of learner that 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.

Many of the induced ARI distributions in Figure 4 are multimodal. Two-sample Kolmogorov–Smirnov (KS) tests indicated that the distribution of ARI given three-formant ADS and teaching data differed under each algorithm; the statistic for each is significant at the $p < 10^{-40}$ level (see Table 2).[4] The categorization outcome differed when the three-formant data were projected onto the F_1–F_2 plane (see Figure 4, bottom row). Categorization performance generally decreased when F_3 was removed. More features (dimensions) provided learners with more information by which they could form categories. For example, Figure 1B and C shows that locating and categorizing /ɝ/ (as in Bert) becomes trivial given F_3.

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.

Learning as a Function of the Number of Examples

Here we investigated 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, . . ., 2,048 examples of each phoneme. The

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

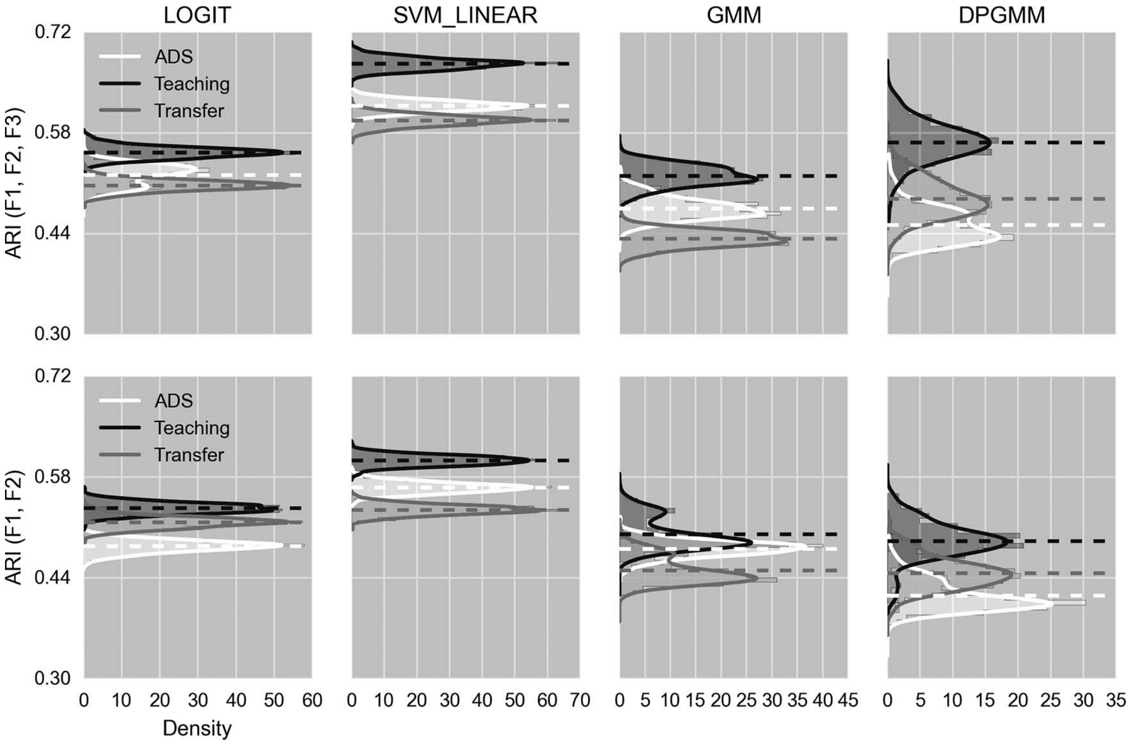

Figure 4. Distributions of adult-directed speech (ADS) for four categorization algorithms (logistic regression [LOGIT], support vector machine [SVM] with linear kernel, finite Gaussian mixture model [GMM] using expectation-maximization, and Dirichlet process Gaussian mixture model [DPGMM]) given ADS data generated from the empirical distribution (light), simulated teaching data (dark), and ADS after having learned from teaching data (transfer; gray). Dashed lines indicate the mean of each distribution. Top row: Adjusted Rand index (ARI) given the original, three-dimensional data. Bottom row: ARI given the data with the third formant removed. F_1, F_2, and F_3 = first, second, and third formants, respectively.

results can be seen in Figure 5. The behavior induced in the DPGMM by the ADS (light) and Teaching (dark) data differ. Adding ADS data appears not to benefit the learner between about 32 and 256 examples per phoneme, whereas adding teaching data continues to improve categorization at an approximately logarithmic rate. This suggests 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 teaching examples decreases; and at 2,048 examples per phoneme the transfer of IDS results in mean performance similar to that for 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 hyperarticulate examples, but the pedagogical usefulness of hypoarticulation and variance increase deserves discussion. Keep in mind that the teacher seeks to increase the likelihood of a globally correct inference. Hypoarticulation can improve categorization when it is the result of disambiguating movement—that is, movement of one cluster away from another cluster that it may be mistaken for. Increased variability can be used to mitigate any negative effects of hypoarticulation by making close or overlapping clusters more distinguishable from each other. Imagine two very closely overlapping, circular clusters: Examples from these clusters may appear to come from one large cluster. If one wishes to express that there are two clusters, one could stretch each cluster perpendicularly so the resulting data manifest as an *X* rather than as a single Gaussian blob; indeed, the teaching model produces this behavior.

The teaching data offer similar examples of how hypoarticulation and increased variability, when employed systematically, do not necessarily reduce learning. For purposes of clarity, we look at only the F_1–F_2 plane (see Figure 1A). The phonemes (/ɝ/; /u/; /ʊ/, as in put; /o/, as in boat) are difficult to distinguish in ADS. In the teaching data, /u/, /ʊ/, and /o/ are pressed into each other (hypoarticulated), 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 hypoarticulation and directional changes in variance reduce the muddling effect of general increases in within-phoneme variance. Looking at the categorization

Table 2
Uncorrected Kolmogorov–Smirnov (KS) Test Statistics for the Distributions of Classification Scores (Adjusted Rand Index) as Seen in Figure 4

Algorithm and comparison	F_1, F_2, F_3		F_1, F_2	
	KS	*p*	KS	*p*
LOGIT				
ADS–Teaching	.894	<.0001	.998	<.0001
ADS–Transfer	.584	<.0001	.972	<.0001
Teaching-Transfer	.996	<.0001	.828	<.0001
SVM (linear)				
ADS–Teaching	1.0	<.0001	.994	<.0001
ADS–Transfer	.822	<.0001	.976	<.0001
Teaching–Transfer	1.0	<.0001	1.0	<.0001
GMM				
ADS–Teaching	.872	<.0001	.434	<0001
ADS–Transfer	.932	<.0001	.69	<.0001
Teaching–Transfer	1.0	<.0001	.830	<.0001
DPGMM				
ADS–Teaching	.946	<.0001	.886	<.0001
ADS–Transfer	.54	<.0001	.596	<.0001
Teaching–Transfer	.858	<.0001	.726	<.0001

Note. F_1, F_2, and F_3 = first, second, and third formants, respectively; ADS = adult-directed speech; LOGIT = logistic regression; SVM = support vector machines; GMM = Gaussian mixture model; DPGMM = Dirichlet process Gaussian mixture model. *p*s range from $\approx 10^{-220}$ to $\approx 10^{-41}$.

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 ($M_{\text{ADS}} = 0.431$, $M_{\text{Teach}} = 0.403$; $KS(500, 500) = 0.716$, $p < .001$; $d = 2.019$), and logistic regression performs similarly on ADS and teaching data ($M_{\text{ADS}} = 0.294$, $M_{\text{Teach}} = 0.292$; $KS(500, 500) = 0.070$, $p = .166$; $d = 0.109$). GMM performs better on the teaching data ($M_{\text{ADS}} = 0.347$, $M_{\text{Teach}} = 0.353$; $KS(500, 500) = 0.184$, $p < .001$; $d = -0.301$), as does DPGMM ($M_{\text{ADS}} = 0.275$, $M_{\text{Teach}} = 0.283$; $KS(500, 500) = 0.14$, $p < .001$; $d = -0.231$). These result show first that hypoarticulation 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 article we 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, infantlike learners using the F_1, F_2, and F_3 formants. In the process, we 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. Hyperarticulation (expansion) of the corner vowels has been identified as a feature that would facilitate learning. However, hypoarticulation such as observed between /I/ and /i/ by Cristia and Seidl (2013), as well as 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, hypoarticulation and increases in variance are indeed consistent with teaching. Our analysis leads to predictions about when and why one may see these surprising properties. Hypoarticulation appears when vowels move away from more confusable alternatives. To compensate for this, hypoarticulated

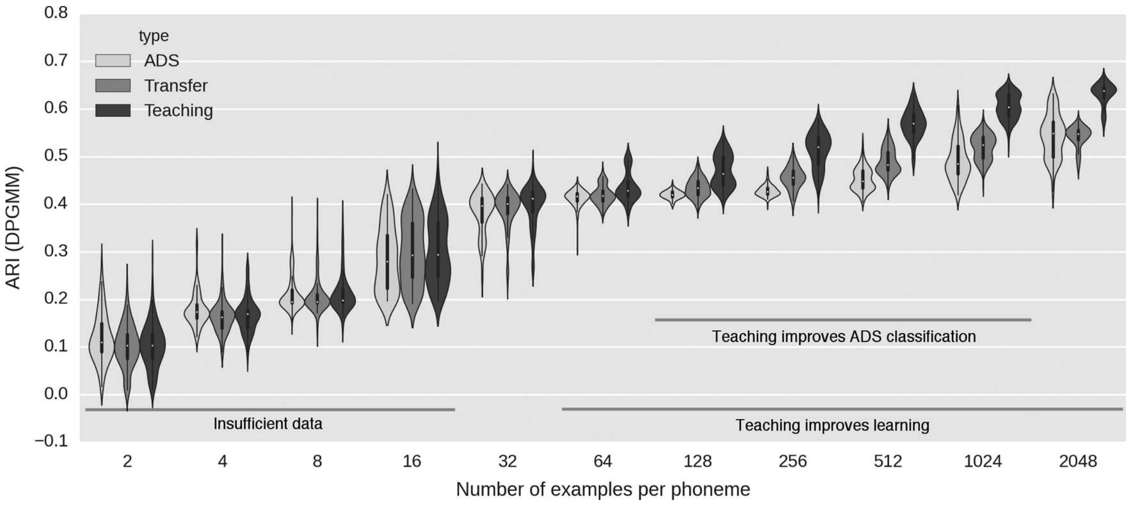

Figure 5. The adjusted Rand index (ARI) as a function of the number of examples per phoneme for the Dirichlet process Gaussian mixture model (DPGMM) given adult-directed speech (ADS) data (light), teaching data (dark), and ADS data after learning from teaching data (transfer; gray). Low ARI scores at 2, 4, 8, and 16 examples per phoneme indicate that the DPGMM has insufficient data. At 32 examples per phoneme, the teaching data begin to improve learning performance. From 128 to 1,024 examples per phoneme, teaching data improve classification of ADS (transfer).

categories appear in conjunction with hyperarticulation 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 hypoarticulated in $F_1 \times F_2$ but are hyperarticulated 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, whereas /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 either that infants 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, whereas algorithms that assume labeled data or a known number of categories do not (see Figure 4). Our results on the basis of 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, in terms of both 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 F_1 and F_2. 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, hypoarticulation 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, hypoarticulation 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 (Cristia & Seidl, 2013; McMurray et al., 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 that 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 for teaching only phonetic categories, then 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 and may also differ in the perceptual dimensions we assume.[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, 2013). 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 multimodal, and Markov chains often find themselves stuck in local maxima. Pilot research has suggested that data from single chains are far more beneficial to learners than are the data aggregated over chains-perhaps due to lower within-phoneme variability compared with aggregated data. We used 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 interspeaker variability).

This work is also relevant to the articulation literature, where the theoretical underpinnings of speakers' speech manipulations are under debate (see Jaeger & Buz, in press). The teaching model, coupled with a temporal model of articulation, could predict hyper- or hypoarticulation, as well as duration increases or decreases. Temporal

[5] Additionally, phonemes in Hillenbrand, Getty, Clark, and Wheeler (1995) were measured only from words beginning with an *h* and ending with a *d;* for example, /ɑ/, /i/, and /u/ were taken from only the words *hod, heed,* and *who'd,* respectively.

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 (Galati & Brennan, 2010; Lindblom, 1990; Munson & Solomon, 2004) may in fact be consistent with pedagogical manipulation (as explicitly suggested by Lindblom, 1990, and Jaeger, 2013). Related research has indicated that speakers adapt when their communications are unsuccessful (Buz, Tanenhaus, & Jaeger, 2016; Schertz, 2013; Stent, Huffman, & Brennan, 2008). 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 (Bonawitz et al., 2011; Gergely, Bekkering, & Király, 2002; Gweon et al., 2014; Koenig & Harris, 2005). 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 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 data sets 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,* 409–429.

Apfelbaum, K. S., & McMurray, B. (2015). Relative cue encoding in the context of sophisticated models of categorization: Separating information from categorization. *Psychonomic Bulletin and Review, 22,* 916–943.

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

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

Bonawitz, E., Shafto, P., Gweon, H., Goodman, N. D., Spelke, E., & Schulz, L. (2011). The double-edged sword of pedagogy: Instruction limits spontaneous exploration and discovery. *Cognition, 120,* 322–330. http://dx.doi.org/10.1016/j.cognition.2010.10.001

Burnham, D., Kitamura, C., & Vollmer-Conna, U. (2002, May 24). What's new, pussycat? On talking to babies and animals. *Science, 296,* 1435. http://dx.doi.org/10.1126/science.1069587

Buz, E., Tanenhaus, M. K., & Jaeger, T. F. (2016). Dynamically adapted context-specific hyper-articulation: Feedback from interlocutors affects speakers' subsequent pronunciations. *Journal of Memory and Language*. Advance online publication. http://dx.doi.org/10.1016/j.jml.2015.12.009

Cole, J., Linebaugh, G., Munson, C. M., & McMurray, B. (2010). Unmasking the acoustic effects of vowel-to-vowel coarticulation: A statistical modeling approach. *Journal of Phonetics, 38,* 167–184.

Cristia, A., & Seidl, A. (2013). The hyperarticulation hypothesis of infant-directed speech. *Journal of Child Language, 13,* 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:* 129. http://dx.doi.org/10.1121/1.1613311

Escobar, M. D. & West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association, 90,* 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,* 751–778. http://dx.doi.org/10.1037/a0034245

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

Gergely, G., Bekkering, H., & Király, I. (2002, February 14). Rational imitation in preverbal infants. *Nature, 415,* 755. http://dx.doi.org/10.1038/415755a

Gergely, G., Egyed, K., & Király, I. (2007). On pedagogy. *Developmental Science, 10,* 139–146. http://dx.doi.org/10.1111/j.1467-7687.2007.00576.x

Gerstman, L. J. (1968). Classification of self-normalized vowels. *IEEE Transactions on Audio and Electroacoustics, 16,* 78–80.

Graf Estes, K., & Hurley, K. (2013). Infant-directed prosody helps infants map sounds to meanings. *Infancy, 18,* 797–824. http://dx.doi.org/10.1111/infa.12006

Gweon, H., Pelton, H., Konopka, J. A., & Schulz, L. E. (2014). Sins of omission: Children selectively explore when teachers are underinformative. *Cognition, 132,* 335–341. http://dx.doi.org/10.1016/j.cognition.2014.04.013

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

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

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

Jaeger, T. F. (2013). Production preferences cannot be understood without reference to communication. *Frontiers in Psychology, 4:* 230. http://dx.doi.org/10.3389/fpsyg.2013.00230

Jaeger, T. F. & Buz, E. (in press). Signal reduction and linguistic encoding. In E. M. Fernandez & H. S. Cairns (Eds.), *Handbook of psycholinguistics. Chichester,* United Kingdom: Wiley-Blackwell.

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,* 158–182.

Jusczyk, P. W. (1992). Developing phonological categories from the speech signal. In C. A. Ferguson, L. Menn, & C. Stoel-Gammon (Eds.), *Phonological development: Models, research, implications* (pp. 17–64). Timonium, MD: York.

Kirchhoff, K., & Schimmel, S. (2005). Statistical properties of infant-directed versus adult-directed speech: Insights from speech recognition. *Journal of the Acoustical Society of America, 117,* 2238. http://dx.doi.org/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,* 148–203. http://dx.doi.org/10.1037/a0038695

Koenig, M., & Harris, P. L. (2005). Preschoolers mistrust ignorant and inaccurate speakers. *Child development, 76,* 1261–1277. http://dx.doi.org/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 1). Cross-language analysis of phonetic units in language addressed to infants. *Science, 277,* 684–686. http://dx.doi.org/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.

Lobanov, B. M. (1971). Classification of Russian vowels spoken by different speakers. *Journal of the Acoustical Society of America, 49,* 606–608.

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). Statistical learning of phonetic categories: Insights from a computational approach. *Developmental Science, 12,* 369–378. http://dx.doi.org/10.1111/j.1467-7687.2009.00822.x

McMurray, B., & Jongman, A. (2011). What information is necessary for speech categorization? Harnessing variability in the speech signal by integrating cues computed relative to expectations. *Psychological Review, 118,* 219–246.

McMurray, B., Kovack-Lesh, K., Goodwin, D., & McEchron, W. (2013). Infant directed speech and the development of speech perception: Enhancing development or an unintended consequence? *Cognition, 129,* 362–378. http://dx.doi.org/10.1016/j.cognition.2013.07.015

Miller, J. D. (1989). Auditory-perceptual interpretation of the vowel. *Journal of the Acoustical Society of America, 85,* 2114–2134.

Monahan, P. J., & Idsardi, W. J. (2010). Auditory sensitivity to formant ratios: Toward an account of vowel normalisation. *Language and Cognitive Processes, 25,* 808–839.

Munson, B., & Solomon, N. P. (2004). The effect of phonological neighborhood density on vowel articulation. *Journal of Speech, Language, and Hearing Research, 47,* 1048–1058.

Murphy, K. P. (2007). *Conjugate Bayesian analysis of the Gaussian distribution*. Unpublished technical report, Department of Computer Science, University of British Columbia. https://www.cs.ubc.ca/~murphyk/Papers/bayesGauss.pdf

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics, 9,* 249–265.

Nittrouer, S. (2004). The role of temporal and dynamic signal components in the perception of syllable-final stop voicing by children and adults. *Journal of the Acoustical Society of America, 115,* 1777–1790.

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.

Peterson, G. E. (1961). Parameters of vowel quality. *Journal of Speech and Hearing Research, 4,* 10–29.

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America, 24,* 175–184.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association, 66,* 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,* 110–120. http://dx.doi.org/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,* 1144–1167. http://dx.doi.org/10.1037/a0020511

Schertz, J. (2013). Exaggeration of featural contrasts in clarifications of misheard speech in English. *Journal of Phonetics, 41,* 249–263.

Shafto, P., & Goodman, N. D. (2008). Teaching games: Statistical sampling assumptions for learning in pedagogical situations. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 1632–1637). Austin, TX: Cognitive Science Society.

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

Singh, L., Nestor, S., Parikh, C., & Yull, A. (2009). Influences of infant-directed speech on early word recognition. *Infancy, 14,* 654–666. http://dx.doi.org/10.1080/15250000903263973

Stent, A. J., Huffman, M. K., & Brennan, S. E. (2008). Adapting speaking after evidence of misrecognition: Local and global hyperarticulation. *Speech Communication, 50,* 163–178.

Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association, 101,* 1566–1581.

Thiessen, E. D., Hill, E. a., & Saffran, J. R. (2005). Infant-directed speech facilitates word segmentation. *Infancy, 7,* 53–71. http://dx.doi.org/10.1207/s15327078in0701_5

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

Vallabha, G. K., McClelland, J. L., Pons, F., Werker, J. F., & Amano, S. (2007). Unsupervised learning of vowel categories from infant-directed speech. *PNAS: Proceedings of the National Academy of Sciences of the United States of America, 104,* 13273–13278. http://dx.doi.org/10.1073/pnas.0705369104

Appendix A

Details of the Model

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

Learner Model

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

$$P(X \mid \theta, \pi) = \prod_{i=1}^{n} \sum_{j=1}^{k} \pi_j \mathcal{N}(x_i \mid \mu_i, \Sigma_i), \tag{A1}$$

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

(*Appendices continue*)

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

$$P(X \mid \theta, Z) = \prod_{i=1}^{n} \sum_{j=1}^{k} \mathcal{N}(x_i \mid \mu_i, \Sigma_i)\delta_{z_i, j}, \tag{A2}$$

where $\delta_{zi,\,j}$ 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. We assumed that μ_j and Σ_j are distributed according to Normal Inverse-Wishart ($\mathcal{NIW}$). Though we made this choice primarily for mathematical convenience, priors of this and similar form have been used successfully to model human behavior (e.g., Feldman et al., 2013; Kleinschmidt & Jaeger, 2015):

$$\mu_j, \Sigma_j \sim \mathcal{NIW}(\mu_0, \Lambda_0, \kappa_0, \nu_0) \quad \forall j \in \{1, \ldots, k\}, \tag{A3}$$

which implies

$$\Sigma_j \sim \text{Inverse-Wishart}_{\nu_0}(\Lambda_0^{-1}), \tag{A4}$$

$$\mu_j \mid \Sigma_j \sim \mathcal{N}(\mu_0, \Sigma_k/\kappa_0) \quad \forall j \in \{1, \ldots, k\}, \tag{A5}$$

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:

$$\nu_0 = 3, \tag{A6}$$

$$\kappa_0 = 1, \tag{A7}$$

$$\mu_0 = \frac{1}{N}\sum_{i=1}^{N} X_i, \tag{A8}$$

$$\Lambda_0 = \frac{1}{K}\sum_{k=1}^{K} \Sigma(X_k), \tag{A9}$$

where $\Sigma(X_k)$ 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 introduced 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 were assigned to components as follows:

$$P(z_i = j \mid Z^{-i}, \alpha) = \begin{cases} \dfrac{n_j}{n - 1 + \alpha} & \text{if } j \in 1, \ldots, k \\ \dfrac{\alpha}{n - 1 + \alpha} & \text{if } j = k + 1 \end{cases}, \tag{A10}$$

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,

$$P(\theta, Z \mid X, \mu_0, \Lambda_0, \kappa_0, \nu_0, \alpha) = \frac{P(Z \mid \alpha)P(\theta \mid \mu_0, \Lambda_0, \kappa_0, \nu_0)P(X \mid \theta, Z)}{P(X \mid \mu_0, \Lambda_0, \kappa_0, \nu_0, \alpha)}. \tag{A11}$$

The Normal Inverse-Wishart prior allowed us to calculate the marginal likelihood, $P(X \mid \mu_0, \Lambda_0, \kappa_0, \nu_0, \alpha)$, 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 13 or fewer), we could exactly calculate the above quantity via enumeration. Expanding the terms, the numerator is

$$P(Z \mid \alpha)\left(\prod_{j=1}^{k} \mathcal{NIW}(\mu_j, \Sigma_j \mid \mu_0, \Lambda_0, \kappa_0, \nu_0)\right) \times \prod_{j=1}^{k} \mathcal{N}(\{x_i \in X : Z_i = j\} \mid \mu_j, \Sigma_j), \tag{A12}$$

where the first term, $P(Z \mid \alpha)$, 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 A11 is calculable by summing over all possible assignment vectors, $\{Z \in \mathfrak{Z}\}$, and integrating over all possible component parameters,

$$P(X \mid \mu_0, \Lambda_0, \kappa_0, \nu_0, \alpha) = \sum_{Z \in \mathfrak{Z}} P(Z \mid \alpha) \prod_{j=1}^{k_Z} \iint_{\theta} \mathcal{N}(\{x_i \in X : Z_i = j\} \mid \theta) \times \mathcal{NIW}(\theta \mid \mu_0, \Lambda_0, \kappa_0, \nu_0)d\theta \tag{A13}$$

$$= \sum_{Z \in \mathfrak{Z}} P(Z \mid \alpha) \prod_{j=1}^{k_Z} P(\{x_i \in X : Z_i = j\} \mid \mu_0, \Lambda_0, \kappa_0, \nu_0), \tag{A14}$$

where k_Z is the number of components in the assignment Z and $P(\{x_i \in X : Z_i = j\} \mid \mu_0, \Lambda_0, \kappa_0, \nu_0)$ 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 normalized over all possible data X,

(*Appendices continue*)

$$P_{\text{opt}}(X|\theta, Z, \mu_0, \Lambda_0, \kappa_0, \nu_0, \alpha) \propto \frac{P(\theta, Z|X, \mu_0, \Lambda_0, \kappa_0, \nu_0, \alpha)}{\int_X P(\theta, Z|X, \mu_0, \Lambda_0, \kappa_0, \nu_0, \alpha)dX}, \tag{A15}$$

$$= \frac{\dfrac{P(Z|\alpha)P(X|\theta, Z)P(\theta|\mu_0, \Lambda_0, \kappa_0, \nu_0)}{P(X|\mu_0, \Lambda_0, \kappa_0, \nu_0, \alpha)}}{\displaystyle\int_X \frac{P(X|\theta, Z)P(\theta|\mu_0, \Lambda_0, \kappa_0, \nu_0)P(Z|\alpha)}{P(X|\mu_0, \Lambda_0, \kappa_0, \nu_0, \alpha)}dX}. \tag{A16}$$

The term

$$P(\theta, Z|X, \mu_0, \Lambda_0, \kappa_0, \nu_0, \alpha) = \frac{P(X|\theta, Z)P(\theta|\mu_0, \Lambda_0, \kappa_0, \nu_0)P(Z|\alpha)}{P(X|\mu_0, \Lambda_0, \kappa_0, \nu_0, \alpha)} \tag{A17}$$

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, $P(X \mid \mu_0, \Lambda_0, \kappa_0, \nu_0, \alpha)$. The teacher's optimization of the choice of data is captured by the normalizing constant,

$$\int_X P(\theta, Z|X, \mu_0, \Lambda_0, \kappa_0, \nu_0, \alpha)dX. \tag{A18}$$

We avoided 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

$$A(X'|X) = \min\left[1, \frac{P(X'|\theta, Z)P(X|\mu_0, \Lambda_0, \kappa_0, \nu_0, \alpha)}{P(X|\theta, Z)P(X'|\mu_0, \Lambda_0, \kappa_0, \nu_0, \alpha)}\right]. \tag{A19}$$

Appendix B

Algorithm for Generating Samples

The normalizing constant in Equation 2 (also Equation A18 in Appendix A) is analytically intractable. We used 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,

$$A = \frac{f(x')q(x|x')}{f(x)q(x'|x)}. \tag{B1}$$

It is easy to see that

$$\frac{f(x')q(x|x')}{f(x)q(x'|x)} = \frac{cf(x')q(x|x')}{cf(x)q(x'|x)} = \frac{p(x')q(x|x')}{p(x)q(x'|x)}. \tag{B2}$$

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

$$A = \frac{f(x')}{f(x)}, \tag{B3}$$

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

$$P(x'|x) = \min[1,\ A]. \tag{B4}$$

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

$$P(X'|X) = \min\left[1, \frac{P(X'|Z, \mu, \Sigma)P(X)}{P(X|Z, \mu, \Sigma)P(X')}\right]. \tag{B5}$$

For the simulations, the sampler simulated one data point for each phoneme (12 total). X comprised 12 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 B5.

The final data comprise samples from 10 independent runs of the sampler. The first 500 samples of each run were discarded, and then each 20th sample was collected until 1,000 samples had been collected. The full set of data thus contained 10,000 total samples of 12 data points each (one for each of the 12 phonemes) for a total of 120,000 examples. Aggregating data over speakers is common practice in the IDS literature; we conducted analyses on data aggregated over independent runs of the sampler.

Received June 24, 2015
Revision received January 28, 2016
Accepted January 30, 2016 ■